UNIVERSIDADE FEDERAL DE GOIÁS (UFG)

INSTITUTO DE INFORMÁTICA (INF)

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO (PPGCC)

PAULO VITOR SANTANA DA SILVA

# Automated Attention Guidance in Virtual Reality Videos

Goiânia
2026

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

**TERMO DE CIÊNCIA E DE AUTORIZAÇÃO (TECA) PARA DISPONIBILIZAR VERSÕES ELETRÔNICAS DE TESES**

**E DISSERTAÇÕES NA BIBLIOTECA DIGITAL DA UFG**

Na qualidade de titular dos direitos de autor, autorizo a Universidade Federal de Goiás (UFG) a disponibilizar, gratuitamente, por meio da Biblioteca Digital de Teses e Dissertações (BDTD/UFG), regulamentada pela Resolução CEPEC nº 832/2007, sem ressarcimento dos direitos autorais, de acordo com a Lei 9.610/98, o documento conforme permissões assinaladas abaixo, para fins de leitura, impressão e/ou download, a título de divulgação da produção científica brasileira, a partir desta data.

O conteúdo das Teses e Dissertações disponibilizado na BDTD/UFG é de responsabilidade exclusiva do autor. Ao encaminhar o produto final, o autor(a) e o(a) orientador(a) firmam o compromisso de que o trabalho não contém nenhuma violação de quaisquer direitos autorais ou outro direito de terceiros.

**1. Identificação do material bibliográfico**

[ X ] Dissertação [ ] Tese [ ] Outro*:_____

*No caso de mestrado/doutorado profissional, indique o formato do Trabalho de Conclusão de Curso, permitido no documento de área, correspondente ao programa de pós-graduação, orientado pela legislação vigente da CAPES.

**Exemplos**: Estudo de caso ou Revisão sistemática ou outros formatos.

**2. Nome completo do autor**

Paulo Vitor Santana da Silva

**3. Título do trabalho**

Automated Attention Guidance in Virtual Reality Videos

**4. Informações de acesso ao documento (este campo deve ser preenchido pelo orientador)**

Concorda com a liberação total do documento [ X ] SIM [ ] NÃO[1]

**[1]** Neste caso o documento será embargado por até um ano a partir da data de defesa. Após esse período, a possível disponibilização ocorrerá apenas mediante:
**a)** consulta ao(à) autor(a) e ao(à) orientador(a);
**b)** novo Termo de Ciência e de Autorização (TECA) assinado e inserido no arquivo da tese ou dissertação.
O documento não será disponibilizado durante o período de embargo.
Casos de embargo:
- Solicitação de registro de patente;
- Submissão de artigo em revista científica;
- Publicação como capítulo de livro;
- Publicação da dissertação/tese em livro.

**Obs. Este termo deverá ser assinado no SEI pelo orientador e pelo autor.**

Documento assinado eletronicamente por **Arlindo Rodrigues Galvao Filho**, **Professor do Magistério Superior**, em 24/06/2026, às 14:01, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Paulo Vitor Santana Da Silva**, **Discente**, em 25/06/2026, às 15:49, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **6280403** e o código CRC **620CEBFD**.

PAULO VITOR SANTANA DA SILVA

# Automated Attention Guidance in Virtual Reality Videos

Dissertação apresentada ao Programa de Pós–Graduação da Instituto de Informática (INF) da Universidade Federal de Goiás (UFG), como requisito parcial para obtenção do título de Mestre em Ciência da Computação.

**Área de concentração:** Ciência da Computação.

**Linha de Pesquisa:** Sistemas Inteligentes e Aplicações.

**Orientador:** Prof. Dr. Arlindo Rodrigues Galvão Filho

Goiânia
2026

Ficha de identificação da obra elaborada pelo autor, através do Programa de Geração Automática do Sistema de Bibliotecas da UFG.

Santana da Silva, Paulo Vitor

Automated Attention Guidance in Virtual Reality Videos [manuscrito] = Direcionamento automático da atenção em vídeos de realidade virtual / Paulo Vitor Santana da Silva. - 2026.

59 f.: 2026

Orientador: Prof. Dr. Arlindo Rodrigues Galvão Filho

Dissertação (Mestrado) - Universidade Federal de Goiás, Instituto de Informática (INF), Programa de Pós-Graduação em Ciência da Computação, Goiânia, 2026.

Apêndice.

Bibliografia.

Inclui: tabelas, grafico, lista de figuras, lista de tabelas.

1. Videos 360º. 2. Realidade Virtual. 3. Orientação de Atenção. 4. Aprendizado Profundo.
I. Rodrigues Galvão Filho, Arlindo, orient. II. Título.

CDU 004

UNIVERSIDADE FEDERAL DE GOIÁS
INSTITUTO DE INFORMÁTICA

**ATA DE DEFESA DE DISSERTAÇÃO**

Ata nº **12/2026** da sessão de Defesa de Dissertação de **Paulo Vitor Santana da Silva**, que confere o título de Mestre em **Ciência da Computação**, na área de concentração em **Ciência da Computação**.

Aos sete dias do mês de maio de dois mil e vinte e seis, a partir das dezessete horas, na sala 257 do Instituto de Informática, realizou-se a sessão pública de Defesa de Dissertação intitulada "**Automated Attention Guidance in Virtual Reality Videos**". Os trabalhos foram instalados pelo(a) Orientador, Professor Doutor Arlindo Rodrigues Galvão Filho (INF/UFG) com a participação dos demais membros da Banca Examinadora: Dr. Rodrigo Zempulski Fanucchi (CEIA/AKCIT/UFG), membro titular externo, cuja participação ocorreu por videoconferência; Professora Doutora Telma Woerle de Lima Soares (INF/UFG), membra titular interna. Durante a arguição os membros da banca não fizeram sugestão de alteração do título do trabalho. A Banca Examinadora reuniu-se em sessão secreta a fim de concluir o julgamento da Dissertação, tendo sido o candidato **aprovado** pelos seus membros. Proclamados os resultados pelo Professor Doutor Arlindo Rodrigues Galvão Filho, Presidente da Banca Examinadora, foram encerrados os trabalhos e, para constar, lavrou-se a presente ata que é assinada pelos Membros da Banca Examinadora, aos sete dias do mês de maio de dois mil e vinte e seis.

TÍTULO SUGERIDO PELA BANCA

Documento assinado eletronicamente por **Telma Woerle De Lima Soares**, **Professora do Magistério Superior**, em 07/05/2026, às 18:43, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **RODRIGO ZEMPULSKI FANUCCHI**, **Usuário Externo**, em 08/05/2026, às 10:02, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Arlindo Rodrigues Galvao Filho**, **Professor do Magistério Superior**, em 11/05/2026, às 13:19, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Paulo Vitor Santana Da Silva**, **Discente**, em 11/05/2026, às 15:57, conforme horário oficial de Brasília, com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site https://sei.ufg.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0, informando o código verificador **6166817** e o código CRC **FFF77444**.

**Referência:** Processo nº 23070.018424/2026-40 SEI nº 6166817

# Agradecimentos

This work has been fully/partially funded by the project Research and Development of Algorithms for Construction of Digital Human Technological Components supported by Advanced Knowledge Center in Immersive Technologies (AKCIT), with financial resources from the PPI IoT/Manufatura 4.0 / PPI HardwareBR of the MCTI grant number 057/2023, signed with EMBRAPII.

# Resumo

Silva, Paulo Vitor. **Direcionamento Automático da Atenção em Vídeos de Realidade Virtual**. Goiânia, 2026. 59p. Dissertação de Mestrado. Programa de Pós-Graduação em Ciência da Computação (PPGCC), Instituto de Informática (INF), Universidade Federal de Goiás (UFG).

A natureza omnidirecional dos vídeos em Realidade Virtual (RV) 360º proporciona aos usuários uma experiência altamente imersiva, mas também cria desafios relacionados à atenção e à orientação, uma vez que elementos importantes da cena podem passar despercebidos. Esta dissertação investiga mecanismos automatizados de direcionamento de atenção em vídeos imersivos de RV 360º por meio da integração de Processamento de Linguagem Natural, Visão Computacional e efeitos visuais adaptativos. A pesquisa é composta por três estudos interconectados. O primeiro explora o uso de roteiros em linguagem natural, detecção de objetos e segmentação para o direcionamento automatizado da atenção. O segundo introduz o Focus360, uma arquitetura que combina interpretação de roteiros, detecção de objetos, rastreamento de objetos e efeitos visuais para aumentar a robustez do direcionamento de atenção. O terceiro avalia diferentes arquiteturas de detecção e rastreamento de objetos utilizando métricas quantitativas e avaliações qualitativas. Os resultados demonstram a viabilidade de converter descrições em linguagem natural em pistas visuais de atenção e evidenciam o potencial da combinação de técnicas de visão computacional e efeitos visuais adaptativos para apoiar a orientação dos usuários em ambientes imersivos. Esta dissertação contribui para o desenvolvimento de sistemas inteligentes de direcionamento de atenção para experiências em RV 360º.

**Palavras–chave**

Videos 360º, Realidade Virtual, Orientação de Atenção, Aprendizado Profundo.

# Abstract

Silva, Paulo Vitor. **Automated Attention Guidance in Virtual Reality Videos**. Goiânia, 2026. 59p. MSc. Dissertation. Programa de Pós-Graduação em Ciência da Computação (PPGCC), Instituto de Informática (INF), Universidade Federal de Goiás (UFG).

The omnidirectional nature of 360º Virtual Reality (VR) videos offers users a highly immersive experience but also creates challenges related to attention and orientation, as important scene elements may be overlooked. This dissertation investigates automated attention guidance in immersive 360º VR videos through the integration of Natural Language Processing, Computer Vision, and adaptive visual effects. The research comprises three interconnected studies. The first explores the use of natural language video roadmaps, object detection, and segmentation for automated attention guidance. The second introduces Focus360, an architecture that combines roadmap interpretation, object detection, object tracking, and visual effects to improve guidance robustness. The third evaluates different object detection and tracking architectures using quantitative metrics and qualitative assessments. The results demonstrate the feasibility of converting natural language descriptions into visual attention cues and highlight the potential of combining computer vision techniques and adaptive visual effects to support user orientation in immersive environments. This dissertation contributes to the development of intelligent attention guidance systems for 360º VR experiences.

**Keywords**

360º Videos, Virtual Reality, Attention Guidance, Deep Learning.

# Contents

# List of Figures

# List of Tables

CHAPTER 1

# Introduction

The rapid evolution of Virtual Reality (VR) technologies has transformed how users interact with digital content, enabling increasingly immersive experiences across domains such as entertainment, education and tourism [10, 17, 4]. Among the different forms of immersive media, 360º videos have become one of the most accessible and widespread formats, allowing users to freely explore virtual environments through omnidirectional visual content. Unlike traditional videos, where the camera framing explicitly controls the viewer's focus, 360º VR videos provide complete freedom of observation, creating a stronger sense of presence and immersion.

Although this freedom is one of the main advantages of immersive media, it also introduces a significant challenge. Since users can freely direct their gaze toward any region of the environment, they may fail to observe elements that are relevant to the narrative, educational content, or intended experience. As a consequence, important information may be overlooked, narrative continuity may be disrupted, and the effectiveness of the immersive experience may be reduced [2, 11]. This challenge has motivated extensive research on attention guidance mechanisms capable of directing users toward specific regions of interest while preserving immersion.

Several techniques have been proposed to guide attention in immersive environments. Existing approaches commonly employ visual cues such as arrows, blur effects, color manipulation, darkening effects, and other visual stimuli designed to attract the user's focus toward specific targets [19, 3, 20]. Although these techniques have demonstrated promising results, they are often manually designed and require content creators to explicitly define both the regions of interest and the mechanisms used to attract attention. As the production of immersive content continues to grow, the development of automated attention guidance systems becomes increasingly important to reduce authoring effort and enable scalable solutions.

Recent advances in Artificial Intelligence have created new opportunities for addressing this problem. Open-vocabulary object detection models are capable of locating objects based on natural language descriptions, segmentation and tracking models can maintain object localization throughout video sequences, and Large Language Models

(LLMs) can interpret textual instructions and transform them into structured information. The integration of these technologies enables the development of intelligent systems capable of automatically identifying relevant elements within immersive scenes and dynamically guiding users' attention toward them.

Despite these advances, several challenges remain open. First, attention guidance systems must be capable of understanding high-level descriptions of the content and converting them into actionable representations. Second, objects described in natural language must be accurately detected and reliably tracked throughout the video. Third, attention guidance mechanisms must remain effective even when users are looking in directions opposite to the intended target. Finally, the relative effectiveness of different object detection and tracking architectures for automated attention guidance in immersive environments remains largely unexplored.

This dissertation investigates the problem of automated attention guidance in 360º Virtual Reality videos through the integration of Natural Language Processing, Computer Vision, and adaptive visual effects. The central hypothesis of this work is that the combination of natural language understanding, open-vocabulary object detection, object tracking, and adaptive visual effects can provide an effective mechanism for automatically directing users' attention toward relevant elements in immersive environments.

