Università di Genova
DIBRIS DIPARTIMENTO DI INFORMATICA, BIOINGEGNERIA, ROBOTICA E INGEGNERIA DEI SISTEMI

Integrating Gaze in Vision-Language Models: Towards Human-Like Visual Reasoning

Dario Valentini

Master Thesis

Università di Genova, DIBRIS Via Opera Pia, 13 16145 Genova, Italy
https://www.dibris.unige.it/

MSc Computer Science
Data Science and Engineering Curriculum

Integrating Gaze in Vision-Language Models: Towards Human-Like Visual Reasoning

Dario Valentini

Advisors:
Matteo Moro
Lucia Schiatti

Examiner:
Viviana Mascardi

March, 2026

Abstract

This thesis investigates the integration of human gaze information to enhance the visual comprehension and reasoning abilities of Vision-Language Models (VLMs) toward a more human-aligned visual understanding. Despite recent progress, current models exhibit limitations not only in high-level cognitive reasoning but also in fundamental visual recognition. Gaze data was collected from 30 participants performing recognition and reasoning tasks, using the images from CogBench, a visual reasoning benchmark. To study the role of gaze, three integration strategies were proposed: (1) a training-free gaze-based reweighting of visual patch embeddings to evaluate gaze as an external attention prior; (2) selective fine-tuning of the multimodal projector to adapt the model to gaze-modulated representations; and (3) joint embedding of image and gaze signals via parameter-efficient adaptation of visual and projection components. These approaches were implemented on several open-source VLMs, and evaluated using the recognition and cognition metrics defined in CogBench. Results indicate that naive gaze weighting yields limited improvements, whereas selective fine-tuning strategies produce dimension-specific performance variations. Analysis across reasoning categories reveals that gaze provides stronger benefits in certain structured reasoning settings. Overall, these findings highlight that human gaze contains informative visual priors, while its effective integration requires targeted architectural adaptation and careful optimization.

Table of Contents

Chapter 1

Introduction

1.1 Motivation

Large Vision-Language Models (LVLMs) are multimodal systems that can process both textual and visual input modalities [36, 33]. They are increasingly adopted in scientific research and real-world applications, with image understanding and description being among their central goals [7].

Several benchmarks and metrics are currently used to evaluate these models [61, 2, 15]. However, many existing evaluation settings mainly capture recognition performance or the fluency of generated captions, and do not explicitly assess whether models can perform high-level visual reasoning about complex scenes [15, 39, 31, 68]. As a result, it is difficult to quantify how well LVLMs support structured interpretation of scene dynamics, relationships, and implicit information that humans typically infer from images [69, 47].

CogBench [51] addresses this limitation by proposing a reasoning-oriented evaluation framework for LVLMs. Rather than focusing only on generic caption quality, it evaluates model outputs on a set of reasoning dimensions defined over image descriptions, while also measuring recognition separately. The results reported for a set of state-of-the-art open source Vision-Language Models (VLMs) show that, although recognition performance is above chance level, cognition scores across several reasoning dimensions remain limited. In particular, reasoning involving events, relationships, and mental state inference is still challenging for these models. These findings indicate that current open source LVLMs do not yet achieve human-like performance in structured visual reasoning.

Human gaze is a behavioral signal that reflects visual attention during scene perception and comprehension [65]. Patterns of fixations can therefore be interpreted as an indicator of which regions humans consider informative when recognizing entities and forming

higher-level interpretations of a scene [43]. Gaze has been explored in several vision and multimodal settings [71], but it has not been systematically studied as an auxiliary signal to support open source Vision-Language Models under a visual reasoning evaluation framework. As a consequence, it is still unclear whether providing gaze information can improve performance on structured reasoning dimensions such as those measured by CogBench.

For these reasons, this thesis investigates whether integrating gaze information into open source Vision-Language Models through multiple integration strategies can improve performance in high-level visual reasoning, by evaluating potential enhancements in recognition and cognition scores under the CogBench evaluation framework.

1.2 Problem Statement

Despite recent progress in multimodal modeling, current open source Vision-Language Models achieve only limited performance in structured visual reasoning. While they can reach acceptable levels of performance in object recognition and entity identification, results on reasoning-oriented benchmarks such as CogBench [51] show that several higher-level cognition dimensions, including event reasoning, event relationship reasoning, and mental state inference, remain challenging. This gap highlights the need to investigate mechanisms that can support more effective and structured interpretation of complex visual scenes.

The central research question of this thesis is whether the integration of human gaze information as an additional input modality can improve the recognition and cognition performance of open source Vision-Language Models under the CogBench evaluation framework. In particular, the objective is to determine whether gaze can contribute to measurable improvements across reasoning dimensions, and whether it can influence model outputs toward more structured and human-like scene interpretation.

The scope of this work is restricted to five open source Vision-Language Models evaluated using the CogBench framework. Recognition and cognition scores are computed according to the benchmark protocol, with recognition derived from entity identification and cognition derived from multiple reasoning dimensions defined over image descriptions. Gaze information is integrated into the models through three architectural strategies, ranging from direct weighting of visual features to learned adaptation mechanisms. Improvement is operationally defined as a quantitative increase in recognition and cognition scores with respect to the baseline models without gaze integration, including analysis at the level of individual reasoning dimensions.

1.3 Contributions

The main contributions of this thesis are the following.

- The collection of gaze data aligned with image descriptions and visual reasoning dimensions for all images in the CogBench dataset. This dataset extends the original image–text benchmark by incorporating human gaze information, and can support future research on multimodal reasoning and attention-guided modeling.
- The design and implementation of three distinct gaze integration strategies within open-source Vision-Language Models, enabling the systematic study of different mechanisms for incorporating gaze as an additional input signal.
- A comprehensive evaluation of the gaze-augmented models under the CogBench framework, aimed at assessing whether the inclusion of gaze information leads to measurable improvements in recognition and cognition scores for high-level reasoning, and whether consistent performance patterns emerge across different model architectures when exposed to the same gaze augmentation.
- An analysis of the impact of gaze on individual reasoning dimensions, aimed at determining whether gaze provides greater benefits for specific categories of reasoning and whether it acts as an effective form of semantic guidance in structured visual understanding.

1.4 Thesis Organization

This thesis is structured into seven chapters.

Chapter 2 provides the background necessary for this work, introducing the main concepts related to human gaze and visual attention, the architecture of Large Vision-Language Models, and common adaptation strategies used in multimodal systems.

Chapter 3 reviews the related literature. It discusses the role of gaze in cognitive science and scene understanding, examining gaze-guided vision-language architectures, and presenting existing benchmarks for image captioning and visual reasoning, with particular focus to the CogBench benchmark.

Chapter 4 describes the methodological framework of the study. It introduces the CogBench dataset and evaluation protocol, and it presents the gaze data collection process, detailing the preprocessing pipeline used to generate gaze heatmaps.

Chapter 5 outlines the experimental setup, including the evaluation protocol, the selected Vision-Language Models, and the three gaze integration scenarios investigated in this work, together with the training procedure used for model adaptation.

Chapter 6 presents the experimental results and discussion, analyzing the impact of gaze integration on recognition and cognition scores across models and reasoning dimensions, and discussing the implications and limitations of the work.

Finally, Chapter 7 concludes the thesis by summarizing the main contributions and findings of the study.

Chapter 2

Background

2.1 Human Gaze and Visual Attention

Since this thesis investigates how human gaze information can be incorporated into vision-language models, it is useful to briefly review the mechanisms that generate eye movement data and the standard representations used in computational studies. This section introduces the main eye movement events recorded by eye trackers, summarizes how gaze is used in cognitive science and computer vision, and motivates the use of gaze heatmaps as a compact representation of spatial attention.

2.1.1 Human Visual System: Eye Movements and Attention

Human vision is *foveated*: high-acuity perception is concentrated in the fovea, a small central region of the retina, while visual resolution rapidly decreases in peripheral vision. As a consequence, observers continuously move their eyes to bring task-relevant regions into foveal view [50]. During viewing, the eyes mainly alternate between two types of movements: fixations and saccades [23, 55].

Fixations are relatively stable periods in which gaze remains within a small region and visual information is acquired. In scene viewing, fixations typically last on the order of a few hundred milliseconds. Their duration is often interpreted as reflecting processing demands, and it can be influenced by multiple factors, including the complexity of the viewed scene and the observation task. For this reason, longer fixations do not necessarily correspond to regions that are inherently more relevant, but may instead indicate that additional cognitive processing was required.

Saccades are rapid, ballistic eye movements that shift the point of fixation. They are

substantially shorter than fixations, often lasting only a few tens of milliseconds, and visual sensitivity is strongly reduced during these movements because of *saccadic suppression*, a brief reduction in visual perception during eye movements.

Visual attention determines which information is prioritized for processing at a given moment. A common distinction is between *overt* attention, in which attentional selection involves an eye movement toward the attended location, and *covert* attention, in which attention shifts without a corresponding gaze shift. In natural scene viewing, these two mechanisms often interact, but eye tracking primarily provides access to overt attentional selection. For this reason, fixation locations are commonly used as an observable proxy for the spatial allocation of visual attention, while acknowledging that they do not capture the full attentional process.

By alternating between fixations and saccades, the human visual system samples a sequence of locations that collectively supports scene perception despite the limited spatial extent of high-acuity foveal vision. The resulting sequence reflects the deployment of overt attention over time and is commonly referred to as a *scanpath*. Importantly, scanpaths depend on both stimulus properties and task demands, and different instructions can lead to substantially different gaze patterns on the same image.

2.1.2 From Fixations to Heatmaps

Raw gaze samples are typically segmented into fixations and saccades using standard event-detection algorithms [13], which commonly rely on thresholding angular velocity to distinguish stable gaze periods from rapid eye movements. After segmentation, gaze can be represented either as scanpaths, that is ordered sequences of fixations that preserve temporal dynamics, or as heatmaps, which encode fixation density over space.

When the objective is to model spatial attention, heatmaps are generally preferred, as they provide a fixed-size representation that can be easily aggregated across observers and directly integrated with vision models. A common procedure consists of accumulating fixation coordinates on a discrete image grid and applying Gaussian smoothing. This smoothing approximates the spatial extent of foveal processing and reduces sensitivity to measurement noise [30, 62].

Fixations can be treated uniformly, emphasizing only where observers looked, or weighted by fixation duration, thereby introducing a temporal component. The choice depends on the intended use. In studies that focus exclusively on the spatial allocation of attention, equal weighting is often preferred, as it avoids conflating spatial distribution with viewing time. Finally, normalization is determined by the downstream application: sum-to-one normalization is appropriate when interpreting the heatmap as a probability distribution, whereas min–max normalization to $[0, 1]$ is more suitable when the map is used as a

bounded weighting signal for visual features [13].

2.2 Large Vision-Language Models

Large Vision-Language Models (**LVLMs**) are multimodal deep learning models designed to jointly process visual and textual information [7]. Given an image and a textual prompt (such as a question or an instruction), LVLMs can generate natural language outputs such as captions, answers, or multi-step explanations grounded in the image content. In recent years, the field has progressed rapidly due to larger-scale pretraining data, stronger vision backbones, and instruction tuning, leading to both closed-source [44, 18] and open-source [36, 33, 5] model families with competitive performance. In this section, we first describe the dominant architectural pattern adopted by most recent LVLMs, and then we briefly review representative open-source models and existing approaches that incorporate human gaze information.

2.2.1 General Architecture

Most recent LVLMs follow a modular architecture composed of a Visual Encoder, a Vision-Language Projector (or multimodal adapter), and a Large Language Model (LLM) [66]. The visual encoder converts the image into visual features, typically represented as a sequence of visual tokens (e.g., patch- or region-level embeddings) that preserve spatial information. The projector then maps these visual tokens into the representation space expected by the LLM. Finally, the LLM performs conditional generation, producing text conditioned on both the prompt and the projected visual tokens.

2.2.1.1 Visual Encoder

The **Visual Encoder** processes the input image and produces a compact representation that captures relevant visual semantics. Depending on the model design, the encoder can output either a single global embedding, or a sequence of embeddings corresponding to spatial regions (e.g., a grid of patch tokens). The latter is particularly common in modern LVLMs because it supports fine-grained grounding and compositional reasoning over localized image content.

In many open-source implementations the term *Vision Tower* (ViT) is used as a practical engineering label for the vision-side stack of the architecture. In the simplest case, it matches the visual encoder backbone itself; alternatively, it may also include lightweight vision-side modules such as token selection, pooling, or resampling components.

Most visual encoders used in LVLMs belong to one out of two broad families:

- **CNN-based encoders.** Convolutional Neural Networks (CNNs) extract visual features through stacked convolutional layers, gradually expanding the receptive field from local patterns (edges, textures) to higher-level semantic concepts (parts, objects, scene cues). CNNs provide strong inductive biases such as locality and translation equivariance, which can be beneficial for data and computational efficiency. Representative backbones include *ResNet* [21], *EfficientNet* [54], and *ConvNeXt* [40]. In LVLM pipelines, CNN feature maps are typically pooled or flattened into a sequence of region embeddings before being passed to the multimodal adapter.
- **Vision Transformers (ViTs).** Vision Transformers adapt the Transformer architecture to vision by splitting an image into fixed-size patches and mapping each patch to a token embedding [12]. A stack of self-attention layers then models relationships among patches, enabling long-range interactions and global context integration. The output is naturally a sequence of patch token embeddings, and, in some variants, an additional special `[CLS]` token, which is intended to represent the global semantics of the image, and was originally introduced for image-level classification tasks. ViT-based encoders are prevalent in modern LVLMs because they produce token sequences that align well with the token-based processing used by LLMs, and because large-scale image-text pretraining allows strong visual representations, that can be used in many downstream tasks. Common backbones include *CLIP-ViT* variants [49], *EVA/EVA-02* [14], and *SigLIP*-based encoders [70].

A practical advantage of ViT-based encoders is that they provide an explicit and stable mapping between input patches and output tokens. This property is useful for patch-level methods such as spatial grounding or patch reweighting, whereas CNN-based encoders typically require an additional step to define region tokens from feature maps.

2.2.1.2 Vision-Language Projector

The **vision-language projector** (also called **multimodal projector**) is the component that bridges the vision encoder and the large language model. Visual encoders typically output image embeddings whose dimensionality and distribution differ from the token embeddings expected by the LLM. The projector maps these visual features into the LLM embedding space, producing **visual tokens** that can be concatenated with text tokens and processed by the LLM as part of its context. In practice, the projector is often implemented as a lightweight module such as a linear layer, a small MLP, or an attention-based pooling block that can also reduce the number of visual tokens [66]. Because it is the main alignment interface between modalities, the projector is frequently the target of

parameter-efficient adaptation when modifying how visual information is presented to the LLM, without retraining the full model.

2.2.1.3 Large Language Model

The **Large Language Model** (LLM) constitutes the language-generation core of a LVLM. It is typically a Transformer-based autoregressive model pretrained on large-scale text corpora to predict the next token [59], which provides strong language modeling capabilities and general reasoning skills. Within a LVLM, the LLM operates over a sequence that includes both textual tokens (e.g., instructions or questions) and projected visual tokens obtained from the image through the visual encoder and multimodal projector [36, 33].

Through the self-attention mechanism of the Transformer architecture, the LLM computes contextualized representations by allowing each token to attend to all other tokens in the sequence. When visual tokens are included in this sequence, textual tokens can attend to visual tokens and vice versa, enabling cross-modal interactions. In this way, the LLM can selectively focus on specific visual regions, as represented by visual tokens, while generating language. This attention-based interaction supports tasks such as image captioning, visual question answering, and multi-step visual reasoning, all within the same next-token prediction framework.

Importantly, the attention distributions learned by the model are optimized solely for predictive performance and are not explicitly constrained to reflect human patterns of visual attention. As a result, the regions that receive high model attention during reasoning do not necessarily correspond to the regions that humans fixate when performing the same task. This motivates investigating whether incorporating human gaze into a LVLM, and potentially aligning the model's attention more closely to human attention, can push the model to move beyond surface-level visual cues, supporting more high-level, structured visual reasoning.

2.2.2 Representative Open-Source VLMs

2.2.2.1 LLaVA

One of the most widely used families of Vision-Language Models is the **LLaVA** family [36]. These models represent a canonical implementation of the common VLM architecture composed of a visual encoder, a lightweight multimodal projector, and a large language model. In most variants, the visual backbone is based on CLIP ViT [49], while different versions adopt different language models, including variants from the Vicuna [9] and Mistral [28] families.

Several iterations of LLaVA have been released over time, including LLaVA 1.5 [37], LLaVA-Next (1.6) [38], and LLaVA-OneVision [32]. LLaVA 1.5 mainly consolidates the original architecture through an improved training procedure and refined instruction tuning. LLaVA-Next (1.6) further improves the framework by supporting higher-resolution inputs and more flexible aspect ratios, enabling better handling of detailed visual content. Later versions such as LLaVA-Next and LLaVA-OneVision introduce support for higher-resolution image processing through the AnyRes strategy. This approach dynamically divides the input image into multiple tiles, each of which is independently encoded by the visual backbone. The number of tiles is selected adaptively based on the image resolution and aspect ratio. For instance, LLaVA-Next supports up to five tiles, while LLaVA-OneVision adopts the AnyRes-Max9 strategy, allowing up to nine tiles. For lower-resolution images, tiling may be skipped entirely and the image is processed using the standard single-image encoding used in earlier LLaVA versions.

Among these variants, LLaVA-OneVision-Chat is architecturally more complex than earlier LLaVA models. In addition to the AnyRes-Max9 tiling mechanism, it adopts a larger and more recent visual backbone based on SigLIP [70] rather than CLIP, producing higher-capacity visual representations. It also relies on a more recent language model backbone, such as Qwen2 [64], instead of Vicuna. These changes increase both the capacity and the complexity of the visual representations and of the multimodal alignment process. As a result, the model relies on a richer visual feature space and a more structured integration between visual and linguistic tokens, making its internal representations more sensitive to modifications of the visual embeddings compared to earlier LLaVA architectures.

