Voila-A: Aligning Vision-Language Models with User's Gaze Attention

Kun Yan[1*], **Zeyu Wang**[2*], **Lei Ji**[3], **Yuntao Wang**[2], **Nan Duan**[3], **Shuai Ma**[1†]
[1] SKLSDE Lab, Beihang University
[2] Key Laboratory of Pervasive Computing, Tsinghua University
[3] Microsoft Research
`kunyan@buaa.edu.cn`, `wang-zy23@mails.tsinghua.edu.cn`
`mashuai@buaa.edu.cn`

Abstract

In recent years, the integration of vision and language understanding has led to significant advancements in artificial intelligence, particularly through Vision-Language Models (VLMs). However, existing VLMs face challenges in handling real-world applications with complex scenes and multiple objects, as well as aligning their focus with the diverse attention patterns of human users. In this paper, we introduce gaze information, feasibly collected by AR or VR devices, as a proxy for human attention to guide VLMs and propose a novel approach, Voila-A, for gaze alignment to enhance the interpretability and effectiveness of these models in real-world applications. First, we collect hundreds of minutes of gaze data to demonstrate that we can mimic human gaze modalities using localized narratives. We then design an automatic data annotation pipeline utilizing GPT-4 to generate the VOILA-COCO dataset. Additionally, we innovate the Voila Perceiver modules to integrate gaze information into VLMs while preserving their pre-trained knowledge. We evaluate Voila-A using a hold-out validation set and a newly collected VOILA-GAZE test set, which features real-life scenarios captured with a gaze-tracking device. Our experimental results demonstrate that Voila-A significantly outperforms several baseline models. By aligning model attention with human gaze patterns, Voila-A paves the way for more intuitive, user-centric VLMs and fosters engaging human-AI interaction across a wide range of applications. Our code is available at https://github.com/naykun/Voila-A

1 Introduction

The integration of vision and language understanding has witnessed significant advancements in recent years, particularly through the development of Vision-Language Models (VLMs). These models have demonstrated remarkable performance in various tasks, such as visual question answering, image captioning, and visual storytelling, among others. Although VLMs exhibit strong performance in various tasks, their applicability in everyday scenarios is hindered by their limited alignment with human users' focus. This misalignment leads to suboptimal performance and decreased user satisfaction. Current VLMs' inability to process these intentional modalities results in imprecise and unhelpful responses. As demonstrated in Figure 1, a user's intent can be communicated through

---

[*] equal contribution

[†] corresponding author

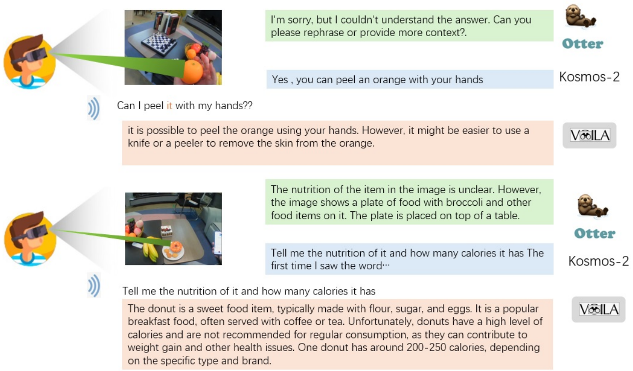
38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-0060

spoken language, multimodal expressions, or even be concealed. The users' attention can clarify vague expressions, meanwhile, uncovering hidden intentions is more challenging.

Most recent VLMs [31, 28, 2] primarily focus on learning alignment between vision input and text tokens for LLMs or designing learnable interaction layers to attend the vision input to the frozen LLM layers. The importance of aligning AI with human attention has been highlighted in previous research, which demonstrates that incorporating visual attention can lead to improved user experience [26, 45, 39]. Additionally, there has been growing interest in grounded VLMs, which investigate the fine-grain grounding capability between region-text pairs instead of image-text pairs and further conduct dense regional prediction tasks [60, 18, 61].

The representation of visual regions within computational models can be achieved through various methodologies, such as bounding boxes [63, 32], discrete points [34], or continuous traces [40, 53]. For the integration of such regional data into models, researchers have developed multiple approaches. These include the concatenation of cropped image patches with the original textual or visual inputs [59, 4], the application of masks or Gaussian maps to highlight user-specified areas [29, 30], and the incorporation of positional encodings to represent points, boxes, or traces [23, 51]. Despite the extensive exploration of bounding boxes and segmentation techniques within VLMs, their application remains suboptimal for end-users to generate input signals.



I'm sorry, but I couldn't understand the answer. Can you please rephrase or provide more context?.

Yes , you can peel an orange with your hands

Can I peel it with my hands??

it is possible to peel the orange using your hands. However, it might be easier to use a knife or a peeler to remove the skin from the orange.

The nutrition of the item in the image is unclear. However, the image shows a plate of food with broccoli and other food items on it. The plate is placed on top of a table.

Tell me the nutrition of it and how many calories it has The first time I saw the word···

Tell me the nutrition of it and how many calories it has

The donut is a sweet food item, typically made with flour, sugar, and eggs. It is a popular breakfast food, often served with coffee or tea. Unfortunately, donuts have a high level of calories and are not recommended for regular consumption, as they can contribute to weight gain and other health issues. One donut has around 200-250 calories, depending on the specific type and brand.

Figure 1: AR and VR scenarios usually involve complex scenes with multiple objects. Users may interested in only one specific object and gaze is the most natural way to interact with the device.