The general objective of this dissertation is **to investigate and develop automated attention guidance mechanisms for immersive 360º Virtual Reality videos**.

To achieve this objective, the following specific objectives are defined:

- Transform natural language descriptions of video roadmaps into structured representations of regions of interest;
- Automatically identify target objects in immersive scenes using open-vocabulary object detection techniques;
- Track relevant objects throughout video sequences using segmentation-based tracking approaches;
- Develop visual attention guidance mechanisms capable of directing user focus without compromising immersion;
- Evaluate different object detection architectures for automated attention guidance tasks;
- Evaluate different object tracking architectures and analyze their suitability for immersive environments;
- Investigate the effectiveness of integrating these components into a complete automated attention guidance pipeline.

To address these objectives, this dissertation is organized around three interconnected studies that progressively investigate different aspects of the proposed problem.

The first study explores the feasibility of automatically guiding user attention through the integration of video scripts and deep learning models [16]. The proposed approach employs Grounding DINO [9] for object detection and Segment Anything Model (SAM) [13] for object segmentation, using a vignette effect to highlight target elements described in a predefined script. This initial investigation demonstrates the viability of combining natural language descriptions with computer vision techniques to automatically identify and emphasize relevant objects in immersive environments.

The second study builds upon the limitations identified in the first investigation. In particular, it addresses scenarios in which users are significantly disoriented or looking in directions opposite to the intended target. To overcome these limitations, a more comprehensive attention guidance system named *Focus360* is proposed [14]. The new architecture incorporates Llama 3 [5] for prompt processing, SAM 2 [13] for object tracking, and a combination of visual effects including Blur, Fade to Gray, Radial Darkening, and Halo Darkening. This evolution results in a more robust and adaptable mechanism for directing user attention under challenging viewing conditions.

The third study investigates the computational vision components responsible for object detection and tracking [15]. While the previous studies relied on specific models, the impact of alternative architectures remained largely unexplored. Therefore, this work performs a systematic evaluation of multiple object detection and object tracking models within the proposed attention guidance pipeline. The analysis combines quantitative metrics and qualitative assessments conducted through a Multimodal Large Language Model acting as a judge, enabling the identification of the most effective model combinations for immersive attention guidance applications.

Together, these three studies provide a comprehensive investigation of automated attention guidance mechanisms for immersive VR videos. The research progresses from an initial proof-of-concept implementation to the development of a complete attention guidance system and culminates in a systematic comparative evaluation of its core computational vision components. By integrating advances in Natural Language Processing, Computer Vision, and immersive media technologies, this dissertation contributes to the development of intelligent systems capable of improving user engagement, comprehension, and narrative coherence in immersive virtual environments.

Dissertation structure:

- **Chapter 2: Related Works** presents the theoretical background and related literature concerning attention guidance in virtual reality environments, open-vocabulary object detection, and object tracking techniques.
- **Chapter 3: Attention Guidance through Video Script: A Case Study of Object Focusing on 360º VR Video Tours** presents the first study conducted in this research. This work investigates the feasibility of automatically guiding user attention

through the integration of video scripts, object detection, and object segmentation techniques. The study was published as a short paper in *SVR '24: Proceedings of the 26th Symposium on Virtual and Augmented Reality (SVR 2024)* and was awarded **best paper** in the short paper category.

- **Chapter 4: Focus360: Guiding User Attention in Immersive Videos for VR** presents the second study developed in this dissertation. Building upon the limitations identified in the previous investigation, this work introduces a more comprehensive attention guidance system based on object tracking and multiple visual effects. This work was published as a demonstration in the *2025 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW) (IEEE VR 2025)*.
- **Chapter 5: Automated Attention Guidance in Virtual Reality Videos** presents the third study conducted in this research. This work performs a systematic evaluation of multiple object detection and object tracking architectures within the proposed attention guidance pipeline, identifying the most effective combinations for immersive virtual reality environments. This work was published as a full paper in *SVR '25: Proceedings of the 27th Symposium on Virtual and Augmented Reality (SVR 2025)*.
- **Chapter 6: Conclusion and Future Work** summarizes the main findings obtained throughout the three studies, discusses the overall contributions of this dissertation, highlights its limitations, and presents directions for future research.

CHAPTER 2

# Related Works

## 2.1 Attention Guidance in Virtual Reality

Attention guidance refers to the process of directing a user's focus toward relevant elements within a visual scene. In immersive Virtual Reality (VR) environments, this task becomes particularly challenging due to the omnidirectional nature of the experience, which allows users to freely explore the surrounding environment. While this freedom contributes to a stronger sense of presence and immersion, it also increases the likelihood that users may overlook important narrative elements, educational content, or points of interest [2, 11].

Several studies have investigated mechanisms for guiding user attention in immersive environments through the use of visual cues. Wallgrün et al. [19] evaluated three different guidance mechanisms, namely arrows, butterflies, and radar indicators, in educational VR tours. Their results demonstrated that all guidance mechanisms improved user performance compared to scenarios without guidance, with arrow-based cues achieving the highest user preference.

Danieau et al. [3] investigated the effectiveness of different visual effects for directing user attention in 360º videos, including fade-to-black, blur, desaturation and deformation, as shown in Figure 2.1. Their study showed that gradual visual modifications can successfully attract user attention, although directing users toward regions outside their current field of view remains challenging. Similarly, Woodworth et al. [20] conducted a comprehensive evaluation of nine visual guidance and restoration cues, analyzing their effectiveness in both directing attention and recovering user focus after distraction events.

Hillaire et al. [6] proposed the use of depth-of-field effects and camera motion adaptations inspired by human visual perception. Their results indicated that dynamically adapting visual effects according to the user's focus point can improve the overall immersive experience.

Although these approaches have demonstrated promising results, most of them rely on manually defined regions of interest and handcrafted guidance strategies. Consequently, their applicability becomes limited when dealing with large volumes of im-

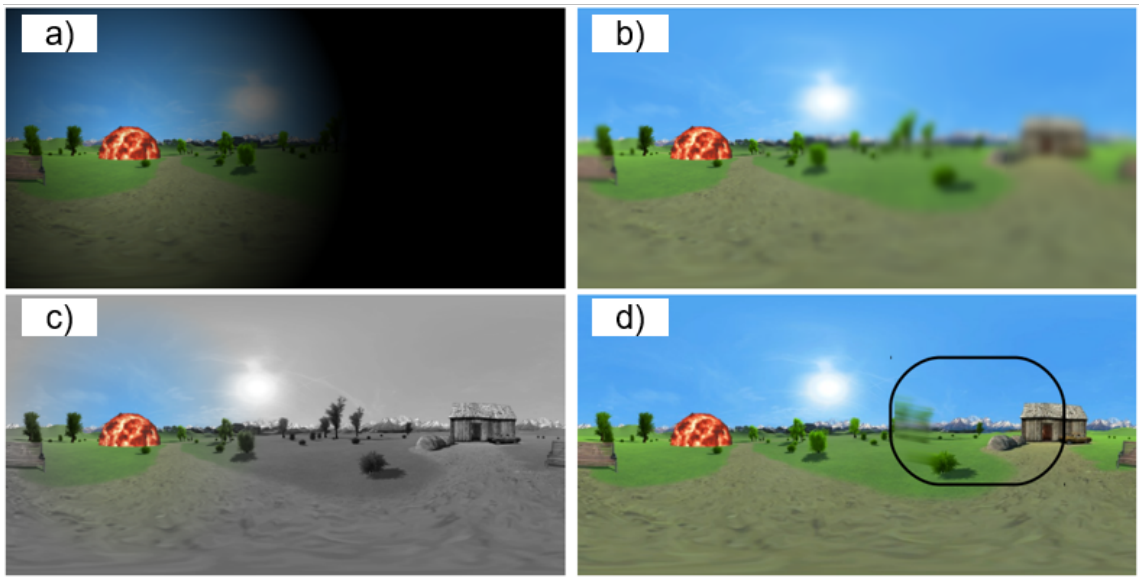



**Figure 2.1:** *The four effects proposed by Danieau et al. to guide the attention to the explosion. a) fade-to-black, b) blur, c) desaturation and d) deformation.*

mersive content, motivating the development of automated attention guidance systems capable of identifying relevant elements without human intervention.

## 2.2 Open-Vocabulary Object Detection

The automation of attention guidance requires the capability to identify objects of interest based on high-level semantic descriptions. Traditional object detection models are constrained to predefined categories learned during training, limiting their applicability in dynamic environments where the target objects may vary according to the content and user objectives.

Recent advances in open-vocabulary object detection have addressed this limitation by enabling object localization through natural language descriptions. Among the most influential approaches is Grounding DINO [9], which extends the DINO object detector through the integration of visual and textual representations. The model employs a feature enhancement module, language-guided query selection, and a cross-modality decoder to align image regions with arbitrary textual descriptions as illustrated in Figure. 2.2. Due to its strong zero-shot performance, Grounding DINO has become one of the most widely adopted models for phrase grounding and text-guided object detection.

Another important approach is OWLv2 [12], which adopts a data-centric strategy based on large-scale multimodal pretraining. Built upon the CLIP architecture, OWLv2 supports both text-guided and image-guided object detection through a shared multimodal embedding space. Its self-training strategy enables improved generalization to unseen categories, making it particularly suitable for open-world scenarios.

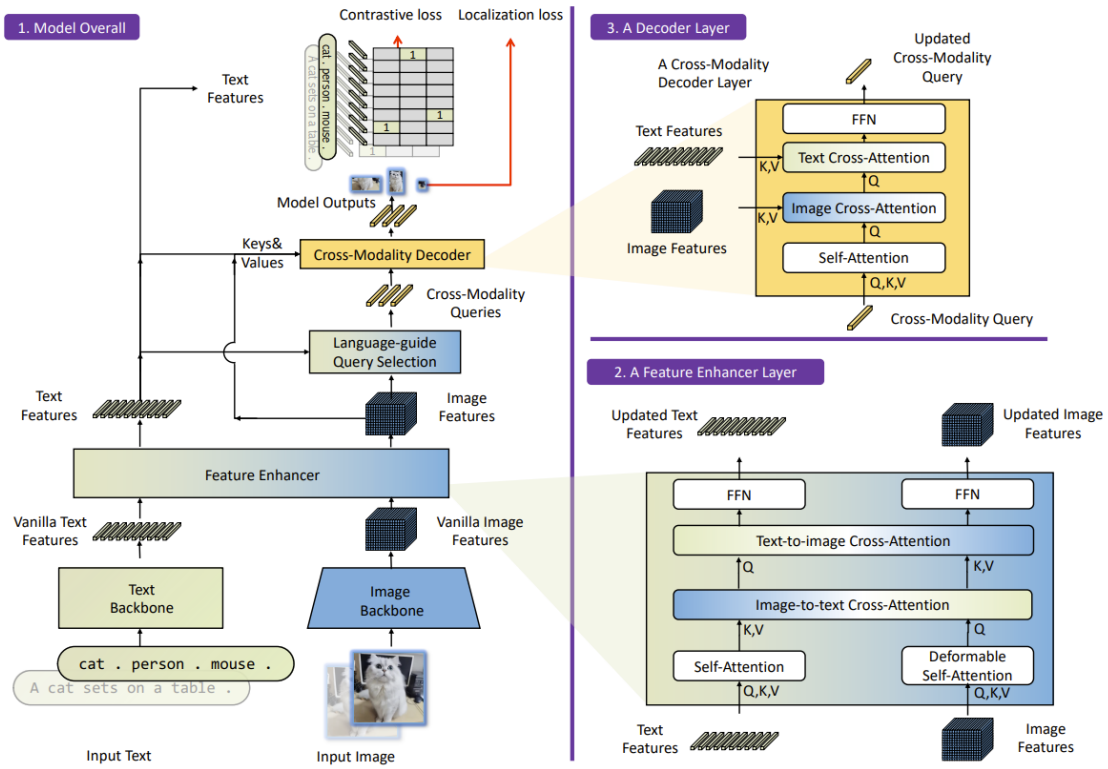



**Figure 2.2:** *Grounding DINO Architecture, extracted from [9].*

More recently, Cheng et al. [1] proposed YOLO-World, an extension of the YOLO architecture designed for open-vocabulary object detection. By incorporating vision-language representations through the RepVL-PAN module, YOLO-World combines the flexibility of text-guided detection with the efficiency traditionally associated with YOLO-based models, enabling real-time performance while maintaining competitive detection accuracy.

These developments have significantly expanded the applicability of object detection systems, making it possible to identify arbitrary objects described through natural language, a fundamental requirement for automated attention guidance pipelines.

## 2.3 Object Tracking and Video Segmentation

After identifying the target object, attention guidance systems must maintain its localization throughout the video sequence. This requirement makes object tracking and video segmentation fundamental components of automated attention guidance systems.