2.2.2.2 Other VLM Families

Although the LLaVA family represents one of the most widely used open-source architectures for large vision-language models, several alternative design paradigms have been proposed in recent years. Some approaches adopt intermediate modules specifically designed to bridge visual and textual representations, such as the Q-Former architecture introduced in BLIP-2 [33] and later extended in models such as InstructBLIP [10]. Other models rely on different strategies for aligning visual features with language models, for instance through alternative multimodal adapters or modified tokenization schemes, as explored in architectures such as Qwen-VL [4] and related systems.

In this thesis, however, the experiments focus exclusively on LLaVA-style architectures. These models provide a simple and modular structure which makes them particularly suitable for controlled modifications of the visual input representation. This property is especially relevant for the present work, where the goal is to study how gaze-derived signals can be injected into the visual processing pipeline without altering the overall architecture of the model.

2.2.3 Adaptation Strategies for VLMs

Large Vision-Language Models are typically pretrained on large-scale image-text corpora with general objectives. Although this pretraining phase provides broad visual and linguistic knowledge, the resulting models are not explicitly optimized for specific downstream tasks, input formats, or additional modalities. Fine-tuning is therefore required to adapt a pretrained model to a new setting [72]. In this context, adaptation refers to the process of updating part or all of the model parameters so that the model can better align with a target task, dataset, or input distribution, while leveraging the knowledge acquired during pretraining.

Different fine-tuning strategies can be adopted, depending on computational constraints and the size of the model [20]. A straightforward approach consists of full fine-tuning, where all model parameters are updated on downstream data. Although this strategy can yield strong performance, it is computationally expensive and often impractical for very large multimodal architectures. An alternative consists of partial fine-tuning, where only specific components are updated, such as the vision-language projector or selected layers of the language model, while the remaining parameters are kept frozen [36, 33]. This approach reduces memory requirements and training time, while still allowing the model to adapt to new input distributions.

In recent years, methods designed to reduce the number of trainable parameters have gained increasing attention [26, 20]. These approaches introduce a limited set of additional trainable weights while keeping the original pretrained parameters fixed. Among them, Low-Rank Adaptation (**LoRA**) [27] has emerged as a simple and effective technique. LoRA injects trainable low-rank decomposition matrices into selected linear layers of the model, typically within attention and feed-forward modules. Instead of updating the full weight matrices, the method learns low-rank updates that approximate the required adaptation. This significantly reduces the number of trainable parameters and computational cost, while preserving most of the pretrained knowledge encoded in the base model.

Chapter 3

Related Work

This chapter examines the intersection of cognitive science and multimodal modeling. It reviews how human gaze serves as a proxy for visual attention, explores existing architectures that integrate gaze into vision-language systems, and analyzes the evolution of benchmarks from surface-level captioning to structured cognitive reasoning.

3.1 Gaze in Cognitive Science and Scene Understanding

In cognitive science, eye tracking has often been used to study how visual attention is allocated during scene understanding. A classic reference is the work of Yarbus [65], later revisited and discussed in more modern terms by Tatler et al. [55], who showed that observers viewing the same image produce clearly different scanpaths depending on the question they are asked to answer, providing evidence that gaze is strongly shaped by task demands rather than by stimulus properties alone. Along the same line, Henderson [23] reviewed evidence from real-world scene perception showing that gaze control is closely tied to ongoing cognitive processing, object selection, and task requirements, rather than being driven only by low-level salience. Together, these works motivate the use of gaze as a behavioral signal of human attention during scene perception.

Within this perspective, fixations are particularly important. Since detailed visual information is acquired mainly when the eyes fixate on a region [23, 24, 50], whereas visual sensitivity is strongly reduced during saccades, fixation locations are commonly used as a practical proxy for overt visual attention [13]. They indicate which regions were selected for detailed processing, while fixation sequences describe how attention was distributed over time

This interpretation has been widely adopted in computer vision. Early work by Judd et al. [30] used eye-tracking data from free viewing on images to train a supervised saliency model combining low-level, mid-level, and high-level visual features, showing that human fixations can serve as effective supervision for predicting where observers tend to look. *SALICON* [29] later extended this line of work at a much larger scale by introducing saliency annotations for *MS COCO* images [35] through a mouse-based paradigm designed to approximate human attention, making saliency prediction more practical at large scale. In these works, human behavioral data are treated as spatial evidence of attention, and aggregated saliency maps become the target representation for attention modeling.

Gaze has also been studied in more explicitly task-driven forms of scene understanding. Chen et al. [8] introduced *COCO-Search18*, a dataset for goal-directed visual search in which observers search for target object categories in complex scenes. In contrast to saliency datasets collected in free viewing scenarios, COCO-Search18 captures how attention is allocated when perception is guided by a concrete objective, and therefore it provides a useful benchmark for modeling top-down attention control. A related observation emerges in multimodal scene description settings. He et al. [22] collected synchronized eye movements and verbal descriptions for image captioning and showed that gaze during description differs from free-viewing, reflecting the specific demands of selecting scene elements that are relevant for the generated descriptions. Their results further support the idea that gaze is sensitive to the observer's cognitive goal and can reveal which parts of a scene are functionally important to the task being performed.

Overall, prior work in both cognitive science and computer vision supports two main hypotheses. First, **gaze during scene viewing is strongly influenced by the observer's task**, making it informative about task-dependent visual processing. Second, **fixations provide a practical proxy for overt attention** because they mark the regions selected for detailed visual analysis. For these reasons, fixations-based gaze representations, such as scanpaths and heatmaps, provide a natural starting point for modeling human attention during scene understanding.

3.2 Gaze-Guided Vision-Language Architectures

Recent work has explored the use of human gaze in vision-language systems to improve visual grounding and reduce ambiguity in complex scenes. Two main strategies emerge. The first uses gaze as an explicit conditioning signal, namely as additional input that guides visual processing during inference. The second uses gaze or human-attention annotations during training to align model attention with human viewing behaviour.

An early example of the first strategy is the gaze-assisted image captioning model of Sugano and Bulling [53]. Although this work predates current large vision-language models, it is

conceptually important because it shows how gaze can be integrated directly into a vision-language generation pipeline. The authors examined the relation between human gaze, bottom-up saliency, and image recognition features, and proposed a captioning model that incorporates gaze maps into an attention-based architecture. Their experiments on *COCO* and *SALICON* showed that this design improves caption generation, suggesting that gaze can complement the model's learned attention in image captioning tasks.

More recent work has extended this idea to modern vision-language models. For example, Voila-A [63] treats gaze as an explicit modality to align a vision-language model with the user's visual focus. The model is built on a CLIP ViT-L/14 visual encoder [49] and an MPT-7B language model [42], and introduces the *Voila Perceiver Resampler*, a series of cross-attention blocks used to inject gaze information into image features while preserving the pretrained knowledge of the base model. The resulting joint gaze-image embeddings are then passed to the language model to generate the output text.

A complementary line of research uses gaze as supervision during training rather than as input at test time. In visual question answering, Das et al. [11] showed that the attention distributions learned by VQA models are only weakly aligned with human attention, challenging the assumption that neural attention is inherently human-like. This finding motivated later work that treated human attention not simply as an analysis tool, but as an explicit supervisory signal. Following this direction, Qiao et al. [48] proposed supervising VQA attention maps with human-like attention annotations. Because manually collected human attention data are limited, they first trained a *Human Attention Network* on the *VQA-HAT* dataset [11] and then used it to generate attention maps for image-question pairs in *VQA v2.0* [19]. These generated maps were used as explicit supervision for an attention-based VQA model. Their results showed that human-like attention supervision improves answer prediction, supporting the idea that external attention guidance can facilitate multimodal reasoning.

More recently, this training-time perspective has been revisited in the context of large vision-language models. Gaze-VLM [45], for instance, proposes a gaze-regularized attention mechanism for egocentric video understanding. Gaze-VLM uses gaze only during training to regularize the model's internal attention: the method is modular and can be applied to different attention-based vision-language architectures, and it is evaluated on tasks such as future event prediction and current activity understanding. This work shows that gaze can also serve as a purely supervisory signal, helping the model attend to regions that are more consistent with human visual behavior even when gaze is not available at inference time.

Overall, these studies show that gaze can support vision-language systems in two complementary ways: as **an additional modality that directly guides visual grounding**, or as **a supervisory signal that encourages model attention to better reflect human viewing patterns**. This distinction is particularly relevant for the present thesis, since

the proposed scenarios include both direct gaze injection into the visual representation and fine-tuning strategies for gaze-informed processing.

3.3 Image Captioning and Visual Reasoning Benchmarks

3.3.1 Limits of Classical Captioning Metrics

Traditional image captioning benchmarks, such as MS COCO [35] and Flickr30k [67], evaluate model outputs using automatic metrics including BLEU [46], METEOR [6], ROUGE [34], and CIDEr [61]. These metrics measure lexical overlap between a generated caption and one or more reference captions. While they provide a convenient and reproducible evaluation protocol, they mainly reward surface-level similarity and correct object mentions. As a result, a model can obtain competitive scores by listing visible entities and simple actions, even if it fails to capture deeper semantic aspects such as event structure, causal relations, or mental states. Moreover, these metrics are insensitive to whether high-level reasoning components are explicitly expressed. Consequently, classical captioning evaluation does not offer a reliable assessment of structured visual understanding or cognitive reasoning.

3.3.2 Visual Reasoning Datasets

To move beyond surface-level caption evaluation, several benchmarks have been proposed to assess visual reasoning more directly. For example, Visual Commonsense Reasoning [69] introduces question answering with justification, encouraging commonsense inference. VisualCOMET [47] focuses on predicting past events, future events, and character intents. Other benchmarks, such as MMBench [39] and MME [15], evaluate multiple capabilities through structured question answering. Although these datasets incorporate reasoning components, they typically assess specific and isolated skills, often in a multiple-choice format. In addition, many images used in existing benchmarks do not require rich narrative interpretation. Therefore, current visual reasoning datasets still lack a comprehensive and fine-grained evaluation of high-level cognitive abilities in image understanding.

3.3.3 CogBench: Cognitive Evaluation Benchmark

To address these limitations, Song et al. [51] proposed **CogBench**, an evaluation benchmark specifically designed to evaluate the high-level cognitive understanding of large vision-

language models rather than only object recognition or isolated reasoning skills.

To ground this objective in established cognitive science practice, CogBench draws inspiration from the **Cookie Theft** picture (Figure 3.1) description task, a widely used clinical instrument for language and cognitive function screening (originally part of the Boston Diagnostic Aphasia Examination [16]). In the Cookie Theft task, participants are asked to freely describe a single, semantically dense image: the quality of the description reflects cognitive status. Individuals with intact cognition tend to reason on high-level events (e.g., interpreting an action as *stealing* rather than merely *taking*), characterize social roles and relationships (e.g., *mother* rather than *lady*), attribute mental states (e.g., *preoccupied*, *happy*), and connect events through causality (e.g., children steal cookies *because* the mother is distracted). In contrast, cognitively impaired individuals more often remain at the level of isolated observations and omit or weaken these higher-level inferences. This contrast provides a clear conceptual template: an image that supports multiple, interacting inferences can serve as a sensitive probe of cognition through description.

CogBench translates this idea into an LVLM benchmark by focusing on Cookie Theft-like images: scenes with an identifiable story theme, multiple entities and interactions, and several plausible but constrained chains of reasoning that connect multiple, small, low-level cues to broader, high-level conclusions. Building on this principle, CogBench defines a set of reasoning dimensions, called **Chains-of-Reasoning** (**CoRs**), and uses them to structure annotations and evaluation.

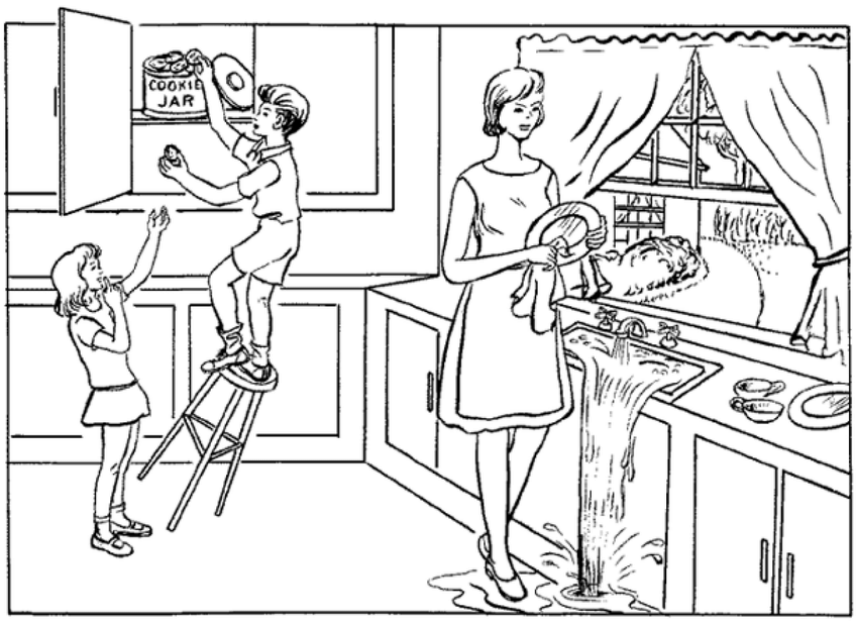


Figure 3.1: The "Cookie Theft" picture, often used in human cognitive tasks by linguists and psychologists, inspired the creation of the CogBench benchmark.

The resulting benchmark is explicitly designed to test whether a model's outputs go beyond enumerating visible elements and instead reconstruct a coherent interpretation of the scene by making and justifying the kinds of inferences that humans naturally produce when describing such images.

3.4 Research Gap and Thesis Positioning

The literature reviewed in this chapter shows that gaze is meaningful to scene understanding and that it can be incorporated into vision-language systems either as an additional input or as supervision signal for model attention. At the same time, recent work has highlighted the limits of standard captioning metrics for assessing higher-level visual understanding, motivating the development of novel evaluation settings such as CogBench.

Despite this progress, it is still unclear whether task-specific human gaze can support the structured scene interpretation targeted by CogBench. In particular, prior work does not clearly establish whether gaze mainly improves the recognition of relevant scene elements or can also contribute to higher-level reasoning about events, relationships, and other semantic aspects of an image.

This thesis addresses this gap by extending the CogBench approach with human gaze data collected on the benchmark's images and aligned with its task structure. It investigates three gaze integration strategies for open-source vision-language models, designed to reflect progressively more invasive forms of architectural intervention, in order to examine how each specific approach for gaze injection influences the corresponding model's behavior. In this framework, gaze is exploited both as an explicit third modality and as a signal for aligning visual processing with human attention. The objective is to assess, through the Recognition and Cognition scores of CogBench, whether and under which conditions gaze can improve both scene recognition and higher-level visual reasoning.

Chapter 4

Methods

In this chapter, we provide an overview of the CogBench evaluation framework, which is central to our experiments and analysis. We then describe the eye-tracking study we conducted to acquire human gaze data during spoken image descriptions. While both gaze and audio data modalities were collected, the experiments presented in this thesis only use the gaze data. The spoken descriptions are rather used to align gaze behavior to the intended reasoning dimension, and incorporated in our dataset for potential use in future work.

4.1 CogBench: Dataset, Task and Metrics

4.1.1 Dataset Creation

CogBench evaluation relies on a dataset of images that depict rich, story-like scenes. Each image is paired with structured annotations that separate low-level visual recognition from higher-level reasoning, enabling a more fine-grained assessment of model outputs. In particular, for each image the dataset provides an entity list, a set of Chains-of-Reasoning covering multiple reasoning categories, and a reference description that summarizes the story expressed by the image.

4.1.1.1 Chains-of-Reasoning (CoRs) Definition

A Chain-of-Reasoning in CogBench is a structured representation of how higher-level conclusions can be inferred from multiple visual cues. Each CoR is written as a set of premises

leading to a conclusion, using a form similar to:

$$A_1 + A_2 + \cdots + A_n \rightarrow B$$

where the A_i terms are observable clues in the image and B is an inferred statement that captures higher-level semantics, such as an event interpretation or an implied relationship. CoRs are used to represent both inferences about single concepts (for example, inferring a season from environmental cues) and relations between events (for example, capturing a cause-effect link).

CogBench defines eight reasoning dimensions, which correspond to the following types of CoR annotations:

- **Special Time Reasoning**: inference about a meaningful time context of the story, such as seasons or special occasions, when it is relevant to understanding the scene.
- **Location Reasoning**: inference about the setting of the story, when it is not directly stated but can be derived from contextual cues.
- **Character Reasoning**: inference about the roles or identities of people in the scene (for example, occupational or social roles) using visual evidence.
- **Character Relationship Reasoning**: inference about relationships between characters (for example, family relations) based on how they interact and appear in context.
- **Event Reasoning**: inference about high-level events in the current or immediately preceding moment, where the conclusion carries more semantic meaning than a direct description of actions.
- **Event Relationship Reasoning**: reasoning about causal or temporal links between events, requiring that both events and their relationship are expressed.
- **Next Moment Event Reasoning**: inference about what is most likely to happen next, restricted to events that have a high probability given the scene cues.
- **Mental State Reasoning**: inference about emotions, intentions, or other mental states of characters, derived from expressions, posture, and context.

This CoR structure is central to CogBench because it operationalizes "cognition" as the ability to express the intended high-level semantics of the story, going beyond the mere description of visible elements and requiring the model to make deductions and infer reasoning about the scene.

4.1.1.2 Image Collection

The CogBench dataset contains **251 Cookie Theft-like images**, manually collected from the web, and it also includes the original Cookie Theft picture. The selected images are designed to be similar in complexity: each image depicts an interesting story and supports multiple Chains of Reasoning that connect low-level visual cues to higher-level inferences or conclusions about the scene. At the same time, the images follow a restricted content complexity criterion, meaning that they contain sufficiently rich content to support the story, while keeping the number of entities limited enough to emphasize the key points and avoid overly chaotic scenes. Figure 4.1 presents several example images from the CogBench dataset.