We propose to introduce a novel approach that leverages gaze tracking as a more intuitive and interactive method for defining visual regions, particularly within augmented reality (AR) and virtual reality (VR) environments. Zhang [58] provides an overview of gaze-related research, outlining a process that begins with collecting human gaze data (further discussed in B.3), followed by building models to predict human attention distribution (i.e., saliency models, as discussed in B.4), and culminating in human-gaze-assisted AI. They conclude that ***AI agents capable of perceiving and understanding human gaze behaviors can better infer user needs and assist in daily tasks***. However, they also note that ***research in this final direction is still limited***. Our work aims to advance this area further. Prior research in gaze-based visual representation includes gaze-directed visual grounding [41] and the implementation of eye-gaze within vision transformers [33]. Nonetheless, these methodologies have encountered obstacles in terms of scalability and adaptability. Despite significant progress, the seamless integration of gaze information into large-scale VLMs is still a formidable challenge.

To tackle this issue, we demonstrate that mouse trace data can be a proxy for gaze behavior modeling and leverage trace data from Localized Narratives [40] to annotate instructional data using GPT-4 [36]. Another critical aspect of the challenge is maintaining the integrity of the pre-trained knowledge within VLMs while effectively assimilating gaze data. We further design Voila-A's attention mechanism to incorporate gaze information while not forgetting pre-trained knowledge. We evaluate Voila-A through a hold-out validation set and a newly collected test set VOILA-GAZE, featuring real-life scenarios with a gaze-tracking device.

In this paper, we make the following contributions:

- We propose Voila-A, a novel approach for aligning VLMs with a user's gaze attention, and design innovative mechanisms to integrate gaze information into VLMs while preserving pre-trained knowledge.

- We leverage trace data from Localized Narratives to annotate instructional data using GPT-4, generating the VOILA-COCO dataset with 72k QA pairs, and demonstrate the scalability of this method.

- We evaluate Voila-A through a hold-out validation set and a newly collected VOILA-GAZE test set of real gaze samples, demonstrating that our approach significantly outperforms several baselines.

By aligning model attention with human gaze patterns and leveraging state-of-the-art techniques, we make a step forward in the development of more intuitive and user-centric VLMs, paving the way for more effective and engaging human-AI interaction in a wide range of applications

2 Leveraging Trace Data as an Alternative Approach to Align VLMs with Gaze Attention

In this section, we discuss the potential of trace data as a proxy for gaze data and propose a method for transforming trace data to make it more gaze-like, ultimately enabling the effective use of trace data for aligning VLMs with user gaze attention.

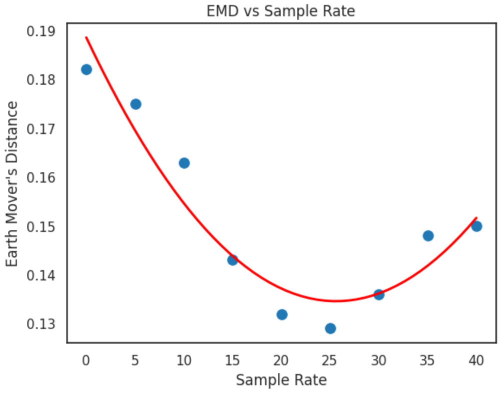


Figure 2: EMD between the mean heatmaps of 1k gaze and trace samples with varying sampling rates.

Obtaining gaze data for training VLMs can be challenging, as it is difficult to annotate and expensive to acquire. We propose that an alternative approach can be employed to align VLMs with user gaze attention: utilizing trace data, such as mouse traces. Localized Narratives [40](LN), a prior work, has annotated 849,000 images with mouse traces that are aligned with each word of the descriptions. The project involved 156 professional annotators who worked full-time, with annotator managers ensuring high-quality annotations through manual inspections and an automatic quality control mechanism. After discarding 23.5% of annotations, the remaining ones demonstrated a semantic accuracy of 98.0% for nouns and verbs. The accuracy of mouse traces in relation to object locations was also analyzed, revealing that most trace points were within the correct bounding box.

To demonstrate the similarities between gaze and mouse traces, we first collect hundreds of minutes of gaze data samples as described in C, and then we further collect mouse trace annotations on those samples. The mouse trace collection was done using the Bubbleview [21] interface, where the annotators were instructed to verbally repeat the questions and register their eye fixations with mouse clicks. **The resulting CC (cross-correlation) score between real gaze fixations and mouse traces was 0.82, and the NSS (normalized scanpath saliency) score was 2.57. These scores are comparable to those reported in the Bubbleview study [21] and are consistent with the performance of SOTA saliency models [7], indicating that gaze and mouse traces exhibit similarities, as users tend to fix their gaze on the target object when asking questions, a behavior also observed with mouse traces.** However, there are minor differences between the two, specifically in terms of gaze fixation continuity and the presence of noise points outside the target object at the end of a query. In the case of mouse traces, points that fell outside the bounding box were attributed to two factors: annotators often circled around objects, causing the traces to be near but not inside the box, and some annotators started moving the mouse before describing the object or vice versa. These observations provide valuable insights for properly leveraging trace data into the alignment process and understanding the relationship between gaze attention and language description.

In order to utilize trace data as a substitute for gaze data, we introduce a method to transform mouse traces, thereby reducing the discrepancies between the two data types and making the trace data more gaze-like. We first address the inherent noise in both trace points and gaze points by converting them into 2D heatmaps using Gaussian blur:

$$H(x,y)=\frac{1}{2\pi\sigma^2}e^{-\frac{x^2+y^2}{2\sigma^2}} \tag{1}$$

where $H(x,y)$ represents the heatmap value at position (x,y), and σ is the standard deviation of the Gaussian kernel.

Since mouse traces are more continuous than gaze fixations, we downsample the trace data to better resemble the distribution of gaze data. We investigate the Earth Mover's Distance (EMD) between the mean heatmaps of 1k gaze and trace samples while varying the sampling rate from 1 to 40:

$$\mathrm{EMD}(P,Q)=\frac{\sum_{i=1}^{n}|F_i(P)-F_i(Q)|}{\sum_{i=1}^{n}F_i(P)} \tag{2}$$

where P and Q are the distributions of the gaze and trace heatmaps, F_i denotes the cumulative distribution function, and n is the number of bins.