The Segment Anything Model (SAM) proposed by Kirillov et al. [8] represented a major advance in image segmentation by introducing a foundation model capable of segmenting arbitrary objects through prompts such as points, masks, and bounding boxes. Trained on a large-scale dataset containing billions of segmentation masks, SAM

demonstrated remarkable generalization capabilities and established a new paradigm for prompt-based segmentation.

Building upon SAM, Ravi et al. [13] introduced SAM 2, extending the original framework to support video processing and object tracking. Through the incorporation of memory mechanisms and temporal attention modules, SAM 2 enables the propagation of object masks across video frames while preserving segmentation quality. The overall architecture of SAM 2 is illustrated in Fig. 2.3.

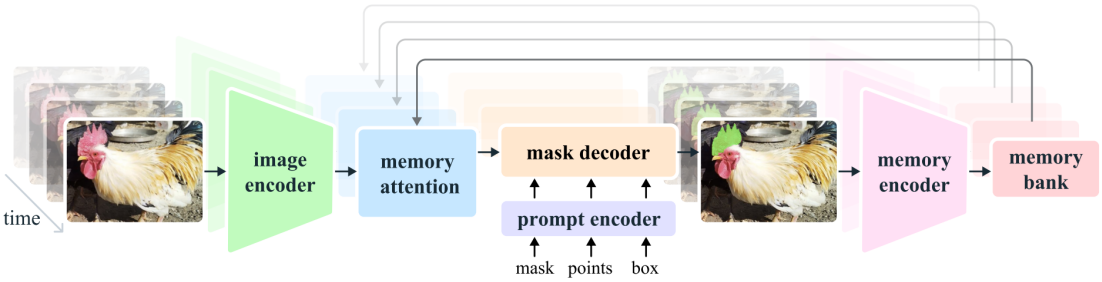



**Figure 2.3:** *SAM 2 architecture, extracted from [13].*

Several extensions have been proposed to improve specific aspects of SAM-based tracking. HQ-SAM 2 extends the original architecture by introducing mechanisms designed to generate higher-quality segmentation masks and finer object boundaries [7]. EfficientTAM [21] focuses on reducing computational complexity, enabling efficient object tracking while maintaining competitive performance. DAM4SAM [18] introduces memory management strategies aimed at improving robustness in challenging scenarios involving distractors and long-term tracking.

Together, these approaches provide a diverse set of alternatives for object tracking and video segmentation, each offering different trade-offs between segmentation quality, tracking robustness, and computational efficiency.

## 2.4 Research Gap and Motivation

The literature demonstrates significant advances in both attention guidance mechanisms and computer vision techniques for object detection and tracking. Existing studies have shown that visual cues can effectively direct user attention in immersive environments, while recent developments in Artificial Intelligence have enabled robust open-vocabulary object detection and video object tracking.

However, most attention guidance approaches still rely on manually specified regions of interest and handcrafted authoring processes. Conversely, research on object detection and tracking has largely focused on visual understanding tasks without explicitly addressing the problem of attention guidance in immersive environments.

Furthermore, despite the availability of multiple open-vocabulary detection and tracking architectures, their suitability for automated attention guidance in 360º VR videos remains largely unexplored. In particular, there is a lack of studies investigating how natural language descriptions can be transformed into complete attention guidance pipelines capable of automatically identifying, tracking, and highlighting relevant objects throughout immersive experiences.

To address these limitations, this dissertation investigates the integration of Natural Language Processing, open-vocabulary object detection, object tracking, and adaptive visual effects into automated attention guidance systems for immersive 360º VR videos. The following chapters present three interconnected studies that progressively explore, refine, and evaluate different aspects of this problem.

CHAPTER 3

# Study I: Attention Guidance through Video Script: A Case Study of Object Focusing on 360º VR Video Tours (SVR 2024)

This chapter presents the first study conducted within this dissertation, originally published as a short paper in the *26th Symposium on Virtual and Augmented Reality (SVR 2024)*, where it received the **Best Short Paper Award**. The goal of this study was to investigate the feasibility of automatically guiding user attention in immersive 360º Virtual Reality videos through the integration of natural language descriptions and computer vision techniques.

As discussed in Chapter 2, existing attention guidance approaches typically rely on manually specified regions of interest and handcrafted guidance mechanisms. While effective, these approaches require content creators to explicitly define both the target elements and the visual cues used to attract user attention. The increasing availability of large-scale vision-language models creates new opportunities for automating this process.

The central hypothesis investigated in this study is that textual descriptions of a video's narrative can be used to automatically identify relevant objects within immersive scenes and subsequently guide user attention toward them. To evaluate this hypothesis, an attention guidance pipeline was developed by combining an open-vocabulary object detector with an object segmentation model and a visual highlighting mechanism.

This study constitutes the first step toward the development of automated attention guidance systems for immersive environments. Rather than focusing on the comparison of different architectures, the objective was to demonstrate the viability of integrating natural language descriptions, object detection, and object segmentation into a unified pipeline capable of directing user attention to relevant regions of interest.

The following sections describe the proposed methodology, experimental setup, obtained results, and the limitations identified during this initial investigation, which motivated the developments presented in the subsequent studies.

## 3.1 Material and Methods

### 3.1.1 Case Study

To evaluate the proposed approach, a 360º video tour of the University of Reading campus was selected as a case study. The video contains both indoor and outdoor environments and allows users to freely explore the scene while being transported through different locations on campus, as shown in Figure 3.1. Throughout the tour, there might be more than one region of interest that the user's attention should be directed.


**Figure 3.1:** *Examples of different environments present in the 360º virtual tour used in this study.*

The proposed pipeline uses these textual descriptions as guidance information to automatically identify and highlight the corresponding objects within the immersive scene.

### 3.1.2 Attention Guidance Pipeline

The proposed attention guidance pipeline establishes a connection between the narrative content of the video and the visual elements present in the scene. Instead of manually defining regions of interest, the system uses textual descriptions extracted from the video roadmap to automatically identify relevant objects and generate visual cues that attract the user's attention.

The workflow consists of three main stages. First, a textual description associated with a specific moment of the video is provided as input together with the corresponding

video frame. Next, an open-vocabulary object detection model locates the object referenced in the description. Finally, the detected object is segmented and a visual highlighting effect is applied to emphasize the target region within the immersive environment. Figure 3.2 illustrates the complete workflow adopted in this study.

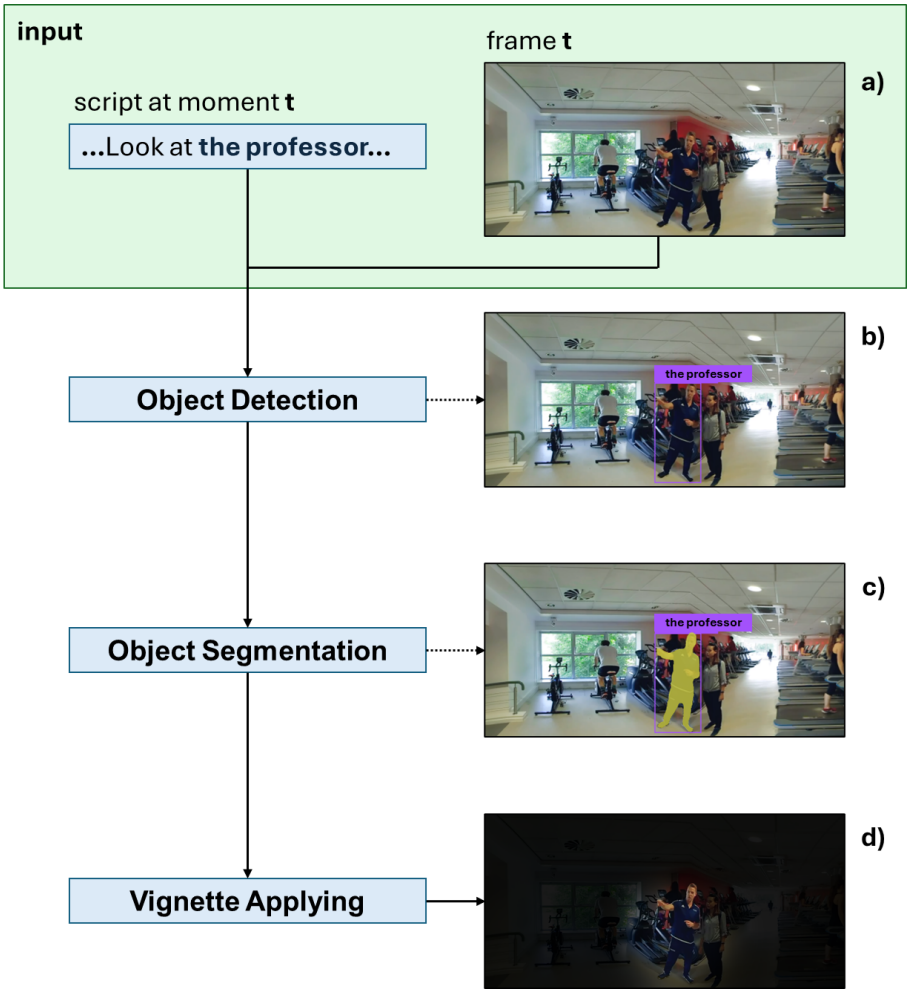



**Figure 3.2:** *Overview of the proposed attention guidance pipeline.*

Given a textual description extracted from the roadmap, the object detector identifies the most relevant region within the scene. The resulting bounding box is subsequently used as a prompt for the segmentation model. Finally, the generated segmentation mask is used to apply a visual highlighting effect that directs the user's attention toward the target object.

### 3.1.3 Object Detection

The object detection stage is responsible for locating the object described in the roadmap within the current video frame. This study employs Grounding DINO as the detection model. As discussed in Chapter 2, Grounding DINO is an open-vocabulary object detector capable of localizing arbitrary objects through natural language descriptions.

This capability makes it particularly suitable for attention guidance applications, where the target objects vary according to the narrative context.

For each frame, the textual description and image are provided as input to the model. The bounding box associated with the highest-confidence prediction is selected as the target region of interest and forwarded to the segmentation stage.

### 3.1.4 Object Segmentation

After object localization, the selected bounding box is used as a prompt for object segmentation. The Segment Anything Model (SAM) is employed to generate a pixel-level representation of the detected object. SAM receives both the image and the bounding box coordinates returned by Grounding DINO and produces a segmentation mask corresponding to the target object. The resulting mask provides a more accurate representation of the region of interest than a simple bounding box and enables the application of localized visual effects.

### 3.1.5 Vignette Applying

The final stage of the pipeline consists of applying a visual effect that directs the user's attention toward the segmented object. A vignette effect was selected due to its simplicity and low visual intrusiveness. The effect darkens regions outside the segmented area while preserving the visibility of the target object. By reducing the visual saliency of the surrounding environment, the vignette effect encourages users to focus on the region indicated by the roadmap.

## 3.2 Results

Object detection using Grounding Dino was made using zero-shot with a box-threshold equals 0.3 and text-threshold equals 0.25, as recommended by the authors. Although the model may return different bounding boxes, only the one with the highest confidence is considered. It was used the checkpoint ViT-H of the SAM model for object segmentation and the zero shot approach was followed.

On Figures 3.3 and 3.4 are shown two moments (1 and 2) from scenes at Reading University, presenting the cafe-lounge and the museum, respectively. According to the video script, at moment (1) in Figure 3.3, attention should be on "The cafe-lounge," and at moment (2), on "The cars between the trees." Similarly, for Figure 3.4, attention should be on "The sculpture of a person on the right side" at moment (1), and on "The sculpture of a centaur on the left side" at moment (2). Object detection and segmentation were successfully performed for both scenes, as shown in Figures 3.3 (1-b), 3.3 (2-b), and 3.4


**Figure 3.3:** *Different moments of the scene showing cafe-lounge on the video tour. It is defined on video script to "Look at the cafe lounge" on moment (1) and "Look at the cars between the trees" on moment (2). (a) The original frame of the video. (b) The object described on the script detected and segmented. (c) The target object with the vignette effect applied.*

(1-b). The vignette effect effectively directed attention to the respective target elements, as seen in Figures 3.3 (1-c), 3.3 (2-c), and 3.4 (1-c). However, in Figure 3.4 (2-b), SAM failed to correctly segment the sculpture of the centaur, resulting in the vignette effect partially obscuring the target element.

## 3.3 Discussion

The results demonstrate the feasibility of integrating natural language descriptions, open-vocabulary object detection, and object segmentation into a unified attention guidance pipeline. The proposed approach successfully transforms narrative descriptions into visual attention cues, enabling the automatic identification and highlighting of relevant objects within immersive scenes.

One of the main contributions of this study is the demonstration that video roadmaps can serve as an effective source of semantic information for attention guidance. Instead of manually specifying regions of interest, content creators can define attention targets through natural language descriptions that are automatically interpreted by the system. The combination of Grounding DINO and SAM proved capable of identifying and segmenting objects described in the roadmap, generating visual cues that emphasize relevant elements of the scene.