4.1.1.3 Image Annotation

Each image is annotated by human annotators using a structured protocol that produces three annotation components in sequence: `[Entities]`, `[CoRs]`, and a final `[Description]`.

Entity annotation requires listing as many clearly identifiable entities as possible (such as people, animals, objects), while omitting entities that are too uncertain to recognize reliably. CoR annotations in turn capture the reasoning processes that connect observations to conclusions under the predefined reasoning categories. Finally, annotators write a reference description that conveys the whole story, guided by the previously listed entities and CoRs.

To reduce subjectivity, three independent annotations are collected for each image and then merged. The merging procedure is based on majority voting, primarily retaining entities and CoRs that appear in at least two annotations, while allowing additional elements when they are judged reasonable. Images with large disagreement across annotators were discarded, to avoid cases where the story interpretation is not sufficiently consistent.

4.1.2 Image Description Task

The evaluation framework proposed by CogBench consists of both an image description task and a visual question answering task, designed to assess different aspects of visual reasoning.

In this work, **we focus exclusively on the image description task**, which is the central component of the benchmark. It evaluates two complementary abilities: the recognition of relevant entities in an image and the ability to derive structured, high-level interpretations from visual evidence.

Figure 4.1: Example images from the CogBench dataset. The images exhibit rich storytelling and multiple Chains of Reasoning, while maintaining a restricted level of content complexity.

Given an image, the model is required to produce a free-form textual description. The objective goes beyond enumerating visible objects or actions. The generated description should convey the overall story represented in the scene, integrating perceptual elements into coherent reasoning patterns. In particular, the model is expected to reflect the types of Chains of Reasoning defined in CogBench, moving from surface-level observations to higher-level conclusions about the situation depicted.

Two evaluation modes are defined:

- **Spontaneous Description**. The model receives a general instruction (i.e. "Describe this image in detail."). This setting evaluates whether the model can autonomously structure a description that includes both recognition and reasoning, without explicit guidance.
- **Directed Reasoning**. The prompt explicitly requires the inclusion of reasoning aspects aligned with the CoR types defined in CogBench. This mode evaluates whether the model can correctly address these reasoning patterns when they are directly specified in the instruction.

The comparison between the two modes provides insight into whether performance differences arise from limitations in reasoning capability or from difficulties in organizing the output without structured guidance.

4.1.3 Evaluation Metrics

CogBench defines two complementary metrics for evaluating the image description task, the Recognition Score and the Cognition Score, which metrics separately quantify entity recognition and high-level reasoning. Both are computed by comparing the generated descriptions with the ground truth annotations provided in the benchmark.

4.1.3.1 Recognition Score

The **Recognition Score** measures the model's ability to mention the entities present in the image. It is defined as the recall of annotated entities, computed as the ratio between the number of recognized entities and the total number of annotated ones.

To compute the number of recognized entities for each image, the model-generated description is processed using the *spaCy* library[1] to extract nouns. Both the extracted nouns and the annotated entities provided in CogBench are embedded using *sentence-transformers*[2]. For each annotated entity, the cosine similarity with all extracted nouns is computed. An entity is considered recognized if at least one similarity value exceeds a threshold of 0.6.

This semantic matching approach allows for flexible comparison beyond exact lexical overlap. The final Recognition Score is obtained by aggregating the recognized entities over all images and dividing by the total number of annotated entities.

4.1.3.2 Cognition Score

The **Cognition Score** evaluates whether the generated description captures the reasoning patterns annotated in the benchmark. Each image is associated with multiple CoRs, corresponding to the CoR types defined in CogBench.

For each annotated CoR, the evaluation verifies whether the semantics of its conclusion are explicitly or implicitly present in the generated description. In the case of CoRs that encode relationships between events, the evaluation additionally requires that the corresponding causal or temporal link be clearly expressed.

GPT-4 [1] is used as an automated evaluator to assign a binary value to each CoR, indicating whether its semantics are reflected in the model's description. The Cognition Score is then computed as a recall measure over all annotated CoRs, both at the level of individual reasoning types and overall. The decision to rely on GPT-4 for evaluating the model outputs was based on an empirical comparison with other cognition evaluation

[1] https://spacy.io
[2] https://www.sbert.net

methods. In this comparison, human judgments were obtained by manually scoring the CoRs of 20 images. Among the tested approaches, GPT-4 achieved the highest evaluation accuracy with respect to these human annotations, indicating a stronger alignment with human assessment than traditional metrics. Although GPT-4 is a probabilistic model and may produce slightly different evaluations across multiple runs on the same descriptions, it provides the most reliable approximation of human evaluation among the considered methods.

Together, the Recognition and Cognition Scores provide a structured assessment framework that distinguishes between the ability to identify visual elements and the ability to organize them into coherent, high-level reasoning consistent with the annotation schema of CogBench.

4.2 Gaze Data Collection

In this section, we describe the eye-tracking study conducted to acquire the gaze data used in this project, together with aligned verbal descriptions. The study was conducted at the Italian Institute of Technology (IIT) in Genoa. The data collection sessions were carried out by the research team, including the author of this thesis, who contributed to the setup of the experimental environment, the initiation and supervision of recording sessions, the resolution of technical issues during gaze acquisition, and final data exportation. The experimental protocol was approved by the Local Ethical Committee (Comitato Etico Regione Liguria), and written informed consent was obtained from all participants prior to their participation in the study.

The goal of the experiment was to enrich the CogBench image dataset with human gaze recordings collected during image description tasks. The study was designed to elicit descriptions that are comparable to the original CogBench annotations: participants were asked to describe each image following the same types of prompts used in CogBench, while their gaze behavior was recorded and their spoken responses were simultaneously captured. Thirty participants took part in the study (mean age: 27.7 ± 3.6; female-to-male ratio: 62%).

4.2.1 Experimental Apparatus

Data were acquired using the Tobii Pro Spectrum[3] eye tracker, while Tobii Pro Lab[4] was used to design the experimental procedure and to record, synchronize, and export the

[3] `https://www.tobii.com/products/eye-trackers/screen-based/tobii-pro-spectrum`

[4] `https://www.tobii.com/products/software/behavior-research-software/tobii-pro-lab`

resulting gaze (and audio) data.

4.2.1.1 Tobii Pro Spectrum

The **Tobii Pro Spectrum** is one of the most widely used eye-tracking systems in academic research. It is a screen based eye tracker equipped with two infrared eye-tracking cameras housed in a separate sensor array positioned below the display. This remote dual camera setup captures stereo images of both eyes, enabling accurate estimation of eye position and movement while remaining robust to moderate head movements. The system supports high temporal resolution, which is important for capturing rapid eye movements as well as stable fixations, and it provides gaze data that can be mapped in a reliable way to screen coordinates for stimulus driven studies. In our setup, the integrated 23.8" display (*EIZO FlexScan EV2451*) has a resolution of 1920 $\times$ 1080 pixels, and gaze was recorded at a sampling rate of 1200 Hz. These characteristics make the Tobii Pro Spectrum suitable for experiments on free viewing of complex images, where both spatial reliability and temporal fidelity are necessary to relate gaze behavior to visual regions of interest.

4.2.1.2 Tobii Pro Lab

Tobii Pro Lab is a dedicated experimental software environment designed to support the design, execution, and analysis of eye tracking studies. It provides an integrated framework for stimulus presentation, participant management, and calibration and validation procedures, and synchronized recording of gaze data. The software allows researchers to define experimental timelines, present static or dynamic visual stimuli, and associate gaze recordings with precise temporal and spatial references. In addition, it supports simultaneous audio recording during stimulus presentation, with the audio stream time-aligned to the gaze recording, which enables reliable synchronization between spoken responses and visual attention.

Tobii Pro Lab also includes built-in tools for processing raw gaze samples into higher-level eye movement events through configurable Gaze Filter presets. In particular, it implements the *Tobii I-VT* (Velocity-Threshold Identification) gaze filter [57], which assigns angular velocity to gaze samples and classifies them as fixations or saccades by applying a velocity threshold. Finally, the software provides tools for visualizing gaze behavior through scanpaths and heatmaps, and standardized export functions in various coordinate systems.

In this context, gaze positions can be expressed either relative to the display surface or relative to a specific stimulus: the Display Area Coordinate System (DACS) represents gaze locations with respect to the active display area, while the Media Coordinate System (MCS) represents gaze locations in the coordinate frame of the presented media (e.g., an image),

enabling direct interpretation in stimulus space. The distinction between DACS and MCS, illustrated in Figure 4.2, is important when processing the raw gaze data recorded, since the conversion of screen coordinates to stimulus coordinates depends on the size and position of the stimulus on the display.

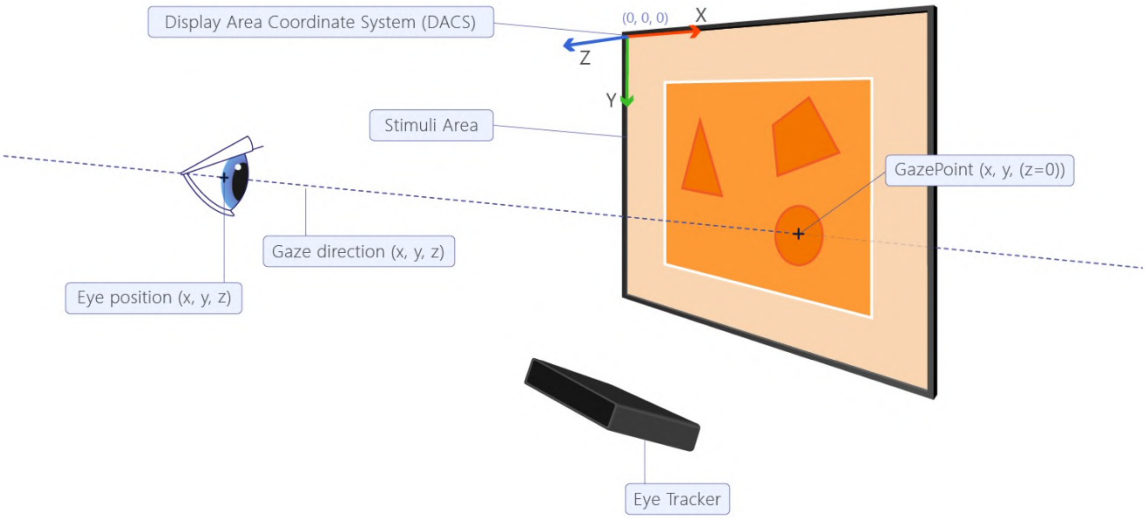
4.2.2 Experimental Environment

The Tobii Pro Spectrum eye tracker was used for this gaze data collection study together with the supplied monitor. Spoken descriptions were recorded with a USB microphone via Tobii Pro Lab. The system was placed on a desk, and participants sat in front of it at an approximate viewing distance of 55-70 cm, as recommended in the user manual [58]. A stable chair without wheels was used to minimize unintended movements and help maintain a consistent seating position. Participants were instructed to remain as still as possible during the recording sessions; nevertheless, the Tobii Pro Spectrum is designed to be robust to moderate head movements, allowing for natural viewing behavior without the need for head stabilization. To reduce interference from external light, window blinds were kept closed during all recording sessions, and the room was illuminated using artificial lighting. Figure 4.3 shows the experimental setup.

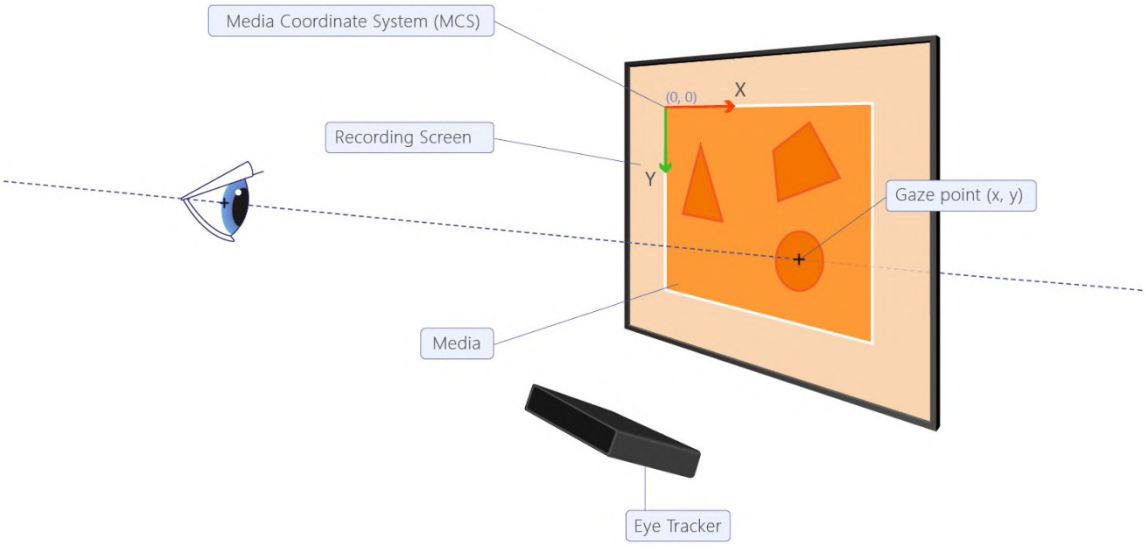
4.2.3 Stimuli and Task Design

We used Tobii Pro Lab to design the stimulus timelines for the experiment. Since CogBench contains 251 images and we aimed to present each image to three different participants, we created 10 timelines: nine containing 25 images each and one containing 26 images. This design allowed us to cover the entire CogBench dataset exactly three times with 30 participants, with each participant viewing a total of 25 images (or 26 for participants assigned to the final timeline).

During data acquisition, participants performed a guided image description task designed to align with CogBench's annotation scheme. In particular, the task covered both entity identification and the full set of CogBench Chain-of-Reasoning (CoR) categories. The instruction prompts shown to participants were exactly the same as those originally provided to the CogBench annotators. Each image in a timeline was presented nine times, once for each description goal; at each presentation, participants were asked to verbally describe the image by answering the current prompt, corresponding to the target entity list or CoR type. Participants could decide when to proceed to the next step of the timeline by pressing the space bar on a keyboard.



(a) Display Area Coordinate System (DACS).



(b) Media Coordinate System (MCS).

Figure 4.2: Illustration of the coordinate systems supported by Tobii Pro Lab. In DACS, gaze positions are referenced to the active display area, while in MCS they are referenced to the stimulus media, enabling gaze interpretation directly in stimulus space. Images adapted from Tobii documentation [56].

Figure 4.3: Close-up and participant's viewing position of the experimental setup used for the data collection study. The participant is seated in front of the Tobii Pro Spectrum eye tracker and the companion monitor, which displays the stimulus images while gaze data and audio descriptions through a microphone are recorded. The keyboard is used to progress through the experiment timeline.

4.2.4 Calibration and Validation

Before the start of each recording session, a calibration procedure was performed to adapt the gaze estimation model to the current participant. A five-point calibration was adopted, in which an animated target was sequentially presented at five predefined screen locations. Participants were asked to fixate the target at each position until it moved to the next point. During this process, the eye tracker collected gaze samples associated with known screen coordinates and used them to compute the participant-specific calibration model.

Following calibration, a four-point validation procedure was conducted to assess the quality of the estimated gaze mapping, using the same procedure as calibration. The validation results were inspected immediately after completion; if one or more validation points showed clear deviations or unstable measurements, the calibration procedure was repeated to ensure reliable gaze data acquisition. This was necessary only a few times, likely due to frequent blinking or larger head movements during either calibration or validation.

4.2.5 Recorded and Exported Data

The raw gaze data were processed and organized in Tobii Pro Lab, which allows exporting recordings in tabular format. For each participant, the exported tables contain one row per sample timestamp; each row reports the current media, gaze coordinates, and additional metadata. From the available fields, we exported the following:

- **Recording Timestamp**: time expressed in ms;
- **Presented Media Name**: the filename of the CogBench image displayed;
- **Fixation Point X** and **Fixation Point Y**: the fixation coordinates expressed in the MCS reference frame;
- **Original Media Height** and **Original Media Width**: the native dimensions of the CogBench image, independent of any scaling applied during stimulus presentation;
- **Eye Movement Type**: the type of the gaze event, as classified by Tobii I-VT gaze filter (e.g., fixation or saccade);
- **Gaze Event Duration**: the duration associated with the classified event.

Gaze samples recorded while instruction prompts were displayed on the screen (i.e., when no CogBench image was shown) were discarded.

Audio descriptions were exported separately from the gaze data. Since audio and gaze recording started simultaneously and remained time-aligned throughout the session, the gaze timestamps can be used to segment the audio into intervals corresponding to each stimulus presentation.

4.2.6 Data Quality Control

Within the project page in Tobii Pro Lab, for each participant a success score is reported, indicating the percentage of successfully recorded eye-position samples out of the maximum 1200 samples per second provided by the 1200 Hz sampling rate. The average success rate across participants was 90%.

Two recording sessions were excluded and subsequently repeated with different participants because the tracking quality reported by Tobii Pro Lab was substantially lower than the overall average. This reduction in tracking performance was likely due to suboptimal acquisition conditions, such as temporary issues with the eye tracker or uncontrolled ambient lighting (e.g., partially open blinds), both of which can negatively affect tracking robustness and accuracy.

4.3 Gaze Data Processing

4.3.1 Fixation Filtering

The exported data were first filtered based on the gaze event labels assigned by the Tobii I-VT filter, retaining only samples classified as fixations. Since our objective was to analyze which regions of the image were visually salient from a spatial perspective, we focused exclusively on fixations, as they are commonly considered a reliable proxy for human visual attention [71, 55].