We observe that the EMD has a local minimum value around a sampling rate of 25 as shown in Figure 2. By selecting this optimal sampling rate, we can approximate the trace heatmap as an alternative to the real gaze heatmap from a statistical perspective. Consequently, this transformation mitigates the differences in inter-relationships, compactness, and noisiness between the trace and gaze data.

3 Method

3.1 Automatic Data Annotation For LN-COCO

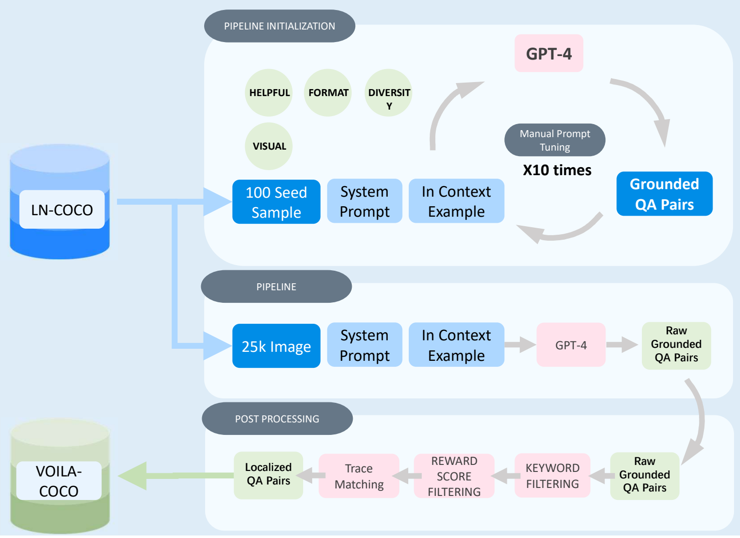


Figure 3: Automatic Data Annotation Pipeline

The automatic data annotation process for Voila-A is driven by the motivation to develop a more intuitive and user-centric VLM by aligning model attention with human gaze patterns. As shown in Figure 3, this process aims to create an effective and engaging human-AI interaction experience across various applications. To achieve this, we have designed an innovative annotating approach that leverages the capabilities of GPT-4 as a visual assistant to annotate trace-aligned instructional data to simulate the user's gaze attention. The data annotation process follows design principles to ensure accurate, relevant, and consistent annotations. These include: 1) focusing on referable sentences and appropriate tags, 2) using a conversational format with specific and general questions, 3) addressing various visual content aspects with definite answers, 4) incorporating complex questions while avoiding uncertainty, and 5) offering detailed, well-organized explanations.

As illustrated in Figure 3, the automatic data annotation pipeline comprises three stages.

Stage 1: Prompt Design Iteration. The first stage focuses on refining the prompt design. Let $S=\{(I_i,N_i,T_i,C_i)\}_{i=1}^{100}$ be a set of 100 samples from the LN-COCO dataset, where I_i represents the image, N_i the localized narrative, T_i the corresponding trace, and C_i the set of five captions from COCO-caption. We initiate the process with a basic system prompt, instructing GPT-4 to generate direct questions $Q_{i,j}^{D}$ and indirect questions $Q_{i,j}^{I}$ and corresponding answers $A_{i,j}$ that specifically reference the localized narratives while considering COCO-caption as background information. The referring portions are annotated with a unique marker $\mathcal{M}$ for trace matching during post-processing. We also provide two in-context examples to guide the model in generating helpful, well-formatted, diverse, and visually grounded QA pairs. Throughout each iteration k, we manually evaluate the quality of the generated grounded QA pairs and adjust the prompt to enhance their helpfulness, formatting, diversity, and visual relevance. After $K = 10$ iterations, we find the quality of most pairs to be satisfactory, and subsequently, we freeze the prompt to initiate the pipeline.

Stage 2: Data Sampling. In the second stage, we sample $N = 25,000$ image pairs from the LN-COCO dataset and obtain approximately $M = 75,000$ QA pairs.

Stage 3: Post-processing. The third stage involves post-processing the raw grounded QA pairs. This includes further filtering based on a set of keywords $\mathcal{K} = \{$"prompt", "this picture", "reference caption", ...$\}$. We define a filtering function $F_k(Q_{i,j}, A_{i,j}, \mathcal{K})$ that identifies and removes QA pairs containing meta descriptions of the prompt. We note that this issue may be further resolved by using GPT-4V, which was not available during our submission date. Additionally, we identify cases where answers are unhelpful, such as "I don't know" or "It's hard to tell." We find that these types of answers have low reward scores, so we further examine all pairs using a reward model [1] and filter the dataset by setting a minimum reward threshold τ. We define a filtering function $F_r(Q_{i,j}, A_{i,j}, \tau)$ that removes QA pairs with reward scores below τ. Finally, we segment each localized narrative into temporally aligned segments with respect to the special marker $\mathcal{M}$. Each segment comprises a grounded fact, a corresponding trace, a direct and indirect question, and an answer. This forms the final VOILA-COCO dataset, denoted as $\mathcal{D} = \{(F_i, T_i, Q^D_{i,j}, Q^I_{i,j}, A_{i,j})\}$. It is worth noting that we did not utilize all localized narratives, leaving room for future exploration. We annotate the COCO subset of localized narratives, resulting in the Voila-COCO dataset, with statistics presented in Table 1.

The finalized prompt can be found in E. We also visualize a sample of our annotated data in Figure 14. Data quality analysis can be found in Section D. By adhering to these design principles, the automatic data annotation process ensures that the resulting dataset is of high quality and effectively aligns the VLM's attention with that of a human user.

3.2 VOILA-GAZE: Real-life gaze-QA pairs

To further demonstrate the effectiveness of our method in aligning VLMs with real-life users' gaze attention, we conduct experiments in two everyday scenarios, encompassing a variety of question types details can be found in Table 5.