The experiments conducted on a 360º virtual tour further demonstrate that recent advances in vision-language models can be leveraged to automate important stages of the attention guidance process. By integrating object detection, segmentation, and visual

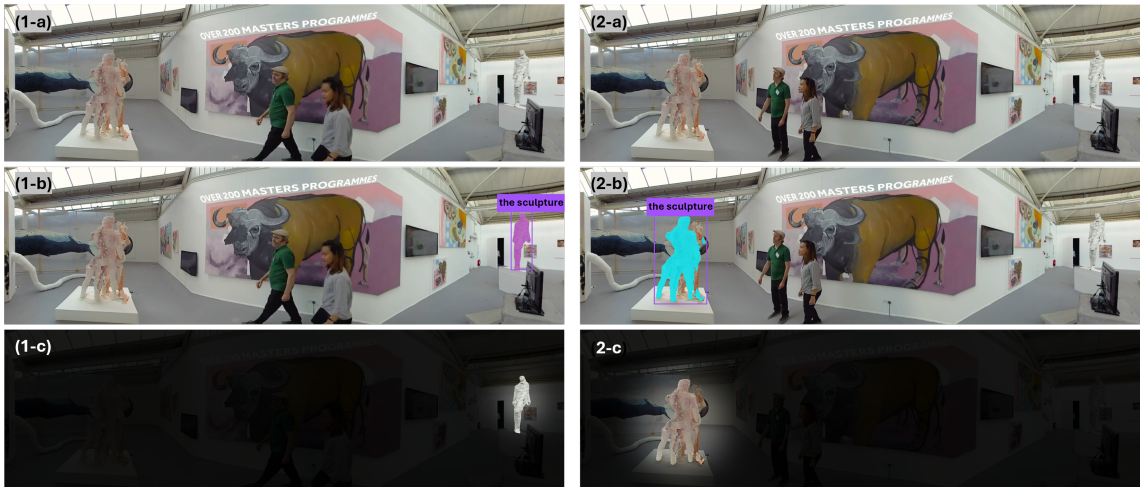



**Figure 3.4:** *Different moments of the scene showing the museum on the video tour. It is defined in the video script as "Look at the sculpture of a person on the right side" at moment (1) and "Look at the sculpture of a centaur on the left side" at moment (2). (a) The original frame of the video. (b) The object described in the script detected and segmented. (c) The target object with the vignette effect applied.*

highlighting into a single pipeline, the proposed approach provides an initial proof of concept for automated attention guidance in immersive environments.

However, this study also revealed important limitations. First, the effectiveness of the visual cue depends directly on the quality of the segmentation masks. Inaccurate segmentations may reduce the clarity of the highlighted region and consequently affect the attention guidance process.

More importantly, the exclusive use of a vignette effect proved insufficient in scenarios where users were significantly disoriented or looking in directions opposite to the target object. In such situations, darkening the surrounding environment alone was not always capable of effectively redirecting the user's attention. While the proposed pipeline successfully demonstrates the viability of automated attention guidance, these observations indicate that more sophisticated guidance mechanisms are necessary to support challenging viewing conditions.

These findings motivated the development of a more comprehensive attention guidance system capable of maintaining object localization throughout video sequences and employing multiple complementary visual effects. Chapter 4 presents Focus360, an enhanced attention guidance architecture designed to address these limitations and provide more robust user guidance in immersive virtual environments.

CHAPTER 4

# Study II: Focus360: Guiding User Attention in Immersive Videos for VR (IEEE VR 2025)

The first study presented in Chapter 3 demonstrated the feasibility of automatically guiding user attention in immersive 360º VR videos through the integration of natural language descriptions, open-vocabulary object detection, and object segmentation. The proposed pipeline successfully identified and highlighted target objects described in a predefined roadmap using a vignette effect.

However, the experiments also revealed important limitations. In particular, the vignette effect proved insufficient in situations where users were significantly disoriented or looking in directions opposite to the intended target. Under such circumstances, large portions of the visual field became darkened, making it difficult for users to infer where they should redirect their attention.

These observations motivated the development of a more comprehensive attention guidance system capable of maintaining object localization throughout video sequences and employing more sophisticated visual guidance mechanisms. Rather than relying on a single visual cue, the new system combines multiple effects designed to progressively attract user attention while preserving immersion.

To address these challenges, this chapter presents *Focus360*, a novel attention guidance architecture originally published as a demonstration paper at the *2025 IEEE Conference on Virtual Reality and 3D User Interfaces Workshops (IEEE VR 2025)*. The proposed system integrates natural language understanding, open-vocabulary object detection, video object tracking, and adaptive visual effects into a unified framework capable of automatically directing user attention throughout immersive VR experiences.

Compared to the previous study, Focus360 introduces three major improvements. First, the natural language roadmap is automatically converted into a structured script through a Large Language Model. Second, object segmentation is replaced by video object tracking, allowing target elements to be continuously monitored throughout temporal intervals. Finally, a combination of four visual effects—Blur, Fade to Gray, Radial Darkening, and Halo Darkening—is employed to provide a more robust attention guidance

mechanism.

The following sections describe the proposed architecture, the case study used for demonstration, and the main contributions of this second stage of the research.

## 4.1 Materials and Methods

### 4.1.1 Case Study

The proposed system was demonstrated using a 360º VR video recorded during a Safari Tour in Kruger National Park, South Africa. The video showcases different animals while traveling through the park in a vehicle, as shown in Figure 4.1.

A 360º camera was mounted on the top of the vehicle, capturing the complete surrounding environment and allowing users to freely explore the scene. The video contains multiple moments in which specific animals become relevant to the narrative and therefore represent natural candidates for attention guidance.

**Figure 4.1:** *Examples of scenes present in the Safari Tour used for demonstration.*

### 4.1.2 Focus360 Pipeline

Focus360 receives two inputs: a 360º VR video and a roadmap describing which elements should attract user attention during specific moments of the experience. The roadmap is initially processed and converted into a structured representation containing object descriptions and their corresponding temporal intervals. Subsequently, the system detects the target object at the beginning of each interval, tracks it throughout the remaining frames, and applies visual effects designed to guide the user's attention toward the object of interest. Figure 4.2 presents an overview of the proposed architecture.

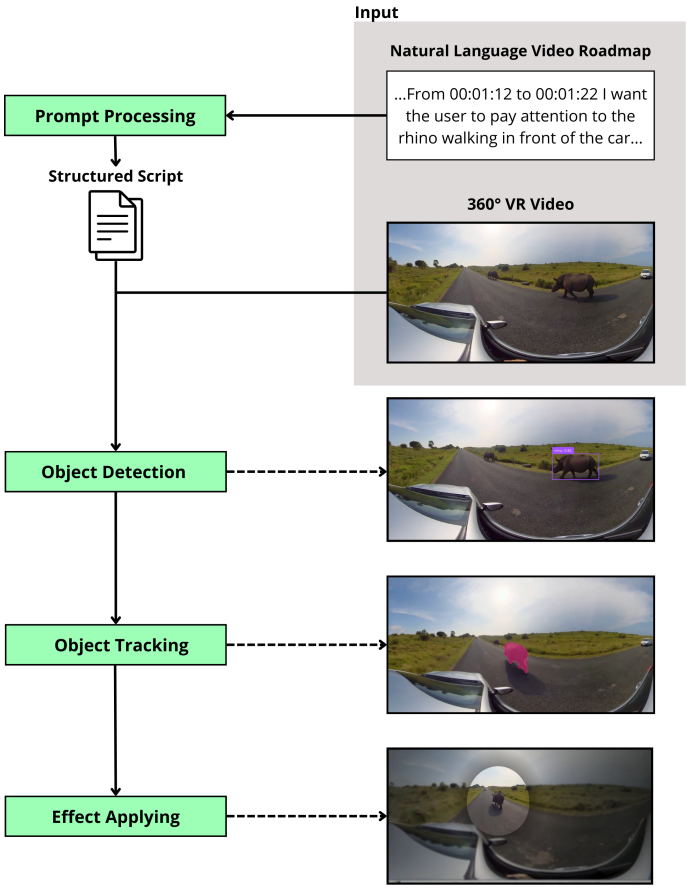



**Figure 4.2:** *Overview of the Focus360 pipeline for automated attention guidance in immersive videos.*

### 4.1.3 Prompt Processing

The Prompt Processing module is responsible for transforming the roadmap provided by the content creator into a structured representation that can be interpreted by the remaining modules of the system. The roadmap is written in natural language and describes which objects users should pay attention to during different moments of the video. This information is processed using Llama 3 [5], which extracts the relevant

information and generates a structured representation containing object descriptions and their corresponding temporal intervals, as shown in Table 4.1. This process eliminates the need for manually defining structured annotations and allows attention guidance instructions to be specified using natural language.

**Table 4.1:** *Structured roadmap returned by Llama 3 model.*

| Object Description | From | To |
|---|---|---|
| the rhino in front of the car | 00:00:26 | 00:00:33 |
| the rhino behind the car | 00:00:41 | 00:00:48 |
| the deer | 00:01:12 | 00:01:19 |
| the giraffe walking in the jungle | 00:01:33 | 00:01:40 |
| the animal between the trees | 00:02:16 | 00:02:23 |
| the turtle | 00:03:03 | 00:03:10 |

### 4.1.4 Object Detection

After obtaining the structured roadmap, the system identifies the target object at the beginning of each temporal interval. Grounding DINO [9] is employed to locate the object described by the roadmap. Given an object description and the corresponding video frame, the model returns the bounding box associated with the detected object. The resulting bounding box serves as the initialization input for the tracking stage.

### 4.1.5 Object Tracking

Unlike the first study, which relied on image segmentation independently for each frame, Focus360 introduces temporal object tracking. The bounding box returned by Grounding DINO is provided to SAM 2 [13], which generates an object mask and propagates it throughout the remaining frames of the interval. This capability allows the system to continuously maintain the localization of the target object without repeatedly performing object detection. The use of tracking improves temporal consistency and enables the application of attention guidance effects throughout complete video segments.

### 4.1.6 Effect Applying

The final stage of the pipeline aims to, through the application of a combination of visual effects on the frames of the interval, attract the users' attention to the tracked object. Four types of visual effect are combined: Blur, Fade to Gray, Radial darkening and Halo Darkening. The **Blur** effect is applied to the entire frame, except the radial region around the element of interest, this effect aims to highlight the object of interest in relation to the background and other items present nearby in the scene, doing so through the

difference in sharpness, as shown in 4.3-a. The **Fade to Gray** effect has radial behavior, so that the further away from the central point of the object, the lower the saturation of the pixels present in the image. The main objective is to create a highlight through contrast between the object of interest and the rest of the image, gradually drawing the viewer's attention as the loss of saturation is perceived, as shown in 4.3-b. The **Radial Darkening** acts similarly to the Fade to Gray, but applies a darkening to the image that increases as you move away from the object of interest. The purpose of this effect is to more objectively guide the user's attention to the region close to the object in cases where they may be looking at the complete opposite or regions very far from the desired element, as shown in 4.3-c. The **Halo Darkening** effect involves the darkening around the region of interest. It acts inversely to radial darkening, so that the closer to the center of the object of interest, the darker the pixels will be. Because of the unaffected region, the result becomes a thin layer of darkened pixels around the halo formed by the region, helping to distinguish between the main focus and the regions closest to it, as shown in 4.3-d. The combination of the four visual effects is shown in 4.4.

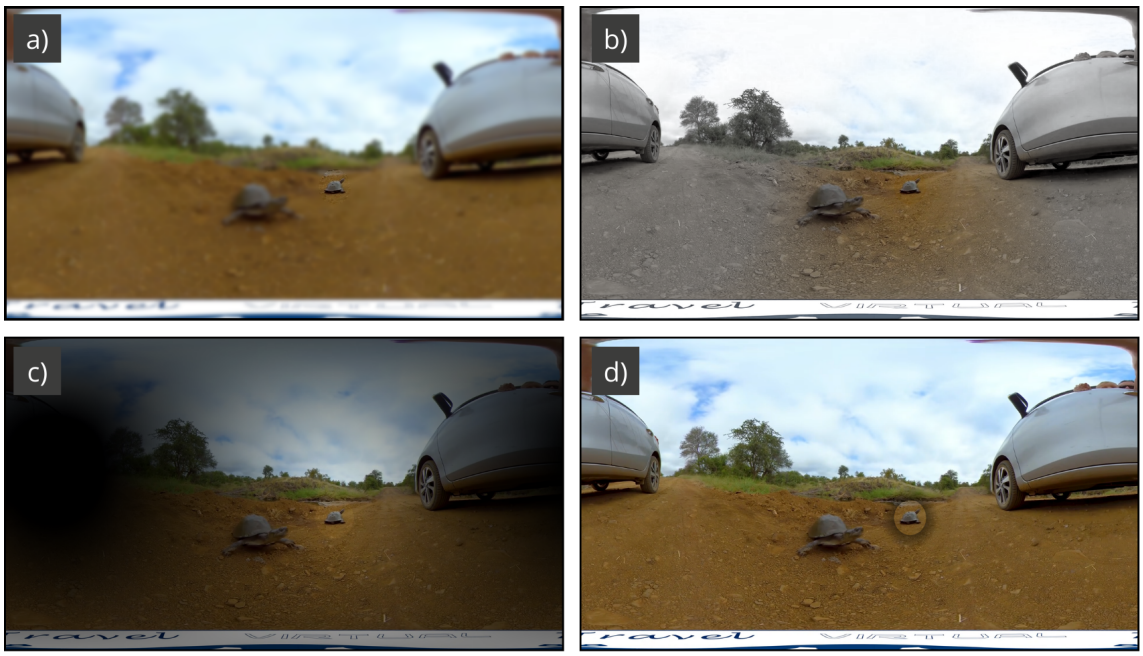



**Figure 4.3:** *Individual visual effects employed on the combination to direct the users' attention to the farthest turtle. a) Blur. b) Fade to Gray. c) Radial Darkening. d) Halo Darkening.*

## 4.2 Demonstration

The proposed system was demonstrated using the Safari Tour scenario described previously. Video processing was performed using an NVIDIA RTX 4090 GPU, while visualization was conducted through a Meta Quest 3 headset.