As an example, Figure 4.4 shows the complete set of raw gaze samples exported from Tobii Pro Lab and the subset of data retained after filtering for fixations only.

The exported fixation coordinates were expressed in the Media Coordinate System, with values in $[0, 1]$. Therefore, mapping them to pixel coordinates simply requires multiplying the normalized horizontal coordinate by the pixel width of the presented stimulus and the normalized vertical coordinate by its pixel height.

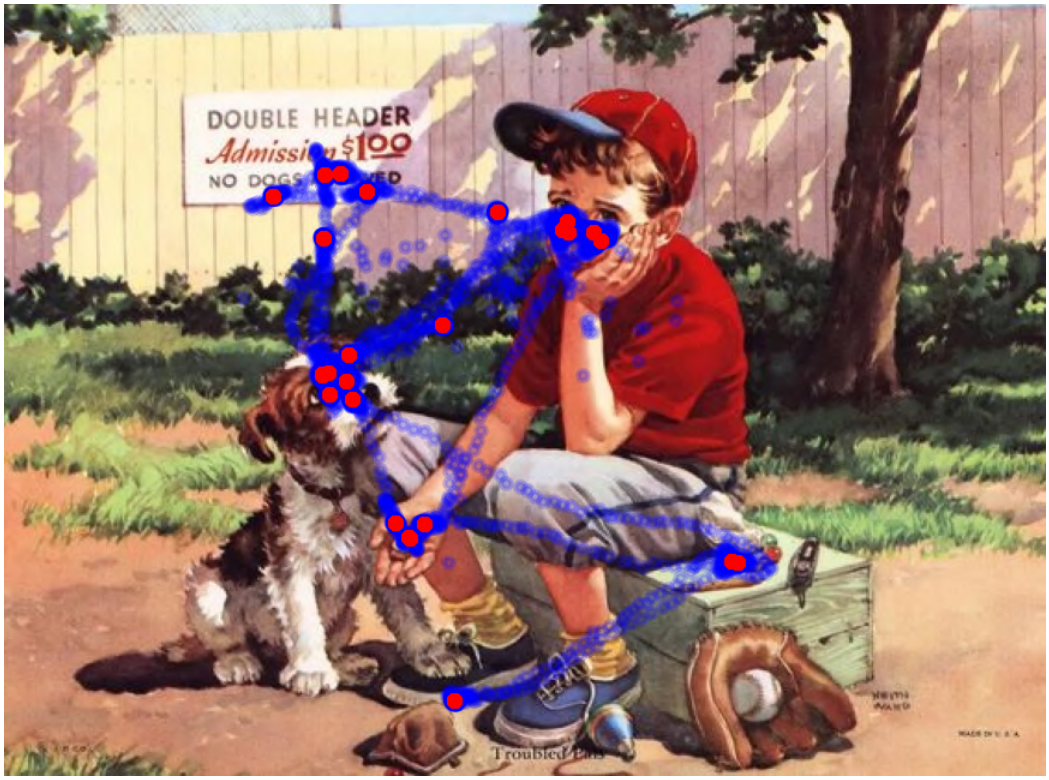


Figure 4.4: All gaze events recorded and classified by the eye tracker during an example recording of the Event Reasoning task. Saccades are shown in blue, forming the scanpath between gaze points, while fixation locations are marked in red.

4.3.2 Heatmap Generation

Since the objective was to model the spatial information conveyed by fixation data, **gaze was represented as a saliency map**, which is a common representation in both psychological and computer vision research. This representation allows fixation locations to be projected onto the image plane, producing a continuous spatial distribution that reflects where observers allocated their attention, also called heatmap.

When generating saliency maps, each fixation was treated equally, independently of its duration. In other words, fixation duration was not used to modulate the contribution of individual gaze points. This decision is consistent with prior studies that aim to model spatial attention patterns only, and therefore intentionally exclude temporal aspects of gaze behavior from their analysis [30, 13, 22]. Additionally, as discussed in Section 2.1.1, longer fixation durations do not necessarily indicate that the attended region is more important. For this reason, the resulting heatmaps encode only *where* participants looked, without modeling *how long* they fixated on specific regions.

To construct a heatmap from fixation coordinates, each fixation point was convolved with a two-dimensional Gaussian kernel, following standard practice to account for spatial uncertainty in gaze localization. A common choice for the standard deviation σ of the Gaussian kernel, used in both cognitive science and computational saliency modeling, is to set σ to approximately 1° of visual angle relative to the displayed stimulus.

This choice is motivated by several factors. First, the exact gaze location is not a pixel-precise point but rather an estimate subject to measurement noise. Second, the human fovea covers approximately 1-2° of the central visual field, meaning that visual processing during a fixation extends over a small but non-negligible spatial region. Finally, eye trackers introduce intrinsic, albeit small, localization errors. Using a Gaussian with a standard deviation corresponding to about 1° of visual angle therefore distributes each fixation over a region that more realistically approximates the area likely attended during that fixation.

For this reason, the appropriate value for σ was computed based on the specific characteristics of our experimental setup. Participants were seated at an average distance of approximately $D = 63$ cm, which lies within the recommended range (55-70 cm) for the Tobii Pro Spectrum. The display used for stimulus presentation had a physical width of $W = 52.8$ cm and a horizontal resolution $R = 1920$ pixels.

Given the geometry of the setup, the standard deviation corresponding to 1° of visual angle, expressed in pixels, was computed as:

$$\sigma = \frac{R}{W} \cdot 2D \cdot \tan\left(\frac{\theta}{2}\right) \tag{4.1}$$

where $\theta = 1°$.

Substituting the values of our setup yields approximately $\sigma \approx 40$ pixels.

In all heatmaps, clamping is applied to values below the 5% percentile and above the 98% percentile in order to improve the stability of the gaze signal distribution and to mitigate the effect of extreme peaks or outliers. This operation reduces the influence of unusually high or low fixation values that may arise from noise or from individual variability across participants.

After clamping, normalization is also applied to the heatmaps. However, different normalization strategies are adopted depending on the specific implementation scenario. The details of these strategies, and their role within each integration approach, are described in Section 5.4.

4.3.3 Aggregation Across Participants

The goal of collecting gaze data on the images from the CogBench dataset was to enrich it with general information about where a human observer would look in order to answer questions requiring specific reasoning about the image content. Since each image was presented to three different participants, for each (image, reasoning type) and (image, entity list) pairs we obtained three distinct heatmaps, one per participant.

To account for inter-subject variability in visual exploration strategies, we aggregated the three corresponding heatmaps by averaging them. Humans do not move their eyes identically nor attend to exactly the same spatial locations when performing a visual task; averaging therefore provides a more stable representation of the spatial patterns that are consistently relevant across observers, and it is a standard technique adopted in similar works in the field.

This choice is also consistent with the construction of the CogBench textual annotations, which are derived through a majority voting procedure among three annotators. By averaging the heatmaps, we obtain a representative estimate of where a human would typically look in an image to answer a given reasoning question, aligning the gaze aggregation strategy with the annotation protocol adopted in the benchmark.

As a result, for each image we computed a total of nine average heatmaps: one corresponding to the entity listing task and one for each of the eight reasoning types. Figure 4.5 shows all the resulting heatmaps for an example image, illustrating how the relevance of image regions varies depending on the specific visual reasoning task performed by participants.

Table 4.1 reports summary statistics of the collected gaze data. For each task, it shows the total number of fixations acquired, the mean number of fixations contributing to a single aggregated heatmap, the mean trial duration, and the mean entropy of an aggregated heatmap. Entropy provides an indication of how fixations are distributed across the image. The Entity Recognition task yields substantially more fixations than any reasoning task. This is likely due both to its position as the first task performed on each image, when participants were seeing the scene for the first time, and to its more exploratory nature, since participants were asked to list as many entities as possible, including elements not central to the main story. The same exploratory pattern may also account for the slightly higher mean entropy observed for this task, suggesting a broader spatial distribution of fixations, as well as for its longer mean trial duration.

Task	Total Fixations	Mean Fixation Count	Mean Duration (s)	Mean Entropy
Entities Recognition	52260	69.40 ± 35.28	17.89 ± 9.64	12.05 ± 0.37
Special Time Reasoning	26837	35.64 ± 29.28	8.85 ± 7.44	11.87 ± 0.38
Location Reasoning	27756	36.86 ± 24.46	9.28 ± 6.18	11.88 ± 0.38
Character Reasoning	26402	35.06 ± 29.16	9.13 ± 7.45	11.68 ± 0.36
Character Relationship Reasoning	29597	39.31 ± 26.07	10.57 ± 7.28	11.66 ± 0.34
Event Reasoning	42967	57.06 ± 29.98	15.56 ± 8.40	11.76 ± 0.36
Event Relationship Reasoning	38182	50.71 ± 32.63	13.67 ± 8.59	11.76 ± 0.36
Next Moment Event Reasoning	23062	30.63 ± 20.55	8.09 ± 5.35	11.70 ± 0.37
Mental State Reasoning	33450	44.42 ± 27.40	13.85 ± 8.74	11.61 ± 0.34

Table 4.1: Summary statistics of the collected gaze data by task category. Compared with the reasoning tasks, Entity Recognition shows a higher number of fixations, longer trial duration, and slightly higher heatmap entropy, suggesting a broader and more exploratory allocation of gaze over the image.

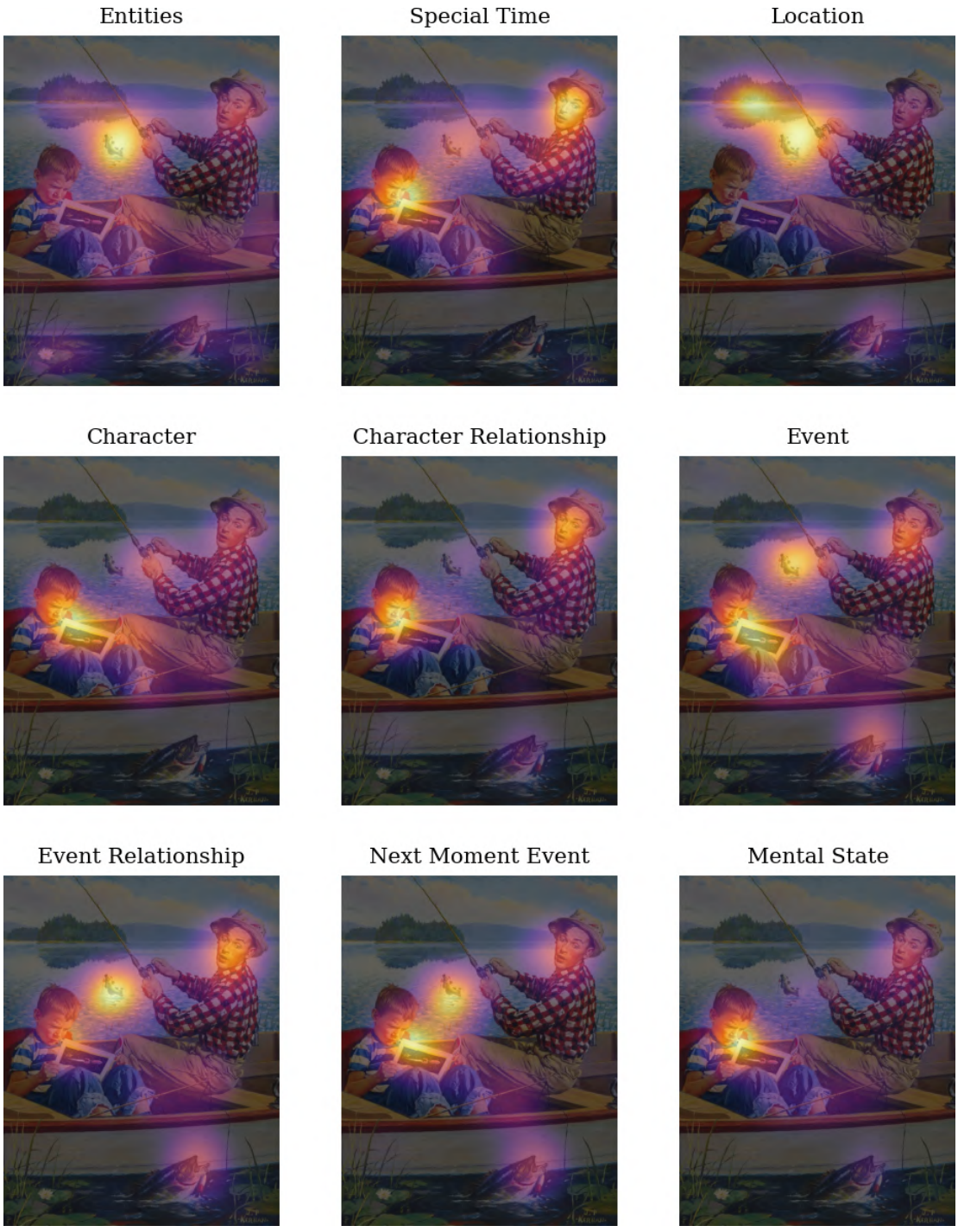


Figure 4.5: Averaged gaze heatmap overlays for an example image. Each map corresponds to one task: entity recognition or one of the eight reasoning categories defined in CogBench, and shows the spatial distribution of gaze fixations aggregated across participants.

Chapter 5

Experiments

In this chapter, the experimental design and evaluation protocol of the proposed approaches is presented. The main objective is to investigate whether providing human gaze information to open-source Vision-Language Models affects their performance on the Recognition and Cognition metrics defined by the CogBench benchmark [51]. In particular, the experiments aim to determine **whether gaze can support models in generating more accurate entity recognition and more human-like high-level reasoning** about image content.

To this end, three architectural scenarios were designed, each corresponding to a different strategy for integrating gaze heatmaps into the visual processing pipeline. These scenarios range from direct modulation of visual features using gaze as a weighting mechanism, to fine-tuned adaptations of the multimodal projector and visual encoder in order to better align model representations with gaze-derived information.

All implemented architectures were evaluated using an adapted version of the CogBench benchmark, enabling a systematic analysis of how different gaze injection strategies influence visual understanding and reasoning performance.

5.1 Evaluation Protocol

5.1.1 Re-Definition of the CogBench Description Task

CogBench evaluates models in the Description task by prompting them to produce a single, comprehensive description of the input image, either in the Spontaneous Description mode or in the Directed Reasoning mode. Both the Recognition and the Cognition scores are computed from this single generated description, by assessing whether annotated entities

and Chains of Reasoning (CoRs) are mentioned in it (Section 4.1.3.1 and Section 4.1.3.2). This evaluation protocol, however, is not directly applicable in our setting. In our case, the dataset is structured as reasoning-task-specific (image, heatmap) pairs. For each image, we do not have a single global gaze heatmap, but rather multiple heatmaps, each associated with a specific reasoning dimension (entities or one of the eight CoR types). Consequently, prompting the model to produce a single, general description of the image would break the alignment between the reasoning task and the corresponding human gaze signal.

For this reason, we reformulate the Description task as follows. Instead of generating one overall description, **the model is prompted separately for each reasoning type**. At inference time, each (image, heatmap) pair is provided together with a CoR-specific prompt aligned with the reasoning dimension that the heatmap refers to. This design ensures explicit alignment between the human attention signal (encoded in the heatmap) and the reasoning task that the model is required to perform. As a consequence, for each image the model produces nine outputs: one for the entity listing task and one for each of the eight reasoning types. The Recognition and Cognition scores are then computed using only the task-specific output corresponding to each reasoning dimension, rather than extracting all information from a single general description.

Despite this modification in the evaluation procedure, **the semantic interpretation of the two metrics remains unchanged**. The Recognition score still measures the model's ability to correctly identify entities present in the image, while the Cognition score still evaluates its ability to perform high-level reasoning beyond surface-level visual content. What changes is only the operational protocol used to obtain and assess such capabilities, which is adapted to ensure consistency with the gaze-conditioned experimental setup.

Figure 5.1 shows the exact prompts used for each reasoning task at inference time. These prompts replace the Spontaneous Description and Directed Reasoning prompts adopted in CogBench. They were custom-designed starting from the instructions originally provided to the human annotators in CogBench and later adapted for the participants in our eye-tracking study, conducted to gather gaze data during the image description and reasoning task.

5.1.2 Recognition Score

The Recognition score is computed using only the model output generated for the entity listing task. Following the procedure defined in CogBench, unique entities are extracted from the generated text, and the score is calculated as the ratio between the correctly recognized entities and the annotated entities in CogBench.

Shared instruction for reasoning prompts containing examples:
Examples mentioned above are illustrative only and should not be treated as categories that must be checked.

Entity Recognition List the entities appearing in the picture, including people, animals, and objects. List only entities that are clearly visible. Mention each entity at most once (no duplicates). Stop as soon as you have covered all salient entities. Do not guess.

Special Time Reasoning Write your reasoning about the special time context of the story depicted in the picture, for example festivals, seasons, or particular times of day. The special time is relevant only if it requires reasoning beyond what is immediately obvious.

Location Reasoning Write your reasoning about the location of the story depicted in the picture, for example near a school, or at home.

Character Reasoning Write your reasoning about the characters in the picture, including their possible roles or identities, for example a teacher, a doctor, or a student.

Character Relationship Reasoning Write your reasoning about the relationships between the characters in the picture, for example a mother-child relationship or friendship.

Event Reasoning Write your reasoning about the events in the current and previous moments of the picture based on the clues provided. You only need to annotate high-level events and can ignore low-level ones. For example, "the woman is looking at the man" is a low-level action. A reasoning process like "The boy is stealing cookies, since he is fetching them behind the mom, while she is busy cooking, and the girl is shushing him." describes a high-level event.

Event Relationship Reasoning Write your reasoning about the relationships between different events in the picture. These relationships are typically causal or temporal (e.g., one event causes or precedes another).

Next Moment Event Reasoning Write your reasoning about the events that are most likely to happen in the next moment. Only include events that have a very high probability of occurring based on the current visual scene.