Table 1: Statistics of Voila-COCO and Voila-Gaze Datasets, SR refers to Survival Rate from raw data after filtering

Dataset	Split	#Images	#Questions	SR
Voila-COCO	Training	20000	70000	93.5%
Voila-COCO	Validation	100	550	71.1%
Voila-COCO	Test	500	1900	75.7%
Voila-GAZE	Real-life	200	200	18.2%

In addition to the recorded gaze trajectory, video, and transcription, each participant is instructed to annotate the key elements of their questions, formulate clear questions based on their interpretations, and choose the best answer from three candidate answers generated by GPT-4 according to their annotations. The experiment includes 16 participants (8 per scenario) with an equal gender distribution, aged between 20 and 30 (with a standard deviation of 2.06). Each participant takes approximately 240 minutes to complete the study. After applying post-filtering and manual checking, we curate a set of 200 QA pairs as our real-life benchmark, VOILA-GAZE. The curation process is conducted by two individuals sequentially, with the second person double-checking the following aspects: **1. The question is related and aligned with gaze. 2. The answer is meaningful and can be considered a proper response to the gaze-directed question. 3. The question is not related to specific brands, prices, or any other objects beyond general knowledge. 4. The question type is not biased towards a few simple patterns.** This two-step process ensures the quality and relevance of the curated data while minimizing potential biases and maintaining a focus on general knowledge. Samples of VOILA-GAZE are shown in Figure 9.

3.3 Model Design

In developing our design, a critical consideration is the adherence to the established architecture of current VLMs. **It is essential to avoid introducing a significant number of new parameters**

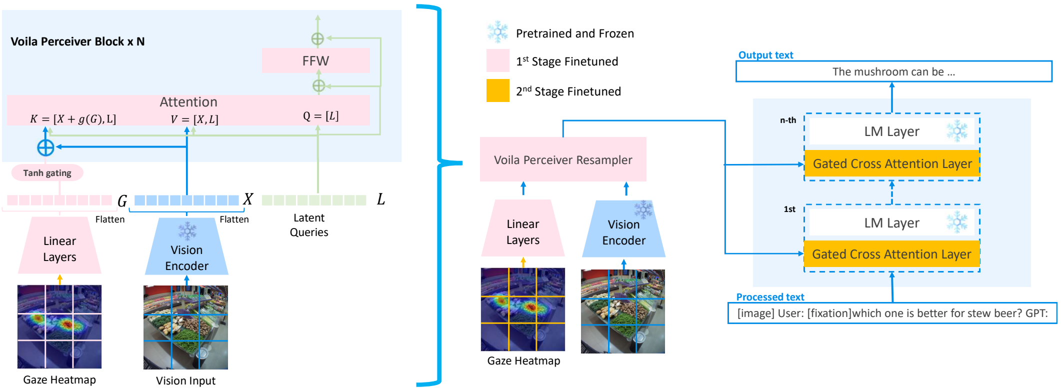


Figure 4: Architecture of the VOILA Model: On the left, gaze fixation is transformed into a heatmap, which is subsequently processed through linear layers to encode the visual attention. This encoded data is then segmented into discrete patches that are spatially correlated with corresponding image patches. These gaze patches are further refined into key embeddings, which undergo modulation by a gating mechanism, designed to incrementally integrate gaze data. The resulting gaze and image key embeddings are then combined and subjected to a self-attention mechanism, which synthesizes the information into a cohesive set of latent perceiver embeddings. On the right, the figure delineates the integration pathway where the gaze heatmap and the image concurrently enter the VOILA Perceiver. This integrated input is subsequently directed through gated cross-attention modules before progressing into the language model layers, culminating in a unified output that encapsulates the interplay between visual attention and linguistic processing.

or making extensive structural modifications. This constraint is due to the limitations of the current gaze dataset, which does not support large-scale pretraining. Additionally, we must be vigilant in preventing catastrophic forgetting during the fine-tuning process. Therefore, we aim to implement only those modifications that are proven to be necessary and beneficial through our ablation studies. These changes are carefully selected to enhance performance without compromising the stability and efficiency of the model.

We employ the model architecture from OpenFlamingo, as illustrated on the right side of Figure 4. This framework consists of a pre-trained vision encoder, language decoder, and gated cross-attention layers, offering flexibility for multi-image and multi-turn conversational interactions. The primary challenge lies in incorporating gaze instructional signals into a pre-trained VLM. To tackle this issue, we initially developed several potential solutions, which are discussed in Sec 4.3.2 and Fig 16.

Based on empirical evidence, we ultimately confirm the effectiveness of the Voila Perceiver Resampler solution. The Voila Perceiver Resampler(VPR) processes the input image features and gaze information, and then feeds them into a series of Voila Perceiver Blocks (VPB):

$$\begin{aligned}\mathbf{VPR}(X,G) =& \mathrm{LN}(\mathbf{VPB}_n(X,G)), \mathbf{VPB}_0(X,G) = \mathrm{VPB}(X, L_0, G),\\ \mathbf{VPB}_i(X,G) =& \mathrm{VPB}(X, \mathbf{VPB}_{i-1}(X,G), G) \quad \text{for } i = 1,2,\ldots,m.\end{aligned} \tag{3}$$

The input hidden states of VPR are denoted as $X \in \mathbb{R}^{B\times H\times L_I\times D}$, $L \in \mathbb{R}^{B\times H\times L_L\times D}$and $G \in \mathbb{R}^{B\times H\times L_H\times D}$, where B is the batch size, $L_L$and$L_I = L_H$ are the lengths of latent tokens and image/heatmap patches, respectively, H is the number of attention heads, and D is the hidden size. In which X represents the image features, G is the gaze heatmap embedding patches. L denotes the latent features, which are introduced from the original Perceiver as a small set of latent units that forms an attention bottleneck through which the inputs must pass. To obtain the gaze information $G \in \mathbb{R}^{B\times H\times L_H\times D}$, we first divide the gaze heatmap $G' \in \mathbb{R}^{B\times h\times w}$ into patches. Then, we apply a linear transformation followed by layer normalization. The process can be represented by the following equation:

$$G = \mathrm{LN}(\mathrm{Linear}(\mathbf{patch}(G'))) \tag{4}$$

The VPR comprises a series of Voila Perceiver Blocks (depicted on the left side of Figure 4). This mechanism leverages gaze information to enhance visual feature perception. Our design adheres

https://doi.org/10.52202/079017-0060

to the principle that the gaze serves as an information aggregator in the attention process without disrupting the original learned distribution. The Voila Perceiver Block(VPB) is defined as follows:

$$\text{VPB}(X, L, G) = \text{LN}(L + \text{FF}(L + \text{GA}(X, L, G))) \tag{5}$$