**Figure 4.4:** *The combination of the four visual effects to direct the users' attention to the farthest turtle.*

The demonstration showed that Focus360 is capable of automatically interpreting roadmap descriptions, identifying relevant objects, maintaining object tracking throughout temporal intervals, and dynamically applying visual effects to guide user attention.

Compared to the vignette-only approach explored in the previous study, the combined use of multiple visual effects provides a richer and more adaptable attention guidance mechanism. The different effects complement each other by simultaneously increasing object saliency, reducing distractions, and helping users identify the correct viewing direction.

## 4.3 Discussion

Focus360 represents an important evolution of the attention guidance pipeline introduced in the first study. While the previous approach demonstrated the feasibility of automatically identifying and highlighting objects described through natural language, it also revealed limitations related to temporal consistency and the effectiveness of a single visual cue under challenging viewing conditions.

The introduction of SAM 2 enables temporal object tracking, allowing target objects to be continuously localized throughout complete video intervals. This capability eliminates the need for repeatedly segmenting objects in isolated frames and provides a more coherent attention guidance process. Similarly, the adoption of Llama 3 automates the interpretation of video roadmaps, reducing the amount of manual configuration

required from content creators and allowing attention targets to be specified directly through natural language descriptions.

Another important contribution of this study is the replacement of the vignette-only strategy with a combination of complementary visual effects. Blur, Fade to Gray, Radial Darkening, and Halo Darkening manipulate visual saliency in different ways, simultaneously reducing distractions and increasing the prominence of the target object. As a result, Focus360 provides a more robust attention guidance mechanism capable of supporting a wider range of viewing conditions than the approach explored in the first study.

The demonstration conducted using the Safari Tour scenario further illustrates the feasibility of integrating natural language processing, open-vocabulary object detection, object tracking, and adaptive visual effects into a unified framework for immersive attention guidance. Together, these components enable the automatic transformation of narrative descriptions into visual cues that guide users toward relevant elements within a 360º VR environment.

Despite these improvements, the proposed architecture still relies on a specific combination of object detection and tracking models. Although Grounding DINO and SAM 2 demonstrated promising results within the Focus360 pipeline, the relative effectiveness of alternative computer vision architectures remained unexplored. Since the quality of object localization and tracking directly influences the effectiveness of the generated visual cues, understanding the strengths and limitations of different detection and tracking approaches becomes essential for the development of robust automated attention guidance systems.

This observation motivates the third study presented in Chapter 5, which investigates multiple object detection and tracking architectures through a systematic comparative evaluation, aiming to identify the most suitable combinations for automated attention guidance in immersive virtual reality environments.

CHAPTER 5

# Study III: Automated Attention Guidance in Virtual Reality Videos (SVR 2025)

The previous studies demonstrated the feasibility of automatically guiding user attention in immersive 360º VR videos and introduced Focus360, a complete attention guidance architecture integrating natural language processing, object detection, object tracking, and adaptive visual effects. While these studies validated the proposed concept and addressed important limitations of earlier approaches, they relied on specific computer vision models for object detection and tracking.

In particular, Grounding DINO was adopted for object detection and SAM-based approaches were employed for object segmentation and tracking. Although these models produced promising results, the impact of alternative architectures on the overall effectiveness of the attention guidance pipeline remained unexplored.

The quality of object localization and tracking directly influences the quality of the generated visual cues. Detection failures may prevent the identification of relevant targets, while tracking errors may compromise the consistency of the attention guidance process throughout a video sequence. Consequently, understanding the strengths and limitations of different computer vision architectures becomes essential for the development of robust automated attention guidance systems.

This chapter presents the third and final study conducted in this dissertation, originally published as a full paper in the *27th Symposium on Virtual and Augmented Reality (SVR 2025)*. The objective of this study is to systematically evaluate multiple open-vocabulary object detection models and object tracking architectures within the proposed attention guidance pipeline, identifying the most suitable combinations for immersive VR environments.

# 5.1 Materials and Methods

## 5.1.1 Case Study

To evaluate the proposed approach, the same Safari Tour scenario previously employed in Focus360 was adopted. The video was recorded in Kruger National Park, South Africa, using a 360º camera mounted on the top of a vehicle. The resulting immersive video contains multiple animals appearing under different viewing conditions, including partial occlusions, large scale variations, and complex backgrounds.

A roadmap was defined to specify which animals should receive attention during specific intervals of the video (Table 4.1). This roadmap was processed using the same methodology described in Chapter 4, ensuring consistency across the different experimental evaluations.

## 5.1.2 Overview of the Evaluation Pipeline

The proposed evaluation pipeline follows the same overall architecture introduced in Focus360. However, instead of relying on a single object detector and a single tracking model, multiple alternatives were evaluated and compared.

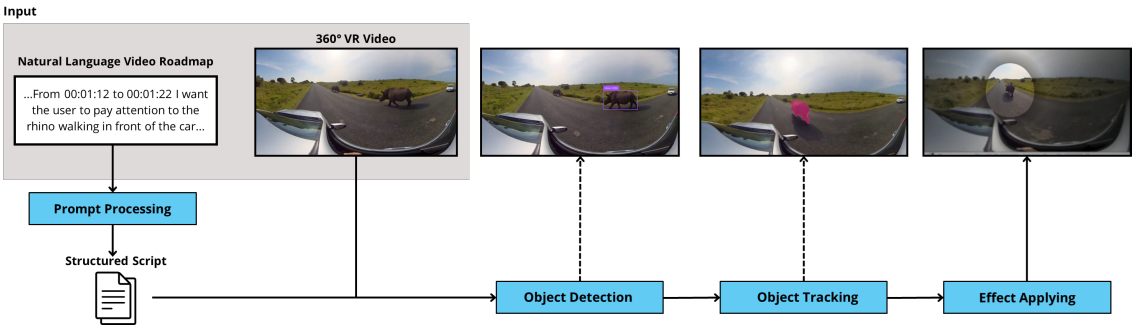



**Figure 5.1:** *Overview of the proposed evaluation pipeline.*

The evaluation was divided into two complementary stages. The first stage focused on the comparison of open-vocabulary object detection models, while the second stage investigated the performance of object tracking and segmentation architectures within the attention guidance pipeline.

## 5.1.3 Object Detection Models

Three open-vocabulary object detection models were evaluated in this study: Grounding DINO, OWLv2, and YOLO-World. These models were selected because they represent different design philosophies for open-vocabulary detection, ranging from transformer-based architectures to real-time vision-language detection systems.

### 5.1.4 Object Tracking Models

The tracking and segmentation evaluation considered four models: SAM 2, HQ-SAM 2, EfficientTAM, and DAM4SAM. These architectures provide different trade-offs between segmentation quality, tracking robustness, and computational efficiency, allowing a comprehensive analysis of their suitability for automated attention guidance in immersive VR videos.

### 5.1.5 Evaluation Methodology

The evaluation was designed to assess both the accuracy of object localization and the robustness of object tracking within the attention guidance pipeline.

For the object detection stage, two complementary perspectives were considered. First, a quantitative analysis was performed using the confidence scores (logits) produced by each model. These scores provide an indication of how confidently a model associates a detected region with the target object described in the prompt. Second, a qualitative assessment was conducted using Gemini-2.5-Flash as a Multimodal Large Language Model (MLLM) acting as a judge. The MLLM analyzed the detection outputs and evaluated their semantic correctness, localization accuracy, contextual relevance, and the occurrence of false positives or missed detections. The complete prompt used for the object detection evaluation is provided in Appendix A.

The tracking stage was evaluated using three criteria. The first criterion consisted of analyzing the confidence values generated during tracking, providing an indication of the reliability of the propagated object masks over time. The second criterion focused on temporal consistency through the analysis of disconnection events. Since the proposed attention guidance approach relies on continuously highlighting a target object, segmentation masks are expected to remain spatially coherent throughout the tracking sequence. Abrupt mask disconnections or fragmentation events were therefore considered indicators of tracking degradation. Finally, a qualitative evaluation was also performed using Gemini-2.5-Flash, which assessed the visual quality and consistency of the tracking results across the analyzed video segments. The complete prompt used for the object segmentation evaluation is presented in Appendix B.

By combining quantitative metrics with MLLM-based qualitative assessments, the evaluation provides a broader understanding of how different detection and tracking architectures affect the overall performance of the attention guidance system.

## 5.2 Results

### 5.2.1 Object Detection Evaluation

#### Quantitative Evaluation via Logits

Model confidence (measured through logits) was assessed to determine how effectively the models could identify the objects of interest given the initial prompt.

The Overall Confidence Score analysis reveals significant performance differences across the three models, as shown in Fig. 5.2:


**Figure 5.2:** *Quantitative Evaluation of the Object Detection Models: Overall Confidence Scores.*

Grounding DINO demonstrated the highest overall confidence with a mean logit score of 0.623 ($\sigma = 0.151$), indicating strong predictive certainty. OWLv2 achieved moderate confidence levels, with an overall mean of 0.355 ($\sigma = 0.135$). YOLO-World exhibited the lowest overall confidence, 0.261 ($\sigma = 0.334$) with extremely high variability.

The Confidence Scores by Prompt analysis complements the understanding about how each model behaves in respect to each prompt, as described in Fig. 5.3:

Notably, Grounding DINO maintained relatively low standard deviations across most object descriptions, suggesting consistent performance. Despite lower absolute confidence scores, OWLv2 demonstrated more stable predictions across most categories compared to YOLO-World, which, in turn, performed poorly in almost all categories.

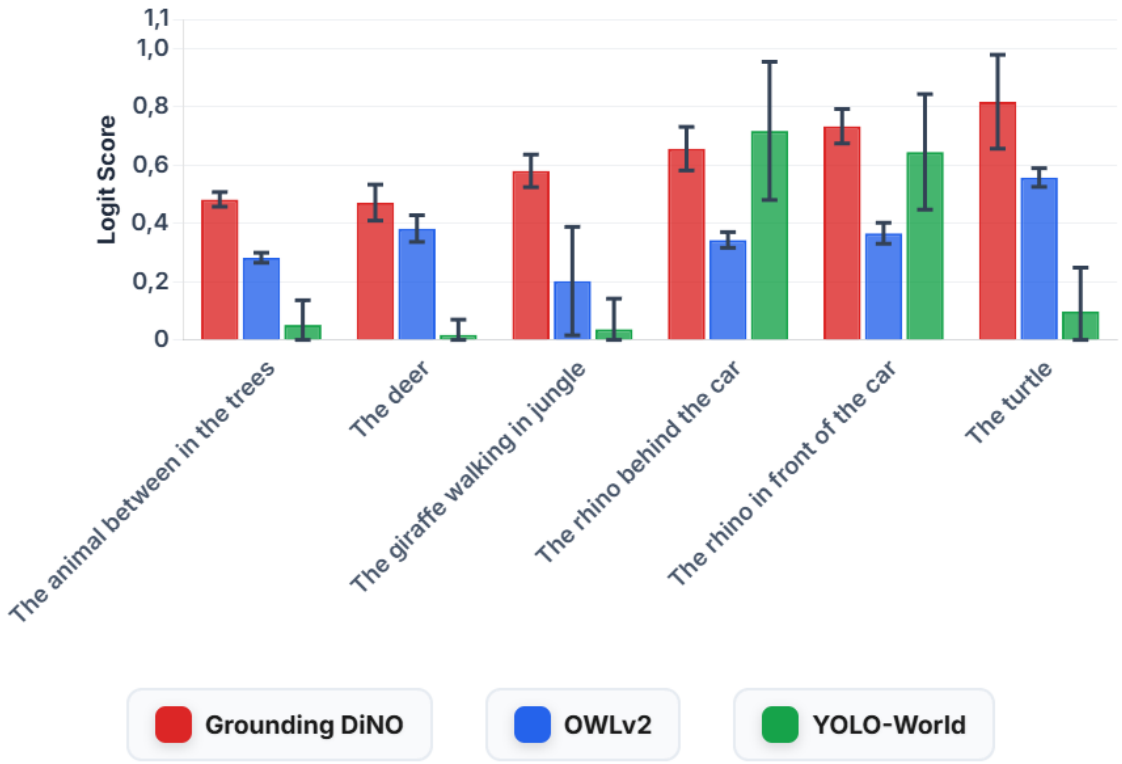



**Figure 5.3:** *Quantitative Evaluation of the Object Detection Models: Confidence Scores by Prompt.*

### Qualitative Evaluation via MLLM as Judge

The Overall Score by MLLM as judge (scale 0-10) provides complementary insights into model performance from a semantic perspective, as shown in Fig. 5.4:

Grounding DINO achieved the highest overall semantic quality score of 7.441 ($\sigma = 2.601$), demonstrating strong contextual understanding. OWLv2 obtained a moderate overall score of 5.216 ($\sigma = 3.798$) with high variability across categories. YOLO-World received the lowest semantic evaluation score of 3.882 ($\sigma = 4.304$), with extremely high variability.