Mental State Reasoning Write your reasoning about the mental states of the subjects in the picture, for example being happy, worried, or daydreaming. You need to reason as best you can about the mental states of all the subjects in the picture. Only omit the subjects that are not showing obvious emotions.

Figure 5.1: Task-specific prompts used at inference time. For each image, the model generates nine independent outputs: one for entity recognition and one for each reasoning dimension. The instruction suffix shared across reasoning prompts that contain examples is reported separately for readability.

5.1.3 Cognition Score

For each reasoning type, the Cognition score is computed from the corresponding task-specific model output as the ratio between the correctly recognized CoRs and the annotated CoRs of that type. The overall Cognition score is then computed as the ratio between the total number of correctly recognized CoRs across all reasoning types and the total number of annotated CoRs.

In Cogbench, GPT-4 is employed as a binary classifier to determine whether a model-generated description contains the ground truth Chains of Reasoning. Due to cost considerations, we instead adopted **Gemini-2.5-Flash** [17] to perform the same evaluation procedure.

Gemini-2.5-Flash was released more than one year after GPT-4 Turbo and represents a more recent generation of large language models. Public benchmark aggregators, such as Chatbot Arena [41] and Artificial Analysis [3], report comparable overall performance between GPT-4 Turbo and Gemini-2.5-Flash across a range of general evaluation settings. In some benchmark settings, Gemini-2.5-Flash achieves equal or higher scores than GPT-4 Turbo, indicating that it belongs to the same performance tier and may even surpass it in specific aggregated evaluations.

Although these benchmarks do not specifically measure binary classification of reasoning content, they provide evidence that the two models exhibit similar levels of general reasoning and language understanding capabilities. On this basis, we consider Gemini-2.5-Flash a suitable and cost-effective alternative for the computation of the Cognition score in our experimental setting.

5.2 Selected Vision-Language Models

Among the architectures evaluated in CogBench, we focused our analysis and implementation on five models belonging to three families of LLaVA architectures. In particular, we used the Hugging Face implementations of **LLaVA v1.5** (7B[1] and 13B[2]), **LLaVA v1.6** (7B[3] and 13B[4]), and **LLaVA-OneVision-Chat**[5].

The models from the LLaVA v1.5 and LLaVA-Next families share the same backbone components, namely the CLIP ViT-L/14 visual encoder and the Vicuna v1.5 language model.

[1] https://huggingface.co/llava-hf/llava-1.5-7b-hf
[2] https://huggingface.co/llava-hf/llava-1.5-13b-hf
[3] https://huggingface.co/llava-hf/llava-v1.6-vicuna-7b-hf
[4] https://huggingface.co/llava-hf/llava-v1.6-vicuna-13b-hf
[5] https://huggingface.co/llava-hf/llava-onevision-qwen2-7b-ov-chat-hf

In contrast, LLaVA-OneVision-Chat relies on a different architectural configuration, based on the SigLIP-400M visual encoder and the Qwen2 language model. These architectural differences may influence how gaze information interacts with the internal representations of each model. Consequently, variations in performance across experiments may not depend solely on the gaze integration strategy, but also on the intrinsic characteristics of the underlying architecture.

5.3 Baseline Without Gaze Integration

To establish a reliable baseline, we first evaluated each pre-trained Vision-Language Model under a standard inference setting, without integrating any gaze information. In this configuration, the models received only the original input image and the corresponding textual prompt required by the task. No additional modality, weighting mechanism, or architectural modification was introduced.

This setting serves two purposes. First, it provides a direct comparison with the results reported in CogBench, where models are evaluated without external auxiliary signals. Second, it establishes a reference point against which the impact of gaze integration can be quantitatively measured in terms of both Recognition and Cognition scores. By isolating the models' original capabilities, this baseline allows us to attribute any performance variation observed in subsequent scenarios specifically to the incorporation of human gaze information, rather than to architectural changes or additional training.

5.4 Gaze-Augmented VLM Scenarios

5.4.1 Scenario 1: Gaze-Weighted Visual Features

The first experimental scenario was designed to investigate the effect of **directly weighting the visual embeddings** produced by the visual encoder using heatmap signal information. The goal was to observe how different models react to this external modulation of visual features, without introducing architectural changes or additional training. A schematic representation of the proposed architecture for this scenario is shown in Figure 5.2.

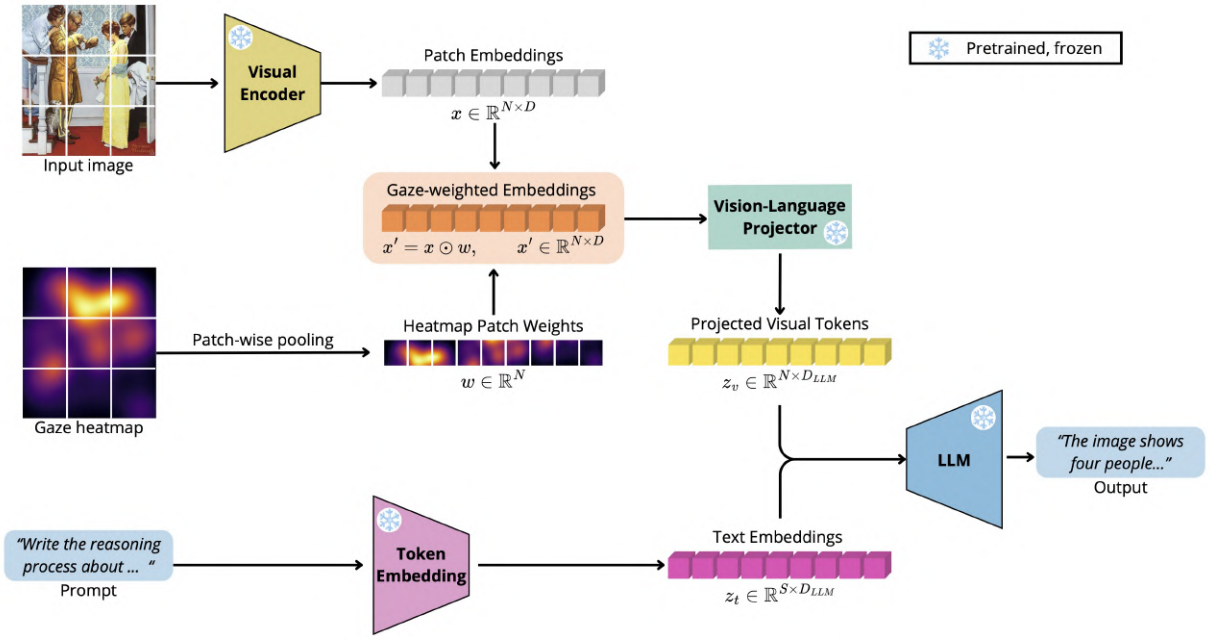


Figure 5.2: **Scenario 1: Gaze-weighted visual features.** Visual patch embeddings extracted from the image are modulated using patch-aligned gaze weights derived from the corresponding heatmap, and the resulting features are provided to the language model through the standard vision-language projector. All model components remain frozen; the figure illustrates the inference-only nature of this scenario.

5.4.1.1 Heatmap as Patch Weights

Heatmaps were processed to ensure alignment with the image patch embeddings produced by the visual encoder. To achieve this, they underwent the same initial pre-processing steps applied to the input images, such as cropping and resizing.

Unlike images, however, heatmaps were not explicitly divided into patches. Instead, two-dimensional adaptive pooling was applied to the full-resolution heatmap in order to obtain an output whose spatial dimensions match the patch grid of the visual encoder. In this way, a **one-to-one correspondence between heatmap weights and image patches** was established without explicitly replicating the patch extraction procedure used for images. Each pooled heatmap cell therefore corresponds to a single image patch and provides one scalar weight per visual token, enabling element-wise modulation of the corresponding embedding.

The only difference with respect to image processing concerned normalization. While images follow the standard normalization required by the visual encoder, heatmaps were

normalized using min-max normalization to constrain their values to the [0,1] range. This choice allows for direct and controlled weighting of the visual embeddings based on gaze intensity. Importantly, since the weights are bounded within [0,1], the element-wise multiplication does not amplify the embeddings of highly attended patches beyond their original magnitude. Instead, it attenuates the embeddings of unattended or weakly attended patches: the weighting mechanism primarily acts as a penalization of low-gaze regions rather than as an explicit amplification of salient ones.

In similar works [52, 45], gaze heatmaps are often normalized such that their values sum to one over the entire image, effectively treating them as probability distributions. In our pipeline, normalization is applied after adaptive pooling, that is, after the heatmap has been reduced to one scalar value per image patch. If a sum-to-one normalization were applied at this stage, the resulting patch-level weights would be numerically very small, since their total across all patches would be constrained to one. When directly used to scale the corresponding visual tokens, these small weights would drastically reduce their magnitude and significantly distort the feature distribution. This strong attenuation would degrade the quality of the visual representations and compromise the stability of the model outputs.

5.4.1.2 Expected Limitations

Given the simplicity of this approach, we expected that the models may not respond effectively to this hard-coded modification of the visual embeddings. By altering the distribution of the feature vectors without introducing any additional adaptation or fine-tuning of the architecture, we anticipated either unstable or undesirable outputs, or even no improvement over the baseline performance. Nevertheless, we considered this experiment to be informative, as it provides insight into how different models react to a direct and externally imposed weighting of their visual representations.

5.4.2 Scenario 2: Learnable Gaze Gating with Projector Fine-Tuning

This scenario represents the first attempt to enable the model to learn how to integrate gaze information within its architecture. Compared to Scenario 1, two modifications are introduced. First, we learn parameters that regulate how gaze-based patch weights are injected into the visual embeddings. Second, we fine-tune the multimodal projector to adapt it to the new gaze-weighted visual representation. Further details on the specific strategies adopted for parameters training and projector fine-tuning are described in Section 5.5, while Figure 5.3 illustrates the architecture adopted in this scenario.

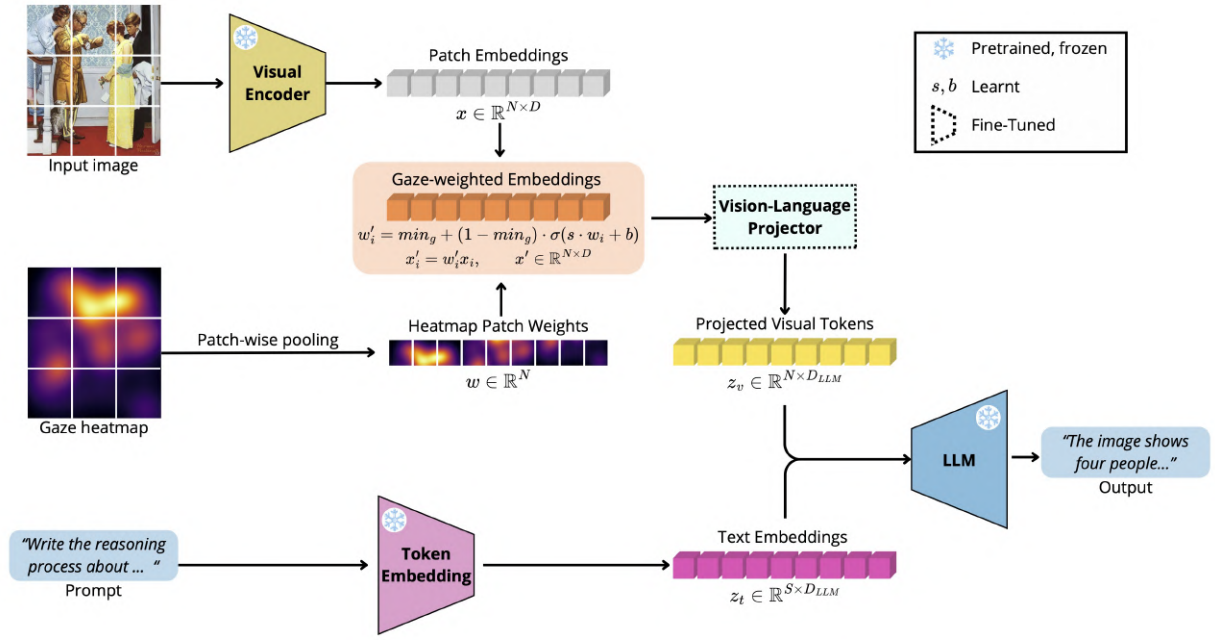


Figure 5.3: **Scenario 2: Learnable gaze gating with projector adaptation.** Gaze-derived patch weights are transformed through a learnable gating function before modulating image patch embeddings; the vision-language projector is fine-tuned to adapt to the resulting distribution shift. The visual encoder and the language model remain frozen.

5.4.2.1 Learnable Gaze Injection Mechanism

By adopting a simple and direct gaze-injection strategy, Scenario 1 remains relatively rigid. In Scenario 2, in turn, the objective is to allow the model to **learn how gaze information should modulate visual embeddings**, starting from a parameterized formulation.

The gaze weights are computed as in Scenario 1. However, each weight is now modulated by a learnable scale factor s and a learnable bias term b. The resulting value is passed through a sigmoid function σ, introducing non-linearity into the gating mechanism. This design increases the flexibility of gaze integration, moving beyond the simple element-wise multiplication between image embeddings and gaze weights.

In practice, the transformed weight is further linearly combined with the identity through a parameter min_{gate}, which ensures that the modulation does not entirely suppress the original visual signal. The final formulation is:

$$w' = min_{\text{gate}} + (1 - min_{\text{gate}}) \cdot \sigma(s \cdot w + b) \tag{5.1}$$

The resulting weights are then applied to the visual patch embeddings before being projected into the language model input space through the fine-tuned multimodal projector. To avoid confusion with the original gaze weights introduced in Scenario 1, throughout the remainder of this document these processed weights w' will be referred to as "gates".

5.4.2.2 Fine-Tuning the Multimodal Projector

Since the visual embeddings produced by the visual encoder are directly modified through gaze-based weighting before being passed to the multimodal projector, the statistical distribution of the visual tokens at the projector input differs from the one observed during pre-training. The multimodal projector was originally trained to map standard visual embeddings, extracted from unaltered images, into the language model input space. By introducing gaze-weighted embeddings, we effectively alter the structure and relative magnitudes of the visual features, thus creating a distribution shift at the interface between the visual encoder and the language model.

This mismatch may worsen the projector's ability to correctly align visual and textual representations: for this reason, **we fine-tune the multimodal projector** so that it can adapt to the new distribution of gaze-enhanced visual features. Through this adaptation, the projector potentially learns to reinterpret the modified embeddings and to preserve meaningful cross-modal alignment, thereby enabling the language model to better exploit the additional semantic signal conveyed by gaze information.

In the tested models, the multimodal projector consists of two linear layers, and both layers are fine-tuned in this scenario. Importantly, this strategy keeps the visual encoder and the language model frozen, and only updates the projector parameters. In this way, we isolate the effect of adapting the cross-modal alignment component, while maintaining the original representational capabilities of the backbone model.

5.4.3 Scenario 3: Dual Encoding with Projector Fine-Tuning

In this third and final experimental scenario, additional architectural modifications are introduced in order to reduce the external constraints imposed on how gaze information is represented, injected, and used within the model. The main change concerns the representation of gaze information. Instead of computing and manipulating patch weights derived directly from the heatmaps, the model is allowed to learn an embedding representation of the heatmaps themselves. This is achieved by encoding each heatmap through a fine-tuned version of the same visual encoder used to process images. Furthermore, the mapping from heatmap embeddings to patch weights is learned, together with the multimodal projector, as in Scenario 2. The architecture corresponding to this scenario is illustrated in Figure 5.4.

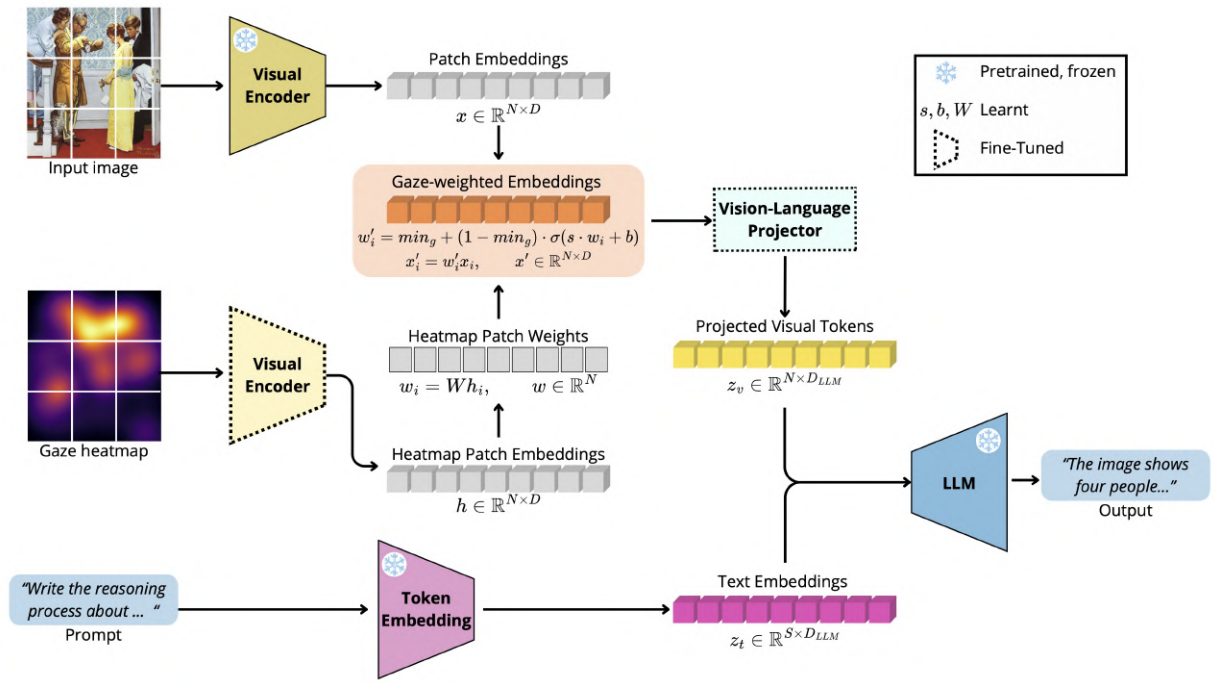


Figure 5.4: **Scenario 3: Dual encoding and projector fine-tuning** The image is encoded into patch embeddings, while the gaze heatmap is processed by a fine-tuned copy of the visual encoder to produce heatmap patch embeddings. These embeddings are mapped to scalar patch weights through a learnable linear layer and used to weight the corresponding image patch embeddings, which are projected into the language space via a fine-tuned vision-language projector. The image visual encoder and the language model remain frozen.