The feed-forward network, $\text{FF}(\cdot)$, is a sequence of layer normalization, linear transformation, GELU activation, and another linear transformation. The attention mechanism, $\text{GA}(X, L, G)$, is computed as follows:

$$\begin{gathered} Q = W_L^Q L,\ V = W_{XL}^V (X \oplus L),\ K = W_{XL}^K (X \oplus L) + W_G^K (G * tanh(\mathbf{g}) \oplus P) \\ \text{GA}(X, L, G) = \text{Softmax}(\frac{QK^\top}{\sqrt{D}})V \end{gathered} \tag{6}$$

Here, $\oplus$ denotes concatenation along the feature dimension, and P is zero padding with the same shape as L. And $\mathbf{g}$ is a learnable gating parameter attributed to each block for fading in gaze guidance gradually. W_L,W_{XL},W_G are the QKV matrices of corresponding input.

3.4 Training

Our approach utilizes the OpenFlamingo training paradigm to train the Voila model, building upon the pre-trained weights of the Otter model, which incorporates an MPT-7B [47] language encoder and a CLIP ViT-L/14 [42] vision encoder. To avoid overfitting and maximize the benefits of pre-trained knowledge, we initially freeze both encoders. As shown in Figure 4, we then train only the linear layers directly related to gaze input at the first stage for one epoch before fine-tuning the entire Perceiver resampler module, the cross-attention layers integrated into the language encoder, and the input/output embeddings of the language encoder in the second stage for an additional epoch. This process results in roughly 1.3 billion trainable parameters for the Otter model. Note that the $\mathbf{g}$ in each **VPB** are initialized as 0.

During training, we adhere to a specific format for preparing our training data. This format combines an image, user instruction, "GPT"-generated answers 1, and a unique token known as the [endofchunk] token. We arrange the training data as follows:

<context> [image] User:[fixation]<instruction> GPT:[answers] <answer>.[endofchunk]

Here, the [image], [answer], [fixation], and [endofchunk] tokens are distinct and serve particular functions. We adopt a chatbot-like format to enhance the instruction-following capabilities and conversational generalizability of our model. The [image] and [endofchunk] tokens originate from the OpenFlamingo training paradigm, while the [answer] token is a new addition by Otter. The [answer] token separates answers from instructions, allowing us to mask all tokens following the [answer] token during training and designate them as the model's prediction objectives. We also introduce the [fixation] token to direct the language model to utilize gaze information. We train our model using a cross-entropy loss function.

4 Experiment

4.1 Evaluation metrics

GPT-4 RANKING We utilize GPT-4 RANKING as our primary automated evaluation metric to assess model performance through a one-to-one comparison. The GPT Ranking represents the language model's evaluation of the quality of the generated response. This score signifies the extent to which the response aligns with the ground truth image description and answer while demonstrating the model's language proficiency. Factors such as grammar, semantics, and fluency are taken into account when comparing the response to that of another model. **It is important to note that GPT-4 exhibits sequence ordering bias.** To mitigate this issue, we implement a dual-setting approach that reverses the order of the models, ensuring that the order does not influence the outcome. The prompt and evaluation procedure can be found in Figure 15.

Reward Score Given that our dataset is automatically annotated using GPT-4, it is crucial to mitigate any potential system bias during model evaluation. To this end, we incorporate human preference by utilizing a reward model score as an auxiliary metric. The reward model, which assesses

https://doi.org/10.52202/079017-0060

the human-like quality of a response, is trained using human feedback to predict the superiority of a generated answer in relation to a given question from a human perspective [1]. This approach allows us to achieve a more balanced and robust evaluation process, ensuring that our model's performance aligns with human expectations and preferences.

4.2 Main Results

4.2.1 VOILA Exhibits a Balanced Capability Between Helpfulness and Fact Grounding

In Figure 5, we observe a notable superiority of VOILA over both Otter and Kosmos-2 on the VOILA-COCO-TESTSET. Regarding the *grounding* capability, both VOILA and Kosmos-2 trained with fine-grained grounded facts outperform Otter model to a large extent. Besides, VOILA surpasses Kosmos-2 marginally. With respect to *helpful* capability, Otter delivers significantly more helpful responses than Kosmos-2. Since Otter is trained on top of Openflamingo with an instruction-following dataset, it can provide a more helpful response, especially for informative queries while Kosmos-2 tends to answer visual observation from the input image. In addition, VOILA trained on gaze dataset demonstrates stronger helpful capabilities over all models.

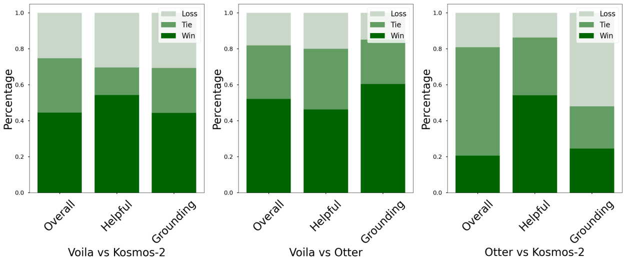


Figure 5: GPT-RANKING ON VOILA-COCO-Testset

In real gaze scenarios, as illustrated in Figure 6, VOILA outperforms the two baseline models as well. These scenarios differ substantially from the collected COCO images and present more challenging questions, necessitating a higher degree of accurate intent understanding and reasoning. Especially from the comparison of Otter vs Kosmos-2, we found that there are many more Tie results due to the hardness of the real cases. Despite these increased demands, VOILA continues to surpass both models, further showcasing its balanced proficiency in both helpfulness and fact grounding.