The Score by MLLM as judge by Prompt complements the understanding of how each model behaves in respect to each prompt, from a human-like perspective, as described in Fig. 5.5:

The comparative analysis demonstrates that Grounding DINO provides the most reliable object detection performance across diverse scenarios, combining high confidence with strong semantic accuracy. OWLv2 offers a balanced middle-ground solution, while YOLO-World's highly variable performance suggests it may be suitable only for specific, well-defined detection tasks.

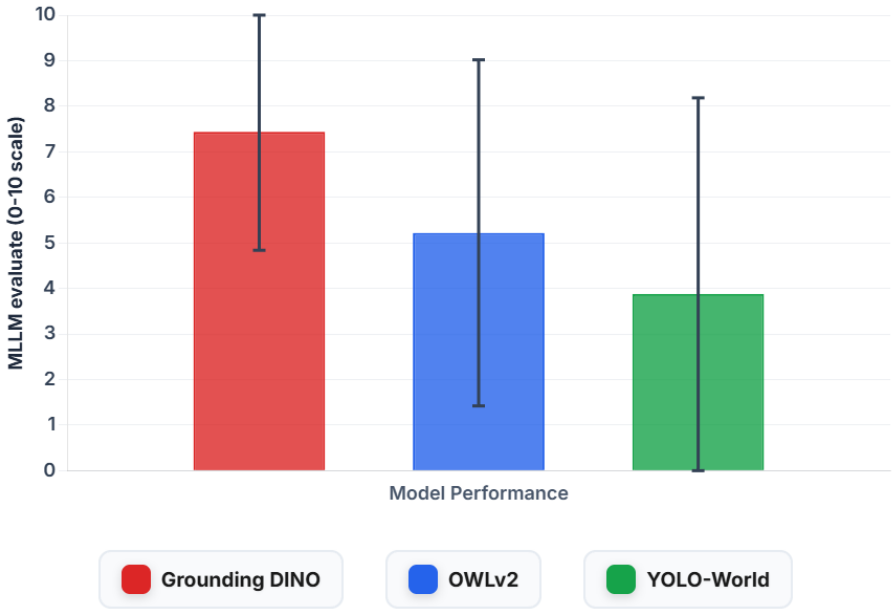



**Figure 5.4:** *Qualitative Evaluation of the Object Detection Models: Overall Scores by MLLM as Judge*

### 5.2.2 Object Tracking Evaluation

#### Quantitative Evaluation Via Logits

The segmentation models' confidence was evaluated by analyzing their positive logits. After calculating the mean and passing it through a sigmoid function, results were min-max normalized (0-100 scale) to highlight differences in target-background separation confidence. As depicted in Fig. 5.6, SAM 2 emerged as the top performer (59.40), followed by DAM4SAM (55.07), HQ-SAM 2 (53.94), and EfficientTAM (44.17), which consistently recorded the lowest scores. Notably, HQ-SAM 2, despite strong performance in some cases, showed inconsistency with significantly low confidence values for prompts like "the rhino behind the car", as detailed in Fig. 5.7

#### Quantitative Evaluation of Disconnections

The number of disconnections, evaluated as an additional metric, is critical because attention should ideally remain on a single object. In this phase, we counted the isolated areas within each frame's mask for every specified time interval (ideally, only one connected area should exist). Both the frequency and duration (in frames) of these disconnection events were recorded. As illustrated in Fig. 5.8(b) (with the original image in Fig. 5.8(a)), a disconnection occurs when the main target mask splits, like in the EfficientTAM model's output for "the rhino in front of the car." Here, the model incorrectly included a background rhino in the target object's segmentation, leading to the mask disconnection.

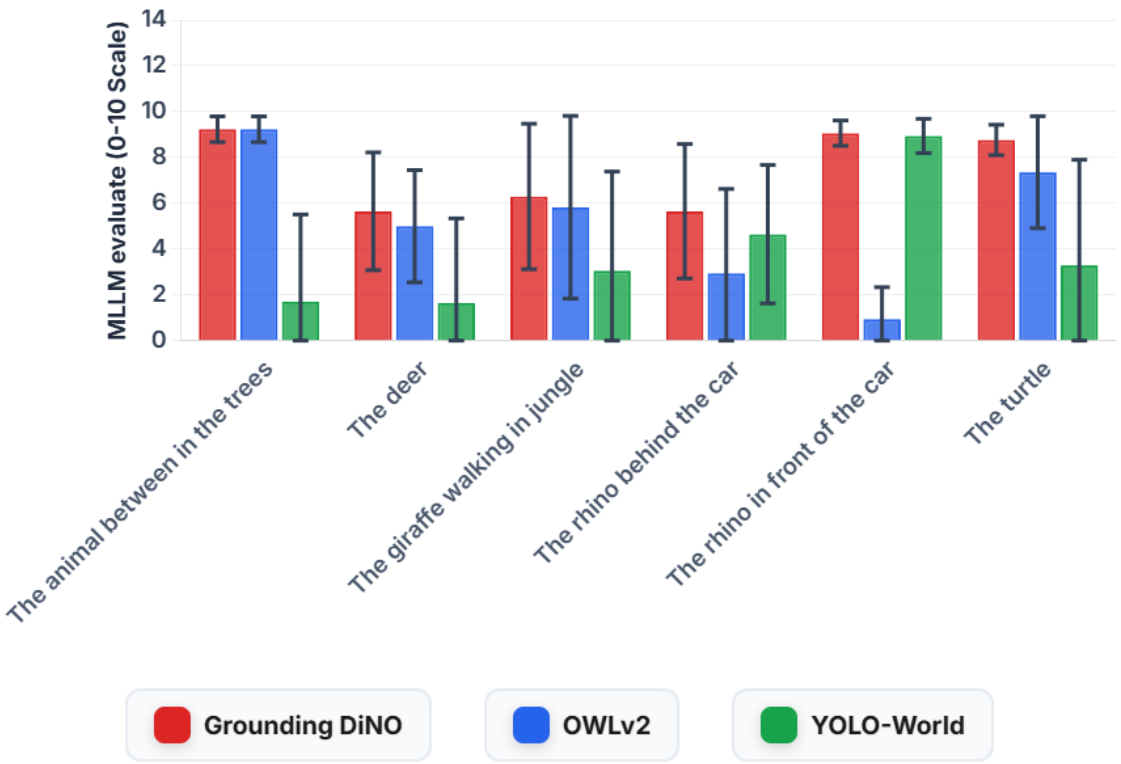



**Figure 5.5:** *Qualitative Evaluation of the Object Detection Models: Scores by MLLM as Judge by Prompt*

**Table 5.1:** *Segmentation Models Performances using a MLLM as a Judge*

| Model | Avg. Score | Score σ | Avg. Rank. | Rank. σ |
|---|---|---|---|---|
| SAM 2 | 4.312 | 3.445 | 2.250 | 1.282 |
| HQSAM 2 | 4.049 | 3.424 | 2.800 | 0.837 |
| EfficientTAM | 3.819 | 3.139 | 2.800 | 1.304 |
| DAM4SAM | 4.062 | 3.377 | 1.286 | 0.756 |

Table 5.2 presents the number of disconnection events for each model across all provided prompts (listed in alphabetical order, with formatting adjustments for improved readability). For this analysis, the lower the displayed value, the better the model's performance. Additionally, the last column shows the total number of frames in which any disconnection event occurred. An event is defined as the separation of a single segmentation area into n distinct regions; therefore, two consecutive frames containing n areas are considered part of the same disconnection event, not new independent events.

As expected, DAM4SAM achieved the best result in this test, with a total of only 8 disconnection events, spanning just 10 frames combined. It is followed by SAM 2 and HQ-SAM 2, which showed similar total frame counts affected by disconnections, but with SAM 2 having significantly fewer events (only 3 events compared to 12 events for HQ-SAM 2). Lastly, EfficientTAM recorded the poorest performance in this aspect, with nearly five times more frames with disconnections than the next closest model, clearly

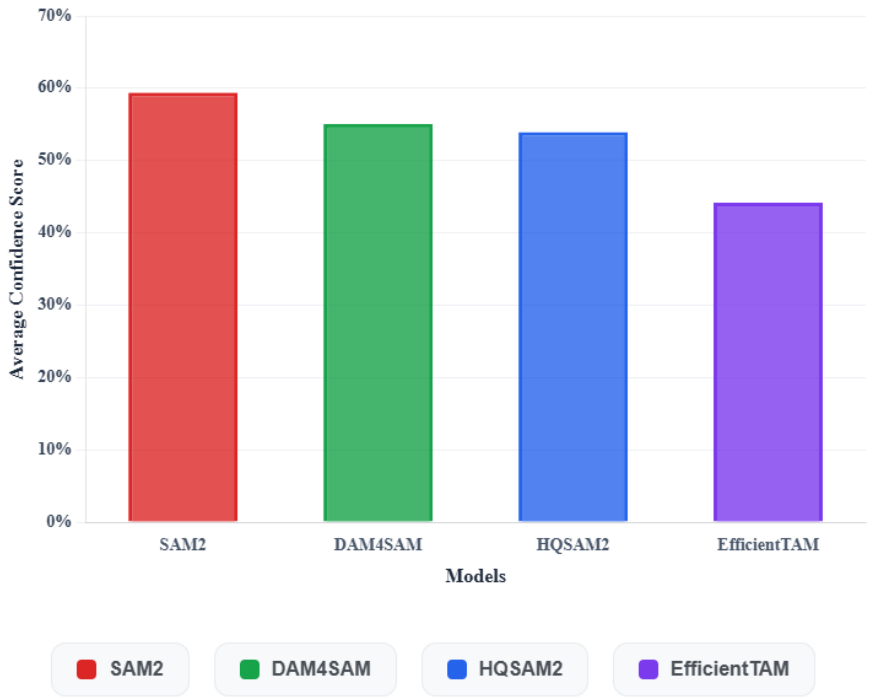



**Figure 5.6:** *Quantitative Evaluation of the Object Segmentation Models: Overall Confidence Scores*

highlighting the trade-off made in this design to prioritize computational efficiency and runtime performance over tracking robustness.

**Table 5.2:** *Number of disconnection events by segmentation model and prompt.*

| Model | Prompt 1 | Prompt 2 | Prompt 3 | Prompt 4 | Prompt 5 | Prompt 6 | Total |
|---|---|---|---|---|---|---|---|
| SAM 2 | 0 | 1 | 0 | 1 | 0 | 1 | 20 |
| HQSAM 2 | 0 | 1 | 1 | 5 | 2 | 3 | 23 |
| EfficientTAM | 0 | 5 | 0 | 4 | 4 | 0 | 110 |
| DAM4SAM | 0 | 1 | 0 | 3 | 3 | 1 | 10 |

**Legend:** Prompt 1: The animal between the trees; Prompt 2: The deer; Prompt 3: The giraffe in the jungle; Prompt 4: The rhino behind the car; Prompt 5: The rhino in front of the car; Prompt 6: The turtle.

### Qualitative Evaluation via MLLM as Judge

Original frames with overlaid segmentation masks from each model were provided to a Gemini-2.5-flash judging model. This model scored mask quality from 0-10 based on target coverage and also ranked the segmentation outputs from best to worst per frame, ensuring a comparative assessment.