5.4.3.1 Fine-Tuning the Heatmap Visual Encoder and the Multimodal Projector

The objective is to allow the model to l**earn an appropriate representation of gaze heatmaps**, rather than relying on weights computed directly from raw heatmap values. Instead of training a new visual encoder from scratch specifically for heatmaps, we **fine-tune a copy of the pre-trained visual encoder** used for image embedding. This choice is motivated both by computational considerations, as training such encoders from scratch is costly and complex, and by representational consistency. By starting from the same pre-trained encoder, we encourage the learned gaze features to remain aligned with the visual features extracted from the input image, since both encoders share the same underlying architecture and initialization. Maintaining the same structure also ensures a direct correspondence between image patch embeddings and heatmap patch embeddings, both in number and dimensionality.

In addition, the multimodal projector is fine-tuned in this scenario as well. As in the previous setting, the distribution of the visual features presented to the projector is altered. The encoded heatmap patch embeddings are integrated into the visual representation using a strategy similar to that adopted in Scenario 2. However, in this case, the mapping from each heatmap patch embedding to its corresponding image patch weight is also learned, as described in the next subsection.

5.4.3.2 Learnable Mapping from Heatmap Embeddings to Patch Weights

Instead of computing heatmap patch weights through two-dimensional adaptive pooling, in this scenario the model learns how to derive gaze weights directly from the heatmap embeddings. After the fine-tuned heatmap visual encoder produces a patch-level representation of the input heatmap, **each patch embedding vector is passed through a trainable linear layer**. This layer maps each heatmap patch embedding to a single scalar value, which represents the corresponding gaze weight.

These learned weights are then used for gaze injection in the same manner as in Scenario 2. Specifically, the weights modulate the image patch embeddings according to Eq. 5.1, ensuring a consistent injection mechanism while allowing the computation of patch weights to be fully learned.

5.5 Training Protocol

This section describes the optimization procedure adopted in Scenarios 2 and 3, including both fine-tuning of existing components and training of newly introduced parameters.

5.5.1 Fine-Tuning Dataset

Since our final goal is to evaluate the modified architectures on CogBench, we do not use CogBench for fine-tuning. Doing so would bias the subsequent evaluation, as the learned adaptations would be influenced by the same data used to compute recognition and cognition scores. Even if only a subset of CogBench were used for training and excluded from evaluation, the results would no longer reflect a semantically valid benchmark setting, and the scores would be computed on a reduced test set, preventing a fair comparison with the baseline.

For these reasons, we fine-tune the components introduced in Scenarios 2 and 3 using an external gaze-annotated dataset. We adopt the **CapGaze1 dataset**, released by He et

al. [22] and available through their GitHub repository[6]. CapGaze1 contains 1,000 images selected from Pascal-50S [60], along with eye fixation data collected from five participants. Participants were asked to produce a one-sentence verbal description while viewing each image, and the corresponding captions were transcribed.

To match the processing applied to our CogBench gaze data, we convert CapGaze fixations into heatmaps by applying two-dimensional Gaussian kernels, as described in Section 4.3.2. Moreover, to preserve semantic and statistical comparability with our setting, we aggregate fixations across participants for each image and average the resulting heatmaps, mirroring the procedure used for CogBench. This choice is intended to make the heatmap distributions comparable across datasets.

Overall, this protocol enables the model to learn how to exploit gaze information without being directly exposed to CogBench during optimization. Therefore, any performance improvement observed on CogBench after fine-tuning on CapGaze1 provides evidence that gaze cues can generalize and effectively guide the generation process.

5.5.2 Optimization Objectives

Although CapGaze and our gaze-enhanced CogBench share similar gaze representations, the underlying description tasks differ substantially. In our data collection, participants were asked to produce detailed and structured reasoning about the images, whereas in CapGaze subjects generated a single-sentence caption. Moreover, after averaging CapGaze heatmaps across participants viewing the same image, it is no longer possible to align them with a corresponding textual description. An "average" heatmap does not correspond to a well-defined "average" caption.

For these reasons, standard cross-entropy loss on the generated text cannot be directly applied for fine-tuning. As an exploratory attempt, we also experimented with cross-entropy loss using the per-participant captions and the corresponding non-aggregated heatmaps. However, this approach led to unsatisfactory results on CogBench at inference time. During fine-tuning, the model adapted to generating short captions, and consequently lost the ability to produce the longer and more structured reasoning required by CogBench.

We therefore adopt a different strategy and design a composite loss function composed of three terms: an **attention alignment loss**, a **distillation loss**, and a **gaze gate regularizer**.

[6] https://github.com/SenHe/Human-Attention-in-Image-Captioning

5.5.2.1 Attention Alignment Loss

This component is based on the idea that gaze can guide the model's attention over image tokens during text generation. The objective is to align the model's internal attention distribution with the human gaze distribution.

The target gaze distribution is computed using the same procedure adopted in Scenario 1 (Section 5.4.1.1), including for Scenario 3, where the embedding strategy is applied only in the injection stage. Heatmaps are reduced to patch-level weights through two-dimensional adaptive pooling. In this case, however, the resulting weights are normalized so that they sum to one, and can therefore be interpreted as a probability distribution over image patches. This distribution represents the relative importance assigned by human gaze to each patch for the given description task.

Let w_i denote the weight associated with patch i. The normalized gaze weight g_i is obtained by applying the following normalization over the N image patches:

$$g_i = \frac{w_i}{\sum_{j=1}^{N} w_j}. \tag{5.2}$$

The target gaze distribution over the image patches is defined by the resulting normalized values, $\mathbf{g} = \{g_i\}_{i=1}^{N}$.

For each image, the model's attention over image tokens is extracted from the attention tensors produced during the forward pass. Specifically, let $\mathbf{A}^{(\ell)} \in \mathbb{R}^{H \times S \times S}$ denote the attention tensor at layer ℓ, where H is the number of attention heads and S is the sequence length. For each sample, we consider the attention of the last valid prompt token, indexed by q, over the set of image token positions $\mathcal{I} = \{i_1, \ldots, i_N\}$, where N is the number of image tokens. In autoregressive language models, the next-token prediction depends on the hidden state at the last valid position. Therefore, the attention of the last prompt token over image tokens provides a proxy for the model's assessment of image token relevance immediately before text generation begins.

For each selected layer ℓ, the attention assigned to image token $i_n \in \mathcal{I}$ is first averaged across heads as

$$\bar{a}_n^{(\ell)} = \frac{1}{H} \sum_{h=1}^{H} A_{h,q,i_n}^{(\ell)}, \qquad n = 1, \ldots, N. \tag{5.3}$$

The resulting values are then normalized over image tokens so that they define a probability distribution:

$$a_n^{(\ell)} = \frac{\bar{a}_n^{(\ell)}}{\sum_{m=1}^{N} \bar{a}_m^{(\ell)}}, \qquad n = 1, \ldots, N. \tag{5.4}$$

Finally, if the last K attention layers are used, the final attention distribution is obtained by averaging these normalized per-layer distributions:

$$\hat{a}_n = \frac{1}{K} \sum_{\ell \in \mathcal{L}} a_n^{(\ell)}, \qquad n = 1, \ldots, N, \tag{5.5}$$

where $\mathcal{L}$ denotes the set of selected last layers. The resulting vector $\hat{\mathbf{a}} = \{\hat{a}_n\}_{n=1}^{N}$ defines the model attention distribution over image tokens.

The attention alignment loss is defined as the Kullback-Leibler divergence between the model's attention distribution and the target gaze distribution.

$$\mathcal{L}_{\text{attn}} = D_{\text{KL}}(\mathbf{g} \,\|\, \hat{\mathbf{a}}) = \sum_{i=1}^{N} g_i \log \frac{g_i}{\hat{a}_i}. \tag{5.6}$$

5.5.2.2 Distillation Loss

The attention alignment loss alone is not sufficient. When optimizing only this objective, the multimodal projector learns to map gaze-weighted visual embeddings into the language model input space, but there is no constraint on how this affects text generation. In practice, we observed catastrophic forgetting, as the language model progressively lost its ability to generate coherent and well-structured sentences.

To mitigate this issue, and to preserve generation capabilities while guiding attention with gaze information, we introduce a distillation loss. This loss is designed to **counteract forgetting by transferring knowledge** from a teacher model to a student model. The objective is to minimize the divergence between the output distributions of the teacher and those of the student.

In our setting, the student corresponds to the fine-tuned model with gaze injection and LoRA adapters, while the teacher is the original pre-trained model without gaze injection and without any additional adaptation.

For a given input image, let $\mathbf{z}_i^{(s)}, \mathbf{z}_i^{(t)} \in \mathbb{R}^V$ denote, respectively, the student and teacher logits at token position i, where V is the vocabulary size.

Following standard knowledge distillation [25], both logits are softened with a temperature parameter $T > 1$. The standard softmax function applied by models to raw logits typically produces highly peaked distributions where the most probable token dominates, effectively masking the relative probabilities of less likely tokens. By applying temperature scaling, the distribution is smoothed to reveal the structural relationships and similarities between all tokens in the vocabulary as learned by the teacher model. This provides a much richer gradient signal for the student than standard hard labels.

The student probability distribution $\mathbf{p}_i^{(s)} = \{p_{i,v}^{(s)}\}_{v=1}^{V}$ is obtained by applying the softmax function with temperature scaling to the model's logits:

$$p_{i,v}^{(s)} = softmax\left(z_{i,v}^{(s)}/T\right) = \frac{\exp\left(z_{i,v}^{(s)}/T\right)}{\sum_{u=1}^{V}\exp\left(z_{i,u}^{(s)}/T\right)}, \qquad v = 1, \ldots, V. \tag{5.7}$$

Similarly, the corresponding teacher probability distribution $\mathbf{p}_i^{(t)} = \{p_{i,v}^{(t)}\}_{v=1}^{V}$ is computed as:

$$p_{i,v}^{(t)} = softmax\left(z_{i,v}^{(t)}/T\right) = \frac{\exp\left(z_{i,v}^{(t)}/T\right)}{\sum_{u=1}^{V}\exp\left(z_{i,u}^{(t)}/T\right)}, \qquad v = 1, \ldots, V. \tag{5.8}$$

For each valid token position, the token-level distillation term is given by the Kullback-Leibler divergence from the teacher distribution to the student distribution:

$$d_i = D_{\mathrm{KL}}\left(\mathbf{p}_i^{(t)} \,\|\, \mathbf{p}_i^{(s)}\right) = \sum_{v=1}^{V} p_{i,v}^{(t)} \log \frac{p_{i,v}^{(t)}}{p_{i,v}^{(s)}}. \tag{5.9}$$

Let $m_i \in \{0, 1\}$ denote the attention mask at position i, where $m_i = 1$ indicates a valid token and $m_i = 0$ corresponds to padding. The distillation loss is computed by averaging the token-level divergences over all valid positions in the sequence and scaling the result by T^2, following the formulation introduced by [25]:

$$\mathcal{L}_{\text{distill}} = T^2 \frac{\sum_{i=1}^{S} m_i \, d_i}{\sum_{i=1}^{S} m_i}, \tag{5.10}$$

where S is the sequence length.

5.5.2.3 L2 Gaze Gate Regularizer

Since training is performed on an external dataset, we introduce an additional regularization term to reduce overfitting and improve generalization. This term applies ℓ_2 regularization to the gaze gates computed by the injection mechanism in Eq. 5.1.

Specifically, we constrain the expected value of gaze gates for each image to remain close to 1. This stabilizes the injection mechanism and ensures that the learned modulation of visual features does not systematically amplify or suppress the overall magnitude of the original image embeddings. As a result, the overall energy of the visual representation remains approximately constant, while still allowing local modifications driven by the gaze signal.

Let w'_i denote the scalar gate applied element-wise to the embedding of patch i, with $i = 1, \ldots, N$, where N is the number of image patches. The regularization loss is defined as the mean squared deviation of these gates from the identity value 1:

$$\mathcal{L}_{\text{gate}} = \frac{1}{N} \sum_{i=1}^{N} (w'_i - 1)^2. \tag{5.11}$$

5.5.2.4 Final Loss Formulation and Cross-Validation

Given the considerations discussed above, the final loss function used for fine-tuning and training in Scenarios 2 and 3 is defined as a linear combination of the three previously introduced loss terms:

$$\mathcal{L} = \lambda_{\text{attn}} \mathcal{L}_{\text{attn}} + \lambda_{\text{dist}} \mathcal{L}_{\text{dist}} + \lambda_{\text{gate}} \mathcal{L}_{\text{gate}} \tag{5.12}$$

where λ_{attn}, λ_{dist}, and λ_{gate} are hyperparameters that control the relative contribution of each loss component to the overall objective.

To estimate suitable values for these hyperparameters, K-fold cross-validation was applied for each tested model, in both Scenario 2 and Scenario 3. In order to reduce computational cost and training time, the following design choices were adopted:

- Three folds were used. The CapGaze dataset was divided into training and validation splits with a 2:1 ratio.
- Since the $\mathcal{L}_{\text{gate}}$ term acts as an ℓ_2 regularizer whose role is to stabilize the model rather than to drive the optimization process, λ_{gate} was excluded from cross-validation. Its value was fixed to 0.05 in all experiments.
- λ_{attn} was selected from the set $[0.05, 0.5, 1]$, while λ_{dist} was selected from the set $[0.5, 1, 1.2]$. These candidate values were chosen after preliminary experiments conducted with *LLaVA-v1.5-7B*, which showed that they produced substantially different validation losses. Therefore, exploring the best combination within these ranges was considered sufficient and representative.
- The optimal pair $(\lambda_{\text{attn}}, \lambda_{\text{dist}})$ was selected as the one minimizing the average of the validation attention alignment loss and the validation distillation loss.
- The temperature parameter T used for the computation of the distillation loss (Section 5.5.2.2) and the minimum gating parameter min_{gate} (Eq. 5.1) were fixed in

order to avoid making cross-validation heavier. Their values were set to 2.0 and 0.05, respectively.

Chapter 6

Results

This chapter presents the results obtained by the modified models under the CogBench evaluation protocol. We report recognition and cognition scores for each configuration. All tables include the recognition score, the overall cognition score, and the cognition scores for each reasoning category defined in CogBench.

6.1 Baseline Performance

Table 6.1 shows the baseline results, obtained by evaluating the original pre-trained models with our redefined CogBench protocol, without any form of gaze integration or alignment.

Although the evaluation procedure slightly differs from the original CogBench setting, the resulting scores are consistent with those reported in the benchmark. Among the tested models, LLaVA-OV-7B-Chat achieves the highest recognition and overall cognition scores, while the remaining models perform significantly worse.

The difference is particularly evident in recognition. Three of the four other LLaVA variants remain below 50% recognition, indicating substantial limitations in entity identification. Cognition scores are generally low across all models, including the best-performing one. LLaVA-OV-7B-Chat consistently achieves the highest scores across all reasoning categories.

In agreement with the observations reported in CogBench, reasoning about events represents the main difficulty. Event Reasoning, Event Relationship Reasoning, and Next Moment Event Reasoning obtain the lowest scores, confirming that models struggle to capture and reason about complex event structures in images.

LLaVA Model	Recognition Score	Overall Cognition Score	Time	Location	Character	Character Relationship	Event	Event Relationship	Next Moment Event	Mental State
v1.5-7B	40.2	18.5	38.3	52.5	21.7	45.2	6.8	0.9	1.9	25.7
v1.5-13B	44.5	18.2	38.3	50.8	17.9	39.2	9.7	1.6	6.5	23.0
v1.6-7B	48.4	21.7	42.6	62.7	25.5	45.2	15.3	4.2	5.6	18.7
v1.6-13B	51.5	23.2	36.2	66.1	32.1	43.7	17.5	4.5	7.5	20.9
OV-7B-Chat	**65.4**	**35.6**	**55.3**	**76.8**	**49.1**	**53.2**	**29.2**	**14.4**	**18.7**	**38.1**

Table 6.1: Baseline Recognition and Cognition Scores (overall and individual reasoning types). Scores are presented in percentages. The highest score in each column is highlighted in gray.

6.2 Scenario 1 Results

The scores obtained in Scenario 1 are reported in Table 6.2. In this setting, visual features are directly modified through gaze weighting before being passed to the multimodal projector, without any fine tuning. Consequently, there is no guarantee that the models can still preserve the semantic information of the original image when processing these altered embeddings.