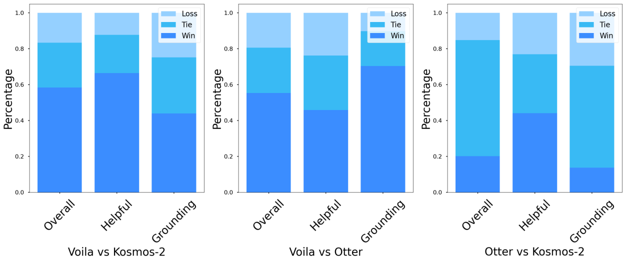


Figure 6: GPT-RANKING ON VOILA-GAZE

4.3 Ablation studies

4.3.1 Query types has a significant impact on Response Quality

Table 2 investigates the varying performance of different question types, specifically direct and implicit/coreference queries. As the base model Openflamingo was pre-trained on direct queries, both Otter and VOILA performed better in this category, as expected. In addition, it is natural for humans to communicate with coreference queries. VOILA maintained strong performance when handling coreference queries with the gaze as guidance while the Otter model greatly decreased. Furthermore, we appended in-context QA pairs prior to the current query and observed that the examples further improved the quality of the responses. In real-life situations, multi-turn conversations are involved in most interactions with many coreference queries. The design

Table 2: Ablation on query types, WR means Winning Rate Over Otter-base

METHODS	QUESTION TYPES	WR	LR	REWARD SCORE
OTTER-BASE	COREFERENCE QUERY	-	-	-1.91
OTTER-BASE	DIRECT QUERY	0.51	0.1	0.02
VOILA	COREFERENCE QUERY	0.41	0.18	-0.79
VOILA	DIRECT QUERY	0.62	0.15	0.14
VOILA	IN-CONTEXT PROMPT + COREFERENCE QUERY	0.46	0.16	-0.02
VOILA	IN-CONTEXT PROMPT + DIRECT QUERY	0.77	0.12	0.20

https://doi.org/10.52202/079017-0060

of an in-context prompt can assist VOILA to demonstrate a superior ability. This improvement is evident across both direct and coreference query types.

4.3.2 Heatmap is a better way to incorporate gaze

To establish the effectiveness of our approach, we implemented several alternative methods for incorporating gaze data into VLMs. These methods include: converting gaze sequences into discrete position tokens for LLMs, using the bounding box position of trace trajectories as additional patch tokens concatenated to VIT image feature token lists, and converting the bounding box coordinates into discrete tokens. We illustrate these methods in Figure 16. However, all these methods failed to outperform the gaze heatmap approach, as shown in Table 3.

Table 3: Ablation on Methods of Integrating Gaze Data, WR means Wining Rate over Otter-base

Methods	WR	LR	Reward Score
Otter-base	-	-	-1.91
Gaze as discrete position tokens	0.19	0.25	-2.44
Gaze *bounding box* as image patch	0.36	0.20	-1.26
Gaze *bounding box* as discrete position tokens	0.21	0.22	-1.72
Voila(Gaze as heatmap)	0.41	0.18	-0.79

4.3.3 Gradual Unfreezing of Parameters Yields Better Results

Table 4 presents empirical findings that demonstrate the effectiveness of gradually unfreezing model parameters. Instead of directly unfreezing the vision perceiver and cross-attention layers, or using LORA to fine-tune the entire model, we first fine-tuned the gaze-related weights and then fine-tuned the other parts of the perceiver and cross-attention layers, which yielded better results. We hypothesize that this improvement is due to the newly added gaze component needing to adapt to the distribution of the pre-trained layers first. This adaptation process can further help mitigate the issue of catastrophic forgetting.

Table 4: Ablation on Training Procedure, WR means Wining Rate over Otter-base

Layers fine-tuned	WR	LR	Reward Score
Otter-base frozen all	-	-	-1.91
Otter-base vision perceiver+cross attention	0.25	0.24	-1.78
Voila gaze weight	0.24	0.20	-1.52
Voila gaze weight+LORA	0.23	0.21	-1.02
Voila gaze weight→perciever+cross attention	0.41	0.18	-0.79

4.4 Qualitative studies

We conducted qualitative studies on randomly selected cases and demonstrated the results of several representative examples in Appendix Figure 8. According to the analysis, the conclusions can be summarized as follows: 1) Existing models are able to generate reasonable results for **explicit** queries. In the 1st row, the object *cakes* and the attributes *color* are explicitly mentioned in the query, and the three models can answer (partially) correctly; 2) Regarding **coreference** queries, the model Otter is hard to understand the pronouns like *it* without spatial guidance as shown in the 2nd row. This requires further context or generates the answer based on the salient object like *plane* instead of the actual human attention; 3) The Kosmos-2 model can take the **bounding box** for grounding as spatial guidance, it is sometimes not accurate compared to the heatmap used in VOILA. As shown in the 3rd row, the bounding box is too coarse and made the model focus on the object *plane* instead of the actual human attention *sky*; 4) Besides, we found that Kosmos-2 tends to describe the detailed visual content and sometimes lacks the **instruction-following** capability; In the 4th row, the Kosmos-2 responses *Keyboard* depicted in the bounding box ignoring the actual query intention; Finally, 5) There are still further challenges for all models to deal with. For instance, counting for objects requires intensive fine-grained recognition of the visual content demonstrated in the last row.