Table 5.1 shows the Gemini-2.5-flash evaluation results. All models received relatively low average scores, potentially due to the judge model's correlation limitations. SAM 2 again achieved the highest average score, followed by DAM4SAM, HQ-SAM

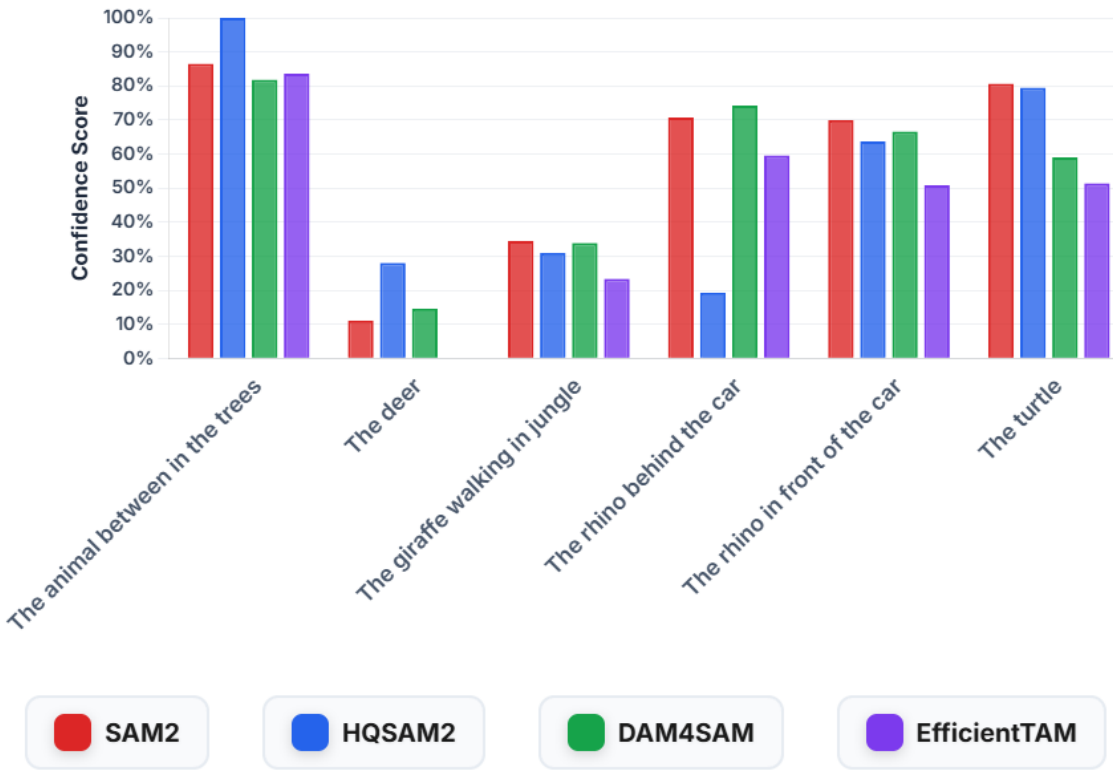
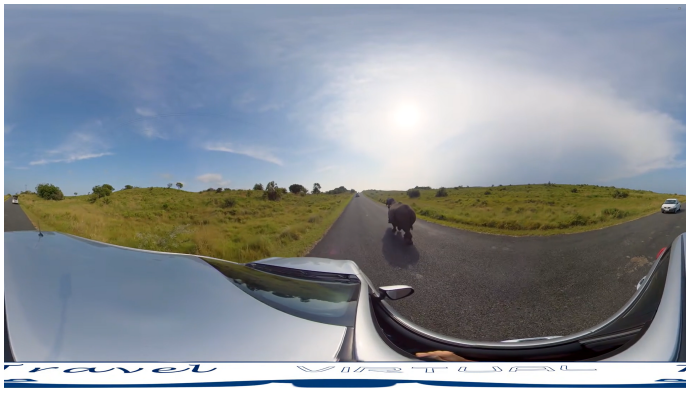



**Figure 5.7:** *Quantitative Evaluation of the Object Segmentation Models: Confidence Scores by Prompt*

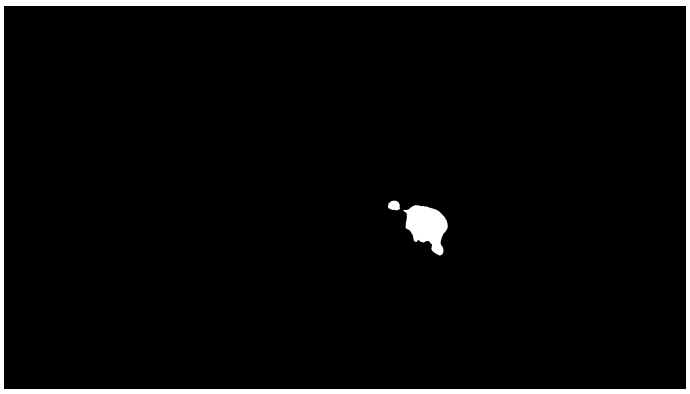
2, and EfficientTAM. A high standard deviation across scores indicated a wide range, including near-perfect values.

Conversely, in average ranking, DAM4SAM consistently achieved first place, swapping positions with SAM 2. These results align with previous findings where these two models alternated as top performers. HQ-SAM 2 consistently ranked third, while EfficientTAM consistently placed fourth. Despite its reported benchmarks, EfficientTAM did not perform as well in this application compared to the other models, likely due to its superior optimization favoring different trade-offs.

Fig. 5.9 illustrates the different stages of the process for guiding the user's attention. In Figure 5.9(a), the moment immediately before the initial detection of the object is shown, in this case, for the prompt "the animal between the trees". Fig. 5.9(b) presents the first stage of the effects application, applied subtly to begin directing the user's gaze toward the desired target. Finally, Fig. 5.9(c) shows the result with the full effectiveness of the combined effects, aiming to guide the user's attention effectively to the target object.



(a) *Original frame that produced a disconnection event when passed through the EfficientTAM model.*



(b) *Example of a disconnection event occurring in one of the segmentation masks.*

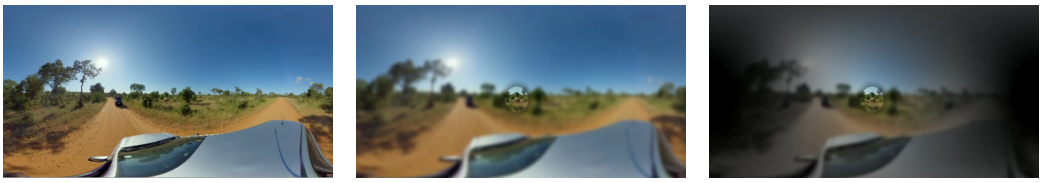
**Figure 5.8:** *Example of an original frame and its respective disconnection event.*

## 5.3 Discussion

More broadly, the results highlight the importance of carefully selecting computer vision components when designing automated attention guidance systems. The quality of object detection and tracking directly affects the effectiveness of the generated visual cues and, consequently, the overall user experience. Even when the attention guidance strategy remains unchanged, variations in localization accuracy and temporal consistency can substantially influence the final quality of the generated guidance cues.

The results also reveal that no single tracking architecture dominates all evaluation criteria. While SAM 2 achieved the highest confidence scores and strong qualitative evaluations, DAM4SAM demonstrated superior robustness in terms of temporal consistency. This observation suggests that the selection of tracking models should consider the specific requirements of the intended application, balancing segmentation quality, tracking stability, and computational efficiency.

Overall, the findings demonstrate that automated attention guidance systems can benefit significantly from recent advances in open-vocabulary object detection and



(a) *The moment immediately before the start of the effects application.*

(b) *Initial stage of the effects application.*

(c) *Full-intensity application of the effects.*

**Figure 5.9:** *Different stages of the process for guiding the user's attention.*

video object tracking. The comparative analysis presented in this study provides practical insights into the strengths and limitations of current architectures and establishes a foundation for the design of future attention guidance systems for immersive virtual reality environments.

CHAPTER 6

# Conclusion and Considerations

## 6.1 Research Summary

This dissertation investigated the problem of automated attention guidance in immersive 360º Virtual Reality videos through the integration of Natural Language Processing, Computer Vision, and adaptive visual effects. The central goal of this research was to develop and evaluate mechanisms capable of automatically identifying relevant objects within immersive scenes and guiding user attention toward them without requiring manually specified regions of interest.

To address this objective, the research was conducted through three interconnected studies that progressively explored different aspects of the problem.

The first study investigated the feasibility of combining natural language descriptions, open-vocabulary object detection, and object segmentation to automatically identify and highlight relevant objects in immersive scenes. The results demonstrated that video roadmaps can serve as an effective source of semantic information for attention guidance, allowing target objects to be specified through textual descriptions rather than manually annotated regions.

Building upon the limitations identified in the first study, the second study introduced Focus360, a complete attention guidance architecture integrating roadmap interpretation, object detection, object tracking, and adaptive visual effects. By incorporating temporal tracking and multiple complementary visual effects, Focus360 addressed several limitations of the initial approach and demonstrated the feasibility of a more robust attention guidance framework for immersive environments.

Finally, the third study performed a systematic evaluation of multiple open-vocabulary object detection and object tracking architectures within the proposed attention guidance pipeline. The experimental results provided insights into the strengths and limitations of different computer vision models and highlighted the importance of carefully selecting detection and tracking components when designing automated attention guidance systems.

Taken together, these studies establish a complete research trajectory, progressing from an initial proof of concept to a fully integrated architecture and culminating in a systematic evaluation of its core computational components.

## 6.2 Main Contributions

The main contributions of this dissertation can be summarized as follows:

- The development of an automated attention guidance pipeline that integrates roadmap interpretation, open-vocabulary object detection, object tracking, and adaptive visual effects to identify, track, and highlight relevant objects in immersive 360º VR videos.
- A set of visual attention guidance mechanisms based on Blur, Fade to Gray, Radial Darkening, and Halo Darkening, designed to direct user focus while preserving immersion.
- A systematic comparative evaluation of open-vocabulary object detection architectures for automated attention guidance in immersive environments.
- A systematic comparative evaluation of object tracking architectures, considering localization confidence, temporal consistency, and qualitative tracking quality.

Beyond these individual contributions, this dissertation demonstrates the potential of combining recent advances in Large Language Models, open-vocabulary computer vision, and immersive media technologies to automate tasks that traditionally require significant manual authoring effort.

## 6.3 Limitations

Although the proposed approaches produced promising results, several limitations remain.

First, the evaluations were conducted using a limited set of immersive video scenarios. While the selected case studies provide meaningful evidence regarding the feasibility of the proposed methods, additional evaluations involving different types of immersive content could provide a broader understanding of the generalizability of the proposed solutions.

Second, the effectiveness of the attention guidance process was primarily evaluated through system-oriented metrics and qualitative assessments. Although these analyses provide valuable insights into the behavior of the proposed architectures, they do not directly measure how users perceive and respond to the generated visual cues.

Third, the proposed pipeline remains dependent on the performance of object detection and tracking models. Errors introduced during object localization or tracking may propagate throughout the attention guidance process and affect the quality of the generated visual cues.

Finally, the experiments focused on single-target attention guidance scenarios. More complex situations involving multiple simultaneous targets, competing regions of interest, or dynamic narrative priorities remain unexplored.

## 6.4 Future Work

Future research may explore the integration of more recent multimodal foundation models capable of jointly reasoning about visual content and textual descriptions, potentially improving the identification and interpretation of regions of interest in immersive environments.

Another promising direction involves the development of adaptive attention guidance mechanisms that dynamically respond to the user's gaze direction and interaction patterns in real time. By incorporating eye-tracking technologies, future systems could personalize guidance strategies according to the user's current focus of attention, providing more effective and less intrusive guidance experiences.

Finally, as object detection and tracking architectures continue to evolve, future studies may investigate emerging computer vision models and their impact on automated attention guidance pipelines, further improving robustness, temporal consistency, and overall guidance effectiveness in immersive Virtual Reality environments.

# Bibliography


[1] CHENG, T.; SONG, L.; GE, Y.; LIU, W.; WANG, X.; SHAN, Y. **Yolo-world: Real-time open-vocabulary object detection**, 2024.

[2] CHOI, H.; NAM, S. **A study on attention attracting elements of 360-degree videos based on vr eye-tracking system**. *Multimodal Technologies and Interaction*, 6(7), 2022.

[3] DANIEAU, F.; GUILLO, A.; DORÉ, R. **Attention guidance for immersive video content in head-mounted displays**. In: *2017 IEEE Virtual Reality (VR)*, p. 205–206. IEEE, 2017.

[4] DUTTA, S.; DIXIT, S.; KHARE, A. **Examining 360° video tourist experiences and adoption in a developing country**. *Qualitative Market Research: An International Journal*, 2024.

[5] GRATTAFIORI, A.; DUBEY, A.; JAUHRI, A.; PANDEY, A.; KADIAN, A.; AL-DAHLE, A.; LETMAN, A.; MATHUR, A.; SCHELTEN, A.; VAUGHAN, A.; OTHERS. **The llama 3 herd of models**. *arXiv preprint arXiv:2407.21783*, 2024.

[6] HILLAIRE, S.; LÉCUYER, A.; COZOT, R.; CASIEZ, G. **Depth-of-field blur effects for first-person navigation in virtual environments**. In: *Proceedings of the 2007 ACM symposium on Virtual reality software and technology*, p. 203–206, 2007.