LLaVA Model	Recognition Score	Overall Cognition Score	Time	Location	Character	Character Relationship	Event	Event Relationship	Next Moment Event	Mental State
v1.5-7B	31.3 (-8.9)	9.5 (-9.0)	10.6 (-27.7)	39.0 (-13.5)	17.0 (-4.7)	19.4 (-25.8)	3.6 (-3.2)	0.9 (+0.0)	0.9 (-1.0)	9.4 (-16.3)
v1.5-13B	43.3 (-1.2)	16.8 (-1.4)	21.3 (-17.0)	50.3 (-0.5)	**28.3** (+10.4)	39.5 (+0.3)	8.0 (-1.7)	**1.9** (+0.3)	**5.6** (-0.9)	**17.5** (-5.5)
v1.6-7B	39.5 (-8.9)	14.5 (-7.2)	25.5 (-17.1)	51.4 (-11.3)	22.6 (-2.9)	31.9 (-13.3)	8.1 (-7.2)	0.9 (-3.3)	2.8 (-2.8)	12.2 (-6.5)
v1.6-13B	**44.9** (-6.6)	**19.1** (-4.1)	**38.3** (+2.1)	**57.6** (-8.5)	25.5 (-6.6)	**41.8** (-1.9)	**12.6** (-4.9)	1.6 (-2.9)	4.7 (-2.8)	17.0 (-3.9)
OV-7B-Chat	-	-	-	-	-	-	-	-	-	-

Table 6.2: Scenario 1 Recognition and Cognition Scores (overall and individual reasoning types), with variation with respect to the baseline. Scores are presented in percentages, variations are presented in percentage points. Note that scores for LLaVA-OV-7B-Chat are not available because direct gaze weighting broke the generational capabilities of the model. The highest score in each column is highlighted in gray.

Despite this uncontrolled modification, models from the LLaVA 1.5 and LLaVA 1.6 families remain capable of generating grammatically correct and semantically coherent text

Baseline Model
The man in the picture is holding a case of beer and a basket, possibly containing a dog. He is standing on a porch or a staircase. **It is not possible to determine his exact role or identity** from the image alone. However, one could speculate that he might be a friend or a family member visiting someone at the location [...].

Scenario 1 Model
The image features a man holding a box of beer and a basket, standing in front of a building. **It is possible that the man is a delivery person** or a worker responsible for transporting goods to the building. The presence of the beer box suggests that he might be delivering beer to a store or a bar [...].

Figure 6.1: Outputs produced by LLaVA-v1.5-13B for the "Character Reasoning" task on the same CogBench image, comparing the baseline configuration (no gaze input) with Scenario 1, where gaze heatmaps are used to weight visual features. The conclusion of the annotated ground-truth CoR regarding the role of the man in the image indicates that *the man is a courier*.

grounded in the input image and task. In contrast, the generation capability of LLaVA-OV-7B-Chat collapses completely under this configuration. The model does not only lose semantic alignment with the input, but also fails at the syntactic level, producing text that is not well-formed English. For this reason, reasoning and cognition scores were not computed for this model in Scenario 1.

A possible explanation is that LLaVA-OV-7B-Chat relies on stricter statistical properties of the visual feature distribution. Directly altering these features without adaptation may disrupt the internal representation expected by the model, leading to instability during generation.

For the remaining models, as anticipated, recognition and overall cognition scores generally do not improve compared to the baseline. However, performance does not collapse entirely, and some degree of reasoning is still observed, albeit at lower levels. Interestingly, a few small improvements appear in specific cases. In particular, LLaVA-v1.5-13B shows gains in three reasoning categories, including a notable increase of +10.4 percentage points in Character Reasoning.

Although this model reports the smallest overall reduction in scores, suggesting that this specific architecture may derive some limited benefit from the direct gaze-weighting strat-

egy, the fact that the improvement in Character Reasoning is not observed in any other model indicates that it is likely an isolated gain. Since the evaluation is performed by Gemini through binary judgments of whether a key point is present in the model's output, the increase may result from minor phrasing differences that led Gemini to classify certain borderline cases as positive.

In other instances, however, the model's outputs change more substantially and become clearer in identifying the characters' roles in the image, as in the example shown in Figure 6.1. For these reasons, it remains unclear whether the model's ability to reason about characters genuinely improved in this scenario, or whether the observed gain is partially due to evaluation variability.

Despite the overall reduction in scores, LLaVA-v1.6-13B achieves the highest performance in this scenario across most categories, including both the recognition score and the overall cognition score. It is surpassed by its predecessor, LLaVA-v1.5-13B, in only a few categories.

6.3 Scenario 2 Results

Scenario 2 shows improvements across most models (Table 6.3). The clearest gains are observed for LLaVA-v1.5-7B, which improves both the recognition score and the overall cognition score. It also increases in five out of the eight reasoning types, while remaining approximately unchanged in one, and decreasing in only two, namely Character Relationship Reasoning and Mental State Reasoning.

The other models in the LLaVA v1.5 and v1.6 families also exhibit improvements in several cognitive reasoning types, although they all do not show higher recognition scores. LLaVA-v1.6-7B achieves the highest recognition score in this scenario, despite still being lower than its baseline value. It also obtains the best results for Location Reasoning and Next Moment Event Reasoning.

With the exception of Mental State Reasoning, where LLaVA-v1.5-13B performs best, LLaVA-v1.6-13B achieves the highest scores across the remaining reasoning types as well as in the overall cognition score. This model also reports the largest improvement observed across all models and reasoning dimensions for Special Time Reasoning, with an increase of +10.6 percentage points.

The only model that shows a marked decrease with respect to the baseline is LLaVA-OV-7B-Chat. Although it partially recovers the ability to produce well-formed English sentences compared to Scenario 1, the generated outputs are often empty or not semantically grounded in the input images across many reasoning dimensions. For this reason, while its scores are reported for completeness, the results indicate that the gaze-injection

LLaVA Model	Recognition Score	Overall Cognition Score	Time	Location	Character	Character Relationship	Event	Event Relationship	Next Moment Event	Mental State
v1.5-7B	41.0 (+0.8)	19.2 (+0.7)	40.4 (+2.1)	53.7 (+1.2)	27.4 (+5.7)	43.0 (-2.2)	9.6 (+2.8)	0.9 (+0.0)	4.7 (+2.8)	23.5 (-2.2)
v1.5-13B	44.5 (+0.0)	18.7 (+0.5)	36.2 (-2.1)	49.7 (-1.1)	17.9 (+0.0)	45.2 (+6.0)	8.4 (-1.3)	2.4 (+0.8)	4.7 (-1.8)	**24.5** (+1.5)
v1.6-7B	**47.1** (-1.3)	21.0 (-0.7)	34.0 (-8.6)	**63.3** (+0.6)	30.2 (+4.7)	43.7 (-1.5)	13.3 (-2.0)	3.1 (-1.1)	**7.5** (+1.9)	19.9 (+1.2)
v1.6-13B	42.9 (-8.6)	**24.2** (+1.0)	**46.8** (+10.6)	62.1 (-4.0)	**31.1** (-1.0)	**49.4** (+5.7)	**16.1** (-1.4)	**8.5** (+4.0)	6.5 (-1.0)	22.1 (+1.2)
OV-7B-Chat	8.6 (-56.8)	3.2 (-32.4)	6.4 (-48.9)	15.3 (-61.5)	6.6 (-42.5)	2.3 (-50.9)	2.3 (-26.9)	0.0 (-14.4)	0.0 (-18.7)	3.1 (-35.0)

Table 6.3: Scenario 2 Recognition and Cognition Scores (overall and individual reasoning types), with variation with respect to the baseline. Scores are presented in percentages, variations are presented in percentage points. The highest score in each column is highlighted in gray.

strategies applied in Scenarios 1 and 2 primarily affect this model's ability to generate coherent and contextually grounded text, rather than its reasoning performance.

Overall, excluding LLaVA-OV-7B-Chat, Scenario 2 introduces consistent improvements over the baseline, particularly in the per-reasoning-type cognition scores. This suggests that learning the gaze-injection parameters and fine-tuning the multimodal projector enable the models to better adapt to gaze-weighted visual embeddings and to exploit gaze as a meaningful semantic guidance signal during reasoning.

6.4 Scenario 3 Results

In Scenario 3 (Table 6.4), LLaVA-OV-7B-Chat shows clear improvements. The model gains across almost all reasoning types, as well as in the overall cognition score, with the largest increase observed in Special Time Reasoning (+6.4 percentage points). Compared to the previous scenarios, this result is notable: the model not only fully recovers its text generation capabilities, but also improves its reasoning performance.

A likely explanation is that, in this configuration, the model is allowed to learn not only the gaze-injection parameters and the projector adaptation, but also an appropriate representation of the gaze heatmaps themselves. Instead of imposing a fixed transformation, the fact that heatmaps are encoded through a fine-tuned visual encoder initialized from the same pre-trained backbone used for images, enables the model to produce gaze embeddings that are more aligned with its internal visual and textual representations, allowing it to extract meaningful semantic guidance from the gaze signal.

Among the remaining models, LLaVA-v1.6-7B reports the largest number of improvements, with gains in the overall cognition score and in six out of eight reasoning types.

In contrast, recognition scores do not generally benefit from this approach. Only LLaVA-v1.5-7B shows a slight increase, while all other models experience a decrease. The largest drop is observed for LLaVA-v1.6-13B, which is subject to a -7.3 percentage points variation compared to the baseline. A more detailed analysis of the reduction in recognition performance is provided in Section 6.5.2.

LLaVA Model	Recognition Score	Overall Cognition Score	Time	Location	Character	Character Relationship	Event	Event Relationship	Next Moment Event	Mental State
v1.5-7B	43.1 (+2.9)	18.2 (-0.3)	38.3 (+0.0)	50.8 (-1.7)	21.7 (+0.0)	43.0 (-2.2)	7.8 (+1.0)	0.9 (+0.0)	3.7 (+1.8)	24.5 (-1.2)
v1.5-13B	44.4 (-0.1)	18.4 (+0.2)	29.8 (-8.5)	50.3 (-0.5)	20.8 (+2.9)	43.7 (+4.5)	9.6 (-0.1)	1.4 (-0.2)	1.9 (-4.6)	23.5 (+0.5)
v1.6-7B	47.5 (-0.9)	21.9 (+0.2)	42.6 (+0.0)	63.8 (+1.1)	26.4 (+0.9)	47.1 (+1.9)	13.6 (-1.7)	4.9 (+0.7)	7.5 (+1.9)	19.7 (+1.0)
v1.6-13B	44.2 (-7.3)	22.6 (-0.6)	42.6 (+6.4)	58.2 (-7.9)	22.6 (-9.5)	44.9 (+1.2)	16.7 (-0.8)	4.7 (+0.2)	11.2 (+3.7)	22.3 (+1.4)
OV-7B-Chat	**65.1** (-0.3)	**36.9** (+1.3)	**61.7** (+6.4)	**78.0** (+1.2)	**49.1** (+0.0)	**57.0** (+3.8)	**29.8** (+0.6)	**14.8** (+0.4)	**20.6** (+1.9)	**39.3** (+1.2)

Table 6.4: Scenario 3 Recognition and Cognition Scores (overall and individual reasoning types), with variation with respect to the baseline. Scores are presented in percentages, variations are presented in percentage points. The highest score in each column is highlighted in gray.

6.5 Discussion

6.5.1 Effect on Models Architectures

The behavior of LLaVA-OV-7B-Chat across the three scenarios highlights the link between the model architecture and the sensitivity to the way gaze information is represented and injected. In Scenarios 1 and 2, where a fixed or externally imposed gaze representation was used, the model either failed or showed degraded performance. Conversely, in Scenario 3, where the model was allowed to learn its own gaze representation in addition to the injection parameters, both text generation and cognition scores improved.

This suggests that for more complex architectures forcing a predefined gaze transformation may disrupt internal feature distributions, whereas learning a representation that is aligned with the model's visual backbone allows gaze to be incorporated more effectively.

At the same time, simpler models, such as those in the LLaVA-v1.5 family, appear to react differently. These models benefit more from the structured and constrained injection strategy of Scenario 2, while the additional flexibility introduced in Scenario 3 does not always yield further gains. This may indicate that smaller or less complex architectures are more stable under externally imposed modifications, but may not fully exploit the added representational capacity introduced in Scenario 3.

Overall, these results suggest that **there is no single gaze-injection strategy that is optimal across architectures**. The effectiveness of gaze integration appears to depend strongly on the model's internal design and on how closely the injected signal aligns with its learned feature distributions.

6.5.2 Effect on Recognition Scores

Incorporating the gaze signal generates a **decrease in the recognition score in all scenarios** for the vast majority of the models, even when the performance in several reasoning dimensions and the overall cognition score improve (see Scenarios 2 and 3). A plausible explanation concerns the data and objective used during fine tuning. In both scenarios, adaptation is performed using the CapGaze dataset, where gaze is collected during an image captioning task. This training setting is closer to the cognition evaluation in CogBench, which rewards the presence of high level semantic content related to the different reasoning types, than to the recognition evaluation, which measures how many annotated entities are explicitly mentioned.

Entity listing is structurally different from captioning. It requires an exhaustive mention of objects and people, often including less salient details. In contrast, description based tasks encourage selecting the most relevant elements to construct a coherent narrative. As a consequence, fine-tuning that aligns the model's attention with gaze during descriptive generation may bias the output toward salient regions while reducing coverage of secondary entities. This can support improvements in reasoning related scores, but at the same time lower the recall of annotated entities.

For this reason, the reduction in recognition scores may indicate that the models' outputs become more selective and more focused on the narrative structure of the scene after fine-tuning. This interpretation is consistent with the optimization objective, which explicitly encourages gaze-conditioned descriptive behavior, while recognition is evaluated through entity coverage, a signal that is not directly optimized.

Moreover, fixations collected for the recognition task are, on average, spatially more distributed than those collected for the reasoning tasks (Table 4.1). Since the averaged heatmaps are obtained from a larger number of fixations, the resulting representation may become less semantically informative, as attention is spread across many regions of the

image. This effect may have negatively affected the model's ability to recognize entities. When fixation patterns are widely distributed, the resulting heatmaps provide weaker guidance about the most relevant regions of the scene and may highlight fewer specific objects that should be recognized.

On the other hand, more complex reasoning dimensions, such as Event Relationship Reasoning, Next Moment Event Reasoning, or Mental State Reasoning, may benefit from the guidance provided by these fixations. Even when they are sparse, fixations located in strategically relevant regions of the image can guide the model toward the parts of the scene that are most informative for higher-level reasoning.

6.5.3 Effect on Reasoning Dimensions

Table 6.5 reports the average variations of all scores with respect to the baseline across the tested models, for each scenario. The average values are reported for the recognition score, the overall cognition score, and for each reasoning dimension. For Scenario 1, the average variation is computed excluding LLaVA-OV-7B-Chat, since this model did not produce meaningful outputs that could be reliably evaluated. For Scenario 2, two averages are reported: either including LLaVA-OV-7B-Chat or excluding it. As discussed earlier, the scores obtained by that model in Scenario 2 mainly reflect a degradation in its ability to generate semantically grounded answers, rather than a genuine change in recognition or reasoning performance. Therefore, excluding this model provides a more representative estimate of the effect of the gaze-injection strategy on the remaining architectures.

In Scenario 2 (excluding LLaVA-OV-7B-Chat) and Scenario 3, a similar pattern emerges across several reasoning dimensions. In particular, the overall cognition score, as well as the scores for Special Time Reasoning, Character Relationship Reasoning, Event Relationship Reasoning, Next Moment Event Reasoning, and Mental State Reasoning, show an average increase in both scenarios.

This pattern suggests that these reasoning dimensions may benefit, on average, from gaze injection and attention alignment. Since these types of reasoning often require integrating contextual cues and understanding relationships between elements in the scene, they may be more sensitive to the semantic guidance provided by human gaze.

Scenario	Recognition Score	Overall Cognition Score	Time	Location	Character	Character Relationship	Event	Event Relationship	Next Moment Event	Mental State
S1 (excl. OV)	-6.4	-5.4	-14.9	-8.4	-0.9	-10.2	-4.2	-1.5	-1.9	-8.0
S2	-13.2	-6.2	-9.4	-13.0	-6.6	-8.6	-5.8	-2.1	-3.4	-6.7
S2 (excl. OV)	-2.3	+0.4	+0.5	-0.8	+2.3	+2.0	-0.5	+0.9	+0.5	+0.4
S3	-1.1	+0.2	+0.9	-1.6	-1.1	+1.8	-0.2	+0.2	+0.9	+0.6

Table 6.5: Average variation in performance relative to the baseline across models for each scenario. Values represent changes in percentage points for both the overall metrics (Recognition and Cognition scores) and the individual reasoning dimensions defined in CogBench. "excl. OV" indicates rows where the scores obtained by LLaVA-OV-7B-Chat are excluded from the average, as the model's generation capabilities were degraded.

6.6 Limitations

6.6.1 Dataset Size and Coverage

The experimental evaluation conducted in this thesis is constrained by the size and structure of the available datasets. CogBench contains a relatively limited number of images and, although each image is enriched with gaze data collected from three participants, this scale remains modest compared to the data volumes typically required for strong statistical generalization in deep learning settings.

Furthermore, while collecting gaze data from three participants per image is consistent with the CogBench ground truth annotation protocol, which relies on majority voting among three annotators, averaging gaze heatmaps across three observers may not fully capture generalized human attention patterns for a given reasoning task. Although aggregated heatmaps provide a more stable and less noisy representation than subject-specific heatmaps considered individually, they may still fail to represent the variability and richness of attention strategies of a larger sample of subjects.

In addition, fine-tuning was performed on an external gaze-annotated captioning dataset that differs from CogBench in both task structure and semantic depth. While this choice was necessary to preserve the validity of the benchmark evaluation and avoid data leakage, it limits the amount of reasoning-specific gaze supervision available during training. Consequently, **the improvements observed in the experimental results should be interpreted as indicative** rather than conclusive evidence of the effectiveness of gaze integration for structured visual reasoning.

6.6.2 Gaze Representation Choices

In this work, gaze information is represented as spatial heatmaps derived from fixation coordinates and aggregated across participants. This representation involves several design choices, including Gaussian smoothing, percentile clamping, normalization strategies, and averaging across observers. Although these operations are standard in gaze analysis, they inevitably introduce inductive biases that influence the signal provided to the model.

Importantly, the adopted heatmap representation discards temporal information such as fixation order and scanpath dynamics. As a result, the model receives only a static approximation of spatial attention, rather than sequential input data, which could better account for how humans explore and interpret a scene during visual reasoning. This design choice is consistent with the objective of the present study, which focuses on providing spatial guidance to the model, based on visual attention patterns to answer a specific reasoning question. However, this representation may not be the most appropriate strategy for fully exploiting the potential of integrating gaze into models' reasoning tasks, and the temporal structure of attention shifts could actually carry additional key information about causal inference or general event understanding.