5 Conclusion

In this study, we presented Voila-A, a cutting-edge approach that aligns Vision-Language Models with user gaze attention. Voila-A can be implemented in HMD AR/VR devices as an egoview copilot, benefiting a wide range of users, including visually impaired individuals who rely on their gaze to communicate their intent. This method surpasses the capabilities of similar mobile apps that necessitate users to lift their phones for scene capture. We successfully utilized trace data to create the

VOILA-COCO dataset, showcasing Voila-A's superior performance in two benchmarks. Our research lays the foundation for more engaging human-AI interactions and encourages further exploration of Voila-A's integration with various modalities and tasks in the realm of multimodal AI systems.

6 Acknowledgements

This work was conducted during Kun Yan's internship at Microsoft Research Asia and supported in part by NSFC No. 61925203, U22B2021, NSFC No. 62472244 and No. 62132010.

References

[1] Openassistant/reward-model-deberta-v3-large-v2. `https://huggingface.co/OpenAssistant/reward-model-deberta-v3-large-v2`, 2023.

[2] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

[3] A. Awadalla, I. Gao, J. Gardner, J. Hessel, Y. Hanafy, W. Zhu, K. Marathe, Y. Bitton, S. Gadre, S. Sagawa, et al. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

[4] L. Bracha, E. Shaar, A. Shamsian, E. Fetaya, and G. Chechik. Disclip: Open-vocabulary referring expression generation. *arXiv preprint arXiv:2305.19108*, 2023.

[5] K. Chen, Z. Zhang, W. Zeng, R. Zhang, F. Zhu, and R. Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

[6] M.-C. Chen, J. R. Anderson, and M.-H. Sohn. What can a mouse cursor tell us more?: correlation of eye/mouse movements on web browsing. *CHI '01 Extended Abstracts on Human Factors in Computing Systems*, 2001.

[7] S. Chen, N. Valliappan, S. Shen, X. Ye, K. Kohlhoff, and J. He. Learning from unique perspectives: User-aware saliency modeling. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2701–2710, Los Alamitos, CA, USA, jun 2023. IEEE Computer Society.

[8] J. Cho, J. Lei, H. Tan, and M. Bansal. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR, 2021.

[9] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.

[10] A. Das, H. Agrawal, C. L. Zitnick, D. Parikh, and D. Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *ArXiv*, abs/1606.03556, 2016.

[11] T. Gong, C. Lyu, S. Zhang, Y. Wang, M. Zheng, Q. Zhao, K. Liu, W. Zhang, P. Luo, and K. Chen. Multimodal-gpt: A vision and language model for dialogue with humans. *arXiv preprint arXiv:2305.04790*, 2023.

[12] Q. Guo and E. Agichtein. Towards predicting web searcher gaze position from mouse movements. *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, 2010.

[13] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[14] J. Huang, R. W. White, and G. Buscher. User see, user point: gaze and cursor alignment in web search. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2012.

https://doi.org/10.52202/079017-0060

[15] J. Huang, R. W. White, and S. T. Dumais. No clicks, no problem: using cursor movements to understand and improve search. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2011.

[16] M. Ilaslan, C. Song, J. Chen, D. Gao, W. Lei, Q. Xu, J. Lim, and M. Shou. Gazevqa: A video question answering dataset for multiview eye-gaze task-oriented collaborations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10462–10479, 2023.

[17] M. Jiang, S. Huang, J. Duan, and Q. Zhao. Salicon: Saliency in context. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[18] W. Jin, S. Mukherjee, Y. Cheng, Y. Shen, W. Chen, A. H. Awadallah, D. Jose, and X. Ren. Grill: Grounded vision-language pre-training via aligning text and image regions. *arXiv preprint arXiv:2305.14676*, 2023.

[19] T. Judd, K. A. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. *2009 IEEE 12th International Conference on Computer Vision*, pages 2106–2113, 2009.

[20] W. Kienzle, F. Wichmann, B. Scholkopf, and M. O. Franz. A nonparametric approach to bottom-up visual saliency. In *Neural Information Processing Systems*, 2006.

[21] N. W. Kim, Z. Bylinskii, M. A. Borkin, K. Z. Gajos, A. Oliva, F. Durand, and H. Pfister. Bubbleview: an interface for crowdsourcing image importance maps and tracking visual attention. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 24(5):36, 2017.

[22] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2017.

[23] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

[24] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human neurobiology*, 4 4:219–27, 1985.

[25] M. Kümmerer, T. S. A. Wallis, and M. Bethge. Deepgaze ii: Reading fixations from deep features trained on object recognition. *ArXiv*, abs/1610.01563, 2016.

[26] M. F. Land. Eye movements and the control of actions in everyday life. *Progress in retinal and eye research*, 25(3):296–324, 2006.

[27] B. Li, Y. Zhang, L. Chen, J. Wang, J. Yang, and Z. Liu. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*, 2023.

[28] J. Li, D. Li, S. Savarese, and S. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[29] Z. Lin, Z. Zhang, L.-Z. Chen, M.-M. Cheng, and S.-P. Lu. Interactive image segmentation with first click attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13339–13348, 2020.

[30] Z. Lin, Z. Zhang, L.-H. Han, and S.-P. Lu. Multi-mode interactive image segmentation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 905–914, 2022.

[31] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.

[32] J. Liu, L. Wang, and M.-H. Yang. Referring expression generation and comprehension via attributes. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4856–4864, 2017.

[33] C. Ma, L. Zhao, Y. Chen, S. Wang, L. Guo, T. Zhang, D. Shen, X. Jiang, and T. Liu. Eye-gaze-guided vision transformer for rectifying shortcut learning. *IEEE Transactions on Medical Imaging*, 2023.

[34] A. Mani, N. Yoo, W. Hinthorn, and O. Russakovsky. Point and ask: Incorporating pointing into visual question answering. *arXiv preprint arXiv:2011.13681*, 2020.

[35] E. Niebur and C. Koch. Control of selective visual attention: Modeling the where pathway. In *Neural Information Processing Systems*, 1995.

[36] OpenAI. Gpt-4 technical report, 2023.