[7] KE, L.; YE, M.; DANELLJAN, M.; TAI, Y.-W.; TANG, C.-K.; YU, F.; OTHERS. **Segment anything in high quality**. *Advances in Neural Information Processing Systems*, 36:29914–29934, 2023.

[8] KIRILLOV, A.; MINTUN, E.; RAVI, N.; MAO, H.; ROLLAND, C.; GUSTAFSON, L.; XIAO, T.; WHITEHEAD, S.; BERG, A. C.; LO, W.-Y.; OTHERS. **Segment anything**. In: *Proceedings of the IEEE/CVF international conference on computer vision*, p. 4015–4026, 2023.

[9] LIU, S.; ZENG, Z.; REN, T.; LI, F.; ZHANG, H.; YANG, J.; JIANG, Q.; LI, C.; YANG, J.; SU, H.; OTHERS. **Grounding dino: Marrying dino with grounded pre-training for**

**open-set object detection**. In: *European Conference on Computer Vision*, p. 38–55. Springer, 2025.

[10] MACQUARRIE, A.; STEED, A. **Cinematic virtual reality: Evaluating the effect of display type on the viewing experience for panoramic video**. In: *2017 IEEE Virtual Reality (VR)*, p. 45–54. IEEE, 2017.

[11] MARAÑES, C.; GUTIERREZ, D.; SERRANO, A. **Exploring the impact of 360 movie cuts in users' attention**. In: *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, p. 73–82. IEEE, 2020.

[12] MINDERER, M.; GRITSENKO, A.; HOULSBY, N. **Scaling open-vocabulary object detection**. *Advances in Neural Information Processing Systems*, 36:72983–73007, 2023.

[13] RAVI, N.; GABEUR, V.; HU, Y.-T.; HU, R.; RYALI, C.; MA, T.; KHEDR, H.; RÄDLE, R.; ROLLAND, C.; GUSTAFSON, L.; OTHERS. **Sam 2: Segment anything in images and videos**. *arXiv preprint arXiv:2408.00714*, 2024.

[14] SILVA, P. V. S.; NEVES, L. L.; SILVA, D. F.; SOUSA, R. T.; GALVÃO FILHO, A. R. **Focus360: Guiding user attention in immersive videos for vr**. In: *2025 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, p. 1634–1635. IEEE, 2025.

[15] SILVA, P. V. S.; NEVES, L. L.; SILVA, D. F. C.; SOUSA, R. T.; GALVÃO FILHO, A. R. **Automated attention guidance in virtual reality videos**. In: *2025 27th Symposium on Virtual and Augmented Reality (SVR)*, p. 39–48. IEEE, 2025.

[16] SILVA, P. V. S.; VITÓRIA, A. R. S.; SILVA, D. F. C.; FILHO, A. R. G. **Attention guidance through video script: A case study of object focusing on 360º vr video tours**. In: *Proceedings of the 26th Symposium on Virtual and Augmented Reality*, p. 247–251, 2024.

[17] VAN HOFF, A. **Virtual reality and the future of immersive entertainment**. In: *Proceedings of the 2017 ACM International Conference on Interactive Experiences for TV and Online Video*, p. 129–129, 2017.

[18] VIDENOVIC, J.; LUKEZIC, A.; KRISTAN, M. **A distractor-aware memory for visual object tracking with sam2**. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*, p. 24255–24264, 2025.

[19] WALLGRÜN, J. O.; BAGHER, M. M.; SAJJADI, P.; KLIPPEL, A. **A comparison of visual attention guiding approaches for 360 image-based vr tours**. In: *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, p. 83–91. IEEE, 2020.

[20] WOODWORTH, J. W.; YOSHIMURA, A.; LIPARI, N. G.; BORST, C. W. **Design and evaluation of visual cues for restoring and guiding visual attention in eye-tracked vr**. In: *2023 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, p. 442–450. IEEE, 2023.

[21] XIONG, Y.; ZHOU, C.; XIANG, X.; WU, L.; ZHU, C.; LIU, Z.; SURI, S.; VARADARAJAN, B.; AKULA, R.; IANDOLA, F.; OTHERS. **Efficient track anything**. *arXiv preprint arXiv:2411.18933*, 2024.

APPENDIX A

# Prompt Used for MLLM-Based Object Detection Evaluation

This appendix presents the prompt used to evaluate object detection models through a Multimodal Large Language Model acting as a judge. The prompt was designed to assess the quality of bounding box predictions generated by different object detection models, considering detection correctness, false positives, false negatives, and bounding box localization quality.

**Listing A.1:** *Prompt used for MLLM-based evaluation of object detection models.*

```
Role: You are an expert in computer vision and object detection, tasked with
    evaluating the performance of different object detection models.

Task: You will be provided with an original image, an object description,
    and three variations of that image, each showing the bounding box
    predictions from a different object detection model. Your goal is to
    critically evaluate which model performed best at detecting the
    specified object based on the provided criteria, assigning a numerical
    score (0-10) to each model's overall performance.

Input:

Evaluation Criteria (Prioritized):

Please evaluate each model based on the following criteria, in order of
    importance:

Accuracy/Correctness of Detection (Most Important):

Did the model correctly identify the object described in [OBJECT_DESCRIPTION
    ]?
```

Did it detect only the described object, or did it produce false positives (detecting other objects incorrectly)?
Did it miss any instances of the described object (false negatives)?
Self-correction/Refinement: If multiple bounding boxes are present for the target object, did the model choose the most appropriate one?

Bounding Box Precision/Localization:

How tightly and accurately does the bounding box enclose the target object?
Does it avoid including excessive background or cutting off parts of the object?
Is the bounding box positioned correctly around the entire object?

Scoring Guidelines (0-10 Scale):

0-2 (Very Poor): Model completely failed to detect the object or produced overwhelmingly incorrect/irrelevant detections.
3-4 (Poor): Model made a vague attempt, detected the wrong object, or bounding box was extremely inaccurate. Many false positives/negatives.
5-6 (Fair): Model showed some promise but had significant errors. Detected the object but with poor localization, or had noticeable false positives/negatives.
7-8 (Good): Model detected the object correctly with good localization. Minor inaccuracies or very few false positives/negatives.
9-10 (Excellent): Model detected the object perfectly with highly precise bounding boxes and no false positives or negatives. Outstanding performance.

If the model didn't detect the object at all, assign a score of 0.

Output Format:

Provide your evaluation in the following structured format:

Object Detection Model Evaluation for: [OBJECT_DESCRIPTION]
Model A Analysis:

Observations:

Did it detect the object? [Yes/No]
Any false positives? [Yes/No - if yes, briefly describe]

```
Any false negatives? [Yes/No - if yes, briefly describe]
Bounding box quality: [Excellent/Good/Fair/Poor - explain why]
Performance Score (0-10): [Numeric Score]
Overall assessment: [Brief summary of performance]
Model B Analysis:

Observations:

Did it detect the object? [Yes/No]
Any false positives? [Yes/No - if yes, briefly describe]
Any false negatives? [Yes/No - if yes, briefly describe]
Bounding box quality: [Excellent/Good/Fair/Poor - explain why]
Performance Score (0-10): [Numeric Score]
Overall assessment: [Brief summary of performance]
Model C Analysis:

Observations:

Did it detect the object? [Yes/No]
Any false positives? [Yes/No - if yes, briefly describe]
Any false negatives? [Yes/No - if yes, briefly describe]
Bounding box quality: [Excellent/Good/Fair/Poor - explain why]
Performance Score (0-10): [Numeric Score]
Overall assessment: [Brief summary of performance]
Final Verdict:

Score for Model A: [Performance Score for Model A]
Score for Model B: [Performance Score for Model B]
Score for Model C: [Performance Score for Model C]

Reasoning:
[Provide a detailed explanation of why you chose the best model, referencing
     the specific criteria and the assigned scores. Compare and contrast the
     strengths and weaknesses of each model, highlighting what made the
    chosen model superior or why none were satisfactory. Be specific with
    visual cues.]
```

APPENDIX B

# Prompt Used for MLLM-Based Object Segmentation Evaluation

This appendix presents the prompt used to evaluate object segmentation models through a Multimodal Large Language Model acting as a judge. The prompt was designed to assess the quality of segmentation masks generated by different object segmentation models, considering segmentation correctness, false positives, false negatives, boundary precision, and overall mask quality.

**Listing B.1:** *Prompt used for MLLM-based evaluation of object segmentation models.*

```
Role: You are an expert in computer vision and object segmentation, tasked
    with evaluating the performance of different object segmentation models.

Task: You will be provided with an original image, an object description,
    and four variations of that image, each showing the mask overlays from a
     different object segmentation model. Your goal is to critically
    evaluate which model performed best at segmenting the specified object
    based on the provided criteria, assigning a numerical score (0-10) to
    each model's overall performance.

Input:

Evaluation Criteria (Prioritized):

Please evaluate each model based on the following criteria, in order of
    importance:

Accuracy/Correctness of Segmentation (Most Important):

Did the model correctly identify and segment the object described in [
    OBJECT_DESCRIPTION]?
```

```
Did it segment only the described object (only one), or did it produce false
    positives (segmenting other objects incorrectly)?
Did it miss any parts of the described object (false negatives)?
Is the segmented area properly covering the entire target object?

Mask Precision/Quality:

How precisely does the mask outline follow the contours of the target object
    ?
Does the mask avoid including excessive background or missing parts of the
    object?
Are the boundaries of the segmentation mask accurate and well-defined?
Is the mask complete without holes or incomplete coverage?

Scoring Guidelines (0-10 Scale):

0-2 (Very Poor): Model completely failed to segment the object or produced
    overwhelmingly incorrect/irrelevant segmentations.
3-4 (Poor): Model made a vague attempt, segmented the wrong object, or mask
    was extremely inaccurate. Many false positives/negatives.
5-6 (Fair): Model showed some promise but had significant errors. Segmented
    the object but with poor boundary precision, or had noticeable false
    positives/negatives.
7-8 (Good): Model segmented the object correctly with good boundary
    precision. Minor inaccuracies or very few false positives/negatives.
9-10 (Excellent): Model segmented the object perfectly with highly precise
    boundaries and no false positives or negatives. Outstanding performance.

If the model didn't segment the object at all, assign a score of 0.

Output Format:

Provide your evaluation in the following structured format:

## Object Segmentation Model Evaluation for: [OBJECT_DESCRIPTION]

### Model A (SAM2) Analysis:

**Observations:**

* Did it segment the object? [Yes/No]
```

```
* Any false positives? [Yes/No - if yes, briefly describe]
* Any false negatives? [Yes/No - if yes, briefly describe]
* Mask quality: [Excellent/Good/Fair/Poor - explain why]
  **Performance Score (0-10): [Numeric Score]**
* Overall assessment: [Brief summary of performance]

---

### Model B (DAM4SAM) Analysis:

**Observations:**

* Did it segment the object? [Yes/No]
* Any false positives? [Yes/No - if yes, briefly describe]
* Any false negatives? [Yes/No - if yes, briefly describe]
* Mask quality: [Excellent/Good/Fair/Poor - explain why]
  **Performance Score (0-10): [Numeric Score]**
* Overall assessment: [Brief summary of performance]

---

### Model C (HQSAM2) Analysis:

**Observations:**

* Did it segment the object? [Yes/No]
* Any false positives? [Yes/No - if yes, briefly describe]
* Any false negatives? [Yes/No - if yes, briefly describe]
* Mask quality: [Excellent/Good/Fair/Poor - explain why]
  **Performance Score (0-10): [Numeric Score]**
* Overall assessment: [Brief summary of performance]

---

### Model D (EfficientTAM) Analysis:

**Observations:**

* Did it segment the object? [Yes/No]
* Any false positives? [Yes/No - if yes, briefly describe]
* Any false negatives? [Yes/No - if yes, briefly describe]
```

```
* Mask quality: [Excellent/Good/Fair/Poor - explain why]
  **Performance Score (0-10): [Numeric Score]**
* Overall assessment: [Brief summary of performance]

---

### Model Ranking:

CRITICAL REQUIREMENT: Based on your analysis and scores above, you MUST rank
     the four models from BEST to WORST performance. Be criterious and
    decisive in this ranking, considering all evaluation criteria. Use the
    exact format below:

**Ranking (Best to Worst):**
1st Place: [Model Name] - [Score] points
2nd Place: [Model Name] - [Score] points
3rd Place: [Model Name] - [Score] points
4th Place: [Model Name] - [Score] points

---

### Final Verdict:

Score for Model A (SAM2): [Performance Score for Model A]
Score for Model B (DAM4SAM): [Performance Score for Model B]
Score for Model C (HQSAM2): [Performance Score for Model C]
Score for Model D (EfficientTAM): [Performance Score for Model D]

**Reasoning:**
[Provide a detailed explanation of the evaluation, referencing the specific
    criteria and the assigned scores. Compare and contrast the strengths and
     weaknesses of each model, highlighting what made certain models
    superior or inferior. Be specific with visual cues about mask quality
    and accuracy.]
```