6.6.3 Evaluation Based on Recall-Only Metrics

The evaluation protocol adopted from CogBench primarily relies on recall-based metrics. Recognition and Cognition scores measure whether annotated entities or Chains of Reasoning are mentioned in the model output, but **they do not penalize over-generation or hallucinated content**. As a result, a model may achieve higher scores by mentioning more elements, even if some of them are incorrect or unsupported by the image.

Furthermore, cognition scoring is performed through automatic binary classification by a large language model. Although this approach enables scalable evaluation, it introduces potential variability related to phrasing sensitivity and classifier uncertainty. The absence of explicit precision measurements and structured reasoning validation limits the interpretability of score improvements, as higher recall alone does not necessarily imply more accurate or faithful reasoning.

6.7 Future Works

6.7.1 Sequential Gaze Representations

A natural extension of this work consists in moving beyond static heatmap representations and incorporating temporal information derived from gaze scanpath sequences. Reasoning about complex scenes may benefit from modeling how attention unfolds over time, especially for tasks involving event structure, causal relationships, or mental state inference. Future work could therefore investigate methods for encoding fixation sequences, attention transitions, or scanpath embeddings, and for aligning these temporal patterns with the attention mechanisms of Vision-Language Models.

6.7.2 Precision-Aware and Structured Evaluation Metrics

Another promising direction concerns the evaluation protocol. Future work could extend the CogBench framework, which is based on recall-based scores, by introducing precision-aware metrics in order to penalize hallucinated or unsupported content.

6.7.3 Synthetic Gaze Data Generation

The collection of human gaze data is resource-intensive and limits scalability. An alternative direction consists in generating synthetic fixation data to approximate human attention patterns. Synthetic gaze maps could be obtained using saliency prediction models or scanpath prediction architectures trained to estimate human-like visual exploration. These synthetic signals could then be used to pre-train or augment gaze-conditioned architectures, either independently or in combination with real gaze data.

Although synthetic fixations would not perfectly replicate authentic human visual behavior, they could provide a scalable approximation of attention guidance and enable training on substantially larger datasets. Future research could systematically compare the effects of synthetic and real gaze supervision on recognition and reasoning performance, and investigate whether hybrid strategies combining both sources of information lead to improved robustness and generalization.

Another promising application of synthetic gaze arises once a stable three-modality architecture (image, text, and gaze) has been successfully trained. In such a setting, a gaze prediction module could be used to generate synthetic gaze maps directly from the input image at inference time. The predicted gaze could then be provided to the model together with the image and textual prompt. This approach would eliminate the need for human

gaze data at inference time, while still leveraging the benefits of attention-guided reasoning. Exploring whether model-predicted gaze can effectively substitute human gaze supervision represents an interesting direction for making gaze-augmented Vision-Language Models practically usable at scale.

Finally, the dataset collected in this work could also serve as a resource for training multimodal scanpath prediction or generation models. Because the gaze data was collected during reasoning tasks aligned with the reasoning dimensions defined in CogBench, it provides gaze supervision associated with cognitively complex visual reasoning processes. Datasets that combine gaze annotations with such structured and high-level reasoning tasks are relatively rare, and this resource could therefore support future research on modeling human-like visual exploration in cognitively demanding image understanding scenarios.

Chapter 7

Conclusion

This thesis investigated whether integrating human gaze information into Vision-Language Models can lead to a more human-like structured visual reasoning, exploiting the CogBench evaluation framework. The latter includes a set of images with rich semantics, annotated with ground-truth human descriptions. Towards the thesis goal, a gaze-augmented version of the CogBench dataset was constructed by collecting human gaze data through a dedicated eye-tracking experiment, resulting in gaze heatmaps aligned with specific tasks, including entity listing (recognition task) and image descriptions across different reasoning dimensions (cognition tasks). This enriched dataset enabled a controlled evaluation of VLMs' gaze-conditioned reasoning abilities.

Three integration approaches were designed and evaluated, ranging from direct gaze-based modulation of visual features (Scenario 1) to learnable injection mechanisms (Scenario 2) and learnable encoding of gaze heatmaps (Scenario 3). The results show that naive gaze weighting without adaptation is generally insufficient, leading to worse or unchanged recognition and cognition scores. Directly modulating visual embeddings using fixed gaze weights (Scenario 1) led to an overall decrease in both recognition and cognition scores for most LLaVA models, and caused severe instability in one architecture, LLaVA-OV-7B-Chat. This model employs a larger visual encoder, which produces more complex and distributionally constrained visual embeddings. This suggests that externally altering the visual feature distribution without allowing model adaption can disrupt cross-modal alignment and degrade performance in architectures with more structured visual representations.

Introducing learnable gaze gating and projector fine-tuning (Scenario 2) led to consistent improvements in cognition scores for most models, particularly when excluding the LLaVA-OV-7B-Chat architecture, which proved unstable under gaze injection. **Gains were observed in several reasoning dimensions**, suggesting that when the model is allowed to

adapt the vision–language interface, gaze acts as a meaningful semantic guidance signal.

Finally, the most flexible integration strategy (Scenario 3), in which gaze heatmaps are encoded through a fine-tuned visual encoder and transformed into patch-level scalar weights through a learnable mapping, yielded **improvements in cognition for several reasoning dimensions across architectures**. In particular, differently from the previous scenarios, LLaVA-OV-7B-Chat recovered its generation capabilities and achieved significant gains in overall cognition as well as in multiple individual reasoning dimensions. This suggests that learning an internal representation of gaze, aligned with the visual backbone and integrated through a trainable transformation, enables more stable and effective incorporation of gaze information, especially for this type of architecture.

Across scenarios, improvements were more pronounced in cognition scores than in recognition scores. In several cases, recognition slightly decreased while reasoning performance improved. This pattern suggests that the gaze-guided adaptation may encourage more selective and semantically structured descriptions, supporting higher-level reasoning rather than simple recognition tasks such as entity listing. However the effectiveness of gaze augmentation strongly depends on the architectural characteristics of the underlying model.

Future work could focus on addressing the limitations of the current recall-based cognitive evaluation defined by CogBench, while also exploring gaze representations that incorporate temporal information, the use of synthetic scanpath data in VLM architectures, and more complex attention alignment strategies between human gaze signals and model attention mechanisms.

Overall, these results indicate that human gaze can support high-level visual reasoning in Vision-Language Models, provided that its integration is learned and aligned with the model's internal feature representations.

Bibliography

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] P. Anderson, B. Fernando, M. Johnson, and S. Gould. SPICE: semantic propositional image caption evaluation. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*, volume 9909 of *Lecture Notes in Computer Science*, pages 382–398. Springer, 2016. doi: 10.1007/978-3-319-46454-1_24. URL https://doi.org/10.1007/978-3-319-46454-1_24.

[3] Artificial Analysis. Model comparison: Gemini 2.5 flash vs gpt-4 turbo. https://artificialanalysis.ai/models/comparisons/gemini-2-5-flash-vs-gpt-4-turbo, 2026. Accessed: 20 February 2026.

[4] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, and J. Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *CoRR*, abs/2308.12966, 2023. doi: 10.48550/ARXIV.2308.12966. URL https://doi.org/10.48550/arXiv.2308.12966.

[5] S. Bai, K. Chen, X. Liu, et al. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

[6] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[7] D. Caffagni, F. Cocchi, L. Barsellotti, N. Moratelli, S. Sarto, L. Baraldi, L. Baraldi, M. Cornia, and R. Cucchiara. The revolution of multimodal large language models: A survey. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13590–13618, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.807. URL https://aclanthology.org/2024.findings-acl.807/.

[8] Y. Chen, Z. Yang, S. Ahn, D. Samaras, M. Hoai, and G. Zelinsky. Coco-search18 fixation dataset for predicting goal-directed attention control. *Scientific reports*, 11 (1):8776, 2021.

[9] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, and I. Stoica. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. `https://vicuna.lmsys.org/`, 2023.

[10] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL `http://papers.nips.cc/paper_files/paper/2023/hash/9a6a435e75419a836fe47ab6793623e6-Abstract-Conference.html`.

[11] A. Das, H. Agrawal, L. Zitnick, D. Parikh, and D. Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? In J. Su, X. Carreras, and K. Duh, editors, *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 932–937. The Association for Computational Linguistics, 2016. doi: 10.18653/V1/D16-1092. URL `https://doi.org/10.18653/v1/d16-1092`.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

[13] A. T. Duchowski. *Eye tracking methodology: Theory and Practice*. Springer, 2017.

[14] Y. Fang, W. Wang, B. Xie, J. Sun, X. Wang, et al. Eva-02: A visual representation for enhanced vision-language alignment, 2023. URL `https://arxiv.org/abs/2303.11331`.

[15] C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, Z. Qiu, W. Lin, J. Yang, X. Zheng, K. Li, X. Sun, and R. Ji. MME: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, abs/2306.13394, 2023. doi: 10.48550/ARXIV.2306.13394. URL `https://doi.org/10.48550/arXiv.2306.13394`.

[16] H. Goodglass, E. Kaplan, and B. Barresi. *Boston Diagnostic Aphasia Examination*. Lippincott Williams & Wilkins, Philadelphia, 3 edition, 2001.

[17] Google. Gemini 2.5 flash model documentation, 2024. URL https://ai.google.dev/gemini-api/docs/models/gemini-2.5-flash.

[18] Google DeepMind. Gemini 3.1 pro model card. https://deepmind.google/models/model-cards/gemini-3-1-pro/, 2026. Latest natively multimodal model with advanced reasoning capabilities.

[19] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 6325–6334. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.670. URL https://doi.org/10.1109/CVPR.2017.670.

[20] Z. Han, C. Gao, J. Liu, J. Zhang, and S. Q. Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey. *Trans. Mach. Learn. Res.*, 2024, 2024. URL https://openreview.net/forum?id=lIsCS8b6zj.

[21] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

[22] S. He, H. R. Tavakoli, A. Borji, and N. Pugeault. Human attention in image captioning: Dataset and analysis. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 8528–8537. IEEE, 2019. doi: 10.1109/ICCV.2019.00862. URL https://doi.org/10.1109/ICCV.2019.00862.

[23] J. M. Henderson. Human gaze control during real-world scene perception. *Trends in cognitive sciences*, 7(11):498–504, 2003.

[24] J. M. Henderson. Regarding scenes. *Current directions in psychological science*, 16(4):219–222, 2007. URL https://journals.sagepub.com/doi/pdf/10.1111/j.1467-8721.2007.00507.x?casa_token=S6dv4iiCtbgAAAAA:i1OqGXrHn8C5wPh_xhiRfd0GZpiZyMbthvD7ZcX8carIjEFmLFuOTnVNijDIOtIASqbRKq1QOjc.

[25] G. Hinton, O. Vinyals, and J. Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[26] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

[27] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL `https://openreview.net/forum?id=nZeVKeeFYf9`.

[28] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

[29] M. Jiang, S. Huang, J. Duan, and Q. Zhao. Salicon: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1072–1080, 2015.

[30] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *2009 IEEE 12th international conference on computer vision*, pages 2106–2113. IEEE, 2009. URL `https://people.csail.mit.edu/torralba/publications/wherepeoplelook.pdf`.

[31] B. Li, Y. Ge, Y. Ge, G. Wang, R. Wang, R. Zhang, and Y. Shan. Seed-bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308, 2024.

[32] B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu, and C. Li. Llava-onevision: Easy visual task transfer. *Trans. Mach. Learn. Res.*, 2025, 2025. URL `https://openreview.net/forum?id=zKv8qULV6n`.

[33] J. Li, D. Li, S. Savarese, and S. Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/li23q.html`.

[34] C.-Y. Lin. Rouge: A package for automatic evaluation of summaries. In *ACL Workshop on Text Summarization Branches Out*, 2004.

[35] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[36] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10*

- *16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

[37] H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26286–26296. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02484. URL https://doi.org/10.1109/CVPR52733.2024.02484.

[38] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

[39] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, K. Chen, and D. Lin. Mmbench: Is your multi-modal model an all-around player? In A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, editors, *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VI*, volume 15064 of *Lecture Notes in Computer Science*, pages 216–233. Springer, 2024. doi: 10.1007/978-3-031-72658-3_13. URL https://doi.org/10.1007/978-3-031-72658-3_13.

[40] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986, 2022. doi: 10.1109/CVPR52688.2022.01167.

[41] LMSYS. Chatbot arena leaderboard. https://huggingface.co/spaces/lmarena-ai/arena-leaderboard, 2026. Accessed: 20 February 2026.

[42] MosaicML NLP Team. Introducing mpt-7b: A new standard for open-source, commercially usable llms, June 2023. URL https://www.mosaicml.com/blog/mpt-7b.

[43] D. Noton and L. Stark. Scanpaths in eye movements during pattern perception. *Science*, 171(3968):308–311, 1971.

[44] OpenAI. Gpt-5.2 pro model card. https://developers.openai.com/api/docs/models/gpt-5.2-pro, 2025. Vision-enabled multimodal model with advanced reasoning and long context support.

[45] A. Pani and Y. Yang. Gaze-vlm:bridging gaze and vlms through attention regularization for egocentric understanding. *CoRR*, abs/2510.21356, 2025. doi: 10.48550/ARXIV.2510.21356. URL https://doi.org/10.48550/arXiv.2510.21356.

[46] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[47] J. S. Park, C. Bhagavatula, R. Mottaghi, A. Farhadi, and Y. Choi. Visualcomet: Reasoning about the dynamic context of a still image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[48] T. Qiao, J. Dong, and D. Xu. Exploring human-like attention supervision in visual question answering. In S. A. McIlraith and K. Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 7300–7307. AAAI Press, 2018. doi: 10.1609/AAAI.V32I1.12272. URL https://doi.org/10.1609/aaai.v32i1.12272.

[49] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

[50] K. Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372, 1998. URL http://andrewd.ces.clemson.edu/courses/cpsc881/papers/reading/Ray98_readingSurvey.pdf.

[51] X. Song, M. Wu, K. Q. Zhu, C. Zhang, and Y. Chen. A cognitive evaluation benchmark of image reasoning and description for large vision-language models. In L. Chiruzzo, A. Ritter, and L. Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6392–6409, Albuquerque, New Mexico, Apr. 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.324. URL https://aclanthology.org/2025.naacl-long.324/.

[52] E. Sood, F. Kögel, P. Müller, D. Thomas, M. Bâce, and A. Bulling. Multimodal integration of human-like attention in visual question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023 - Workshops, Vancouver, BC, Canada, June 17-24, 2023*, pages 2648–2658. IEEE, 2023. doi: 10.1109/CVPRW59228.2023.00265. URL https://doi.org/10.1109/CVPRW59228.2023.00265.

[53] Y. Sugano and A. Bulling. Seeing with humans: Gaze-assisted neural image captioning. *arXiv preprint arXiv:1608.05203*, 2016.

[54] M. Tan and Q. V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2019. URL http://proceedings.mlr.press/v97/tan19a.html.

[55] B. W. Tatler, N. J. Wade, H. Kwan, J. M. Findlay, and B. M. Velichkovsky. Yarbus, eye movements, and vision. *i-Perception*, 1(1):7–27, 2010.

[56] Tobii AB. Display area coordinate system (dacs), 2023. URL `https://connect.tobii.com/s/article/Display-Area-Coordinate-System-DACS`.

[57] Tobii AB. *Tobii Pro Lab User Manual*. Tobii AB, 2023. URL `http://andrewd.ces.clemson.edu/courses/cpsc881/manuals/Tobii/Tobii-Pro-Lab-User-Manual.pdf`. Appendix C: Gaze Filter functions and effects.

[58] Tobii AB. *Tobii Pro Spectrum User Manual*. Tobii AB, v3.2 edition, March 2025. URL `https://go.tobii.com/tobii-pro-spectrum-user-manual`. Section 10.2: Setup.

[59] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[60] R. Vedantam, C. L. Zitnick, and D. Parikh. Collecting image description datasets using crowdsourcing. *CoRR*, abs/1411.3041, 2014. URL `http://arxiv.org/abs/1411.3041`.

[61] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[62] D. S. Wooding. Fixation maps: quantifying eye-movement traces. In *Proceedings of the 2002 symposium on Eye tracking research & applications*, pages 31–36, 2002. URL `https://dl.acm.org/doi/pdf/10.1145/507072.507078`.

[63] K. Yan, Z. Wang, L. Ji, Y. Wang, N. Duan, and S. Ma. Voila-a: Aligning vision-language models with user's gaze attention. *Advances in neural information processing systems*, 37:1890–1918, 2024.

[64] A. Yang, B. Yang, B. Hui, Z. Zheng, B. Zhang, J. Zhou, Q. Li, Z. Wang, D. Yu, X. Li, et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024.

[65] A. L. Yarbus. *Eye Movements and Vision*. Plenum Press, 1967.

[66] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024.

[67] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2:67–78, 2014.

[68] W. Yu, Z. Yang, L. Li, J. Wang, K. Lin, Z. Liu, X. Wang, and L. Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. In R. Salakhutdinov, Z. Kolter, K. A. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, editors, *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*, volume 235 of *Proceedings of Machine Learning Research*, pages 57730–57754. PMLR / OpenReview.net, 2024. URL `https://proceedings.mlr.press/v235/yu24o.html`.

[69] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731, 2019.

[70] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer. Sigmoid loss for language image pre-training, 2023. URL `https://arxiv.org/abs/2303.15343`.

[71] R. Zhang, A. Saran, B. Liu, Y. Zhu, S. Guo, S. Niekum, D. Ballard, and M. Hayhoe. Human gaze assisted artificial intelligence: A review. In *IJCAI: Proceedings of the Conference*, volume 2020, page 4951, 2020.

[72] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2020.