[37] J. Pan, E. Sayrol, X. G. i Nieto, K. McGuinness, and N. E. O'Connor. Shallow and deep convolutional networks for saliency prediction. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 598–606, 2016.

[38] Z. Peng, W. Wang, L. Dong, Y. Hao, S. Huang, S. Ma, and F. Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

[39] R. Piening, R. Piening, K. Pfeuffer, A. Esteves, T. Mittermeier, S. Prange, P. Schröder, and F. Alt. Looking for info: Evaluation of gaze based information retrieval in augmented reality. *IFIP TC13 International Conference on Human-Computer Interaction*, 2021.

[40] J. Pont-Tuset, J. Uijlings, S. Changpinyo, R. Soricut, and V. Ferrari. Connecting vision and language with localized narratives. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 647–664. Springer, 2020.

[41] K. Qian, Z. Zhang, W. Song, and J. Liao. Gvgnet: Gaze-directed visual grounding for learning under-specified object referring intention. *IEEE Robotics and Automation Letters*, 2023.

[42] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision, 2021.

[43] E. Sood, F. Kögel, F. Strohm, P. Dhar, and A. Bulling. Vqa-mhug: A gaze dataset to study multimodal neural attention in vqa. In *Proc. ACL SIGNLL Conference on Computational Natural Language Learning (CoNLL)*, pages 27–43. Association for Computational Linguistics, 2021.

[44] Y. Sugano and A. Bulling. Seeing with humans: Gaze-assisted neural image captioning. *ArXiv*, abs/1608.05203, 2016.

[45] V. Tanriverdi and R. J. Jacob. Interacting with eye movements in virtual environments. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 265–272, 2000.

[46] H. R. Tavakoli, F. Ahmed, A. Borji, and J. T. Laaksonen. Saliency revisited: Analysis of mouse movements versus fixations. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6354–6362, 2017.

[47] M. N. Team. Introducing mpt-7b: A new standard for open-source, commercially usable llms, 2023. Accessed: 2023-05-05.

[48] M. Tonsen, C. K. Baumann, and K. Dierkes. A high-level description and performance evaluation of pupil invisible. *arXiv preprint arXiv:2009.00508*, 2020.

[49] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[50] A. B. Vasudevan, D. Dai, and L. V. Gool. Object referring in videos with language and human gaze. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4129–4138, 2018.

[51] P. Voigtlaender, S. Changpinyo, J. Pont-Tuset, R. Soricut, and V. Ferrari. Connecting vision and language with video localized narratives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2461–2471, 2023.

[52] P. Wang, A. Yang, R. Men, J. Lin, S. Bai, Z. Li, J. Ma, C. Zhou, J. Zhou, and H. Yang. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR, 2022.

[53] K. Yan, L. Ji, H. Luo, M. Zhou, N. Duan, and S. Ma. Control image captioning spatially and temporally. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2014–2025, 2021.

[54] Y. Yao, Q. Chen, A. Zhang, W. Ji, Z. Liu, T.-S. Chua, and M. Sun. Pevl: Position-enhanced pre-training and prompt tuning for vision-language models. *arXiv preprint arXiv:2205.11169*, 2022.

[55] A. L. Yarbus. Eye movements and vision. In *Springer US*, 1967.

[56] Q. Ye, H. Xu, G. Xu, J. Ye, M. Yan, Y. Zhou, J. Wang, A. Hu, P. Shi, Y. Shi, et al. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023.

[57] R. Zhang, J. Han, A. Zhou, X. Hu, S. Yan, P. Lu, H. Li, P. Gao, and Y. Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023.

[58] R. Zhang, A. Saran, B. Liu, Y. Zhu, S. Guo, S. Niekum, D. H. Ballard, and M. M. Hayhoe. Human gaze assisted artificial intelligence: A review. *IJCAI : proceedings of the conference*, 2020:4951–4958, 2020.

[59] S. Zhang, P. Sun, S. Chen, M. Xiao, W. Shao, W. Zhang, K. Chen, and P. Luo. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv e-prints*, pages arXiv–2307, 2023.

[60] Y. Zhong, J. Yang, P. Zhang, C. Li, N. Codella, L. H. Li, L. Zhou, X. Dai, L. Yuan, Y. Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022.

[61] Q. Zhou, C. Yu, S. Zhang, S. Wu, Z. Wang, and F. Wang. Regionblip: A unified multi-modal pre-training framework for holistic and regional comprehension. *arXiv preprint arXiv:2308.02299*, 2023.

[62] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[63] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004, 2016.

https://doi.org/10.52202/079017-0060

A Limitation and Discussion

A.1 Future work

Despite its promising potential, there are limitations and future directions to consider. These include enhancing inference efficiency for real-time responses, integrating voice modalities for seamless interaction, and supporting higher resolutions for OCR and screen or UI understanding.

A.2 Hallucination

During inference, our method occasionally exhibits hallucinations related to image content. This issue may stem from the limited number of training samples and the imperfect integration of visual hidden distributions into the language decoding process during vision language pretraining. Recognizing the potential for further scaling of our method, we consider addressing these hallucinations as a future research direction.

A.3 Comparison to GPT-4-V

With the release of GPT-4-V, which features vision capabilities, during the submission of this paper, we find it pertinent to include a brief discussion on the relationship between our work and GPT-4-V. Although GPT-4-V demonstrates remarkable visual capabilities in its demos, it surpasses our method in terms of visual understanding abilities. Nonetheless, our work remains valuable as it presents an effective approach to incorporating user sensory information for generating more relevant responses.

As reported, GPT-4-V occasionally struggles to accurately capture a user's intent when referencing specific elements within an image, prompting the design of an interface that allows users to directly draw sketches for highlighting purposes. Our method has the potential to enhance this user experience and can be extended to more dynamic scenarios such as virtual reality and augmented reality.

B Related Work

B.1 Multimodal Large Language Model

Recent research works for multimodal vision and language tasks relied on multimodal large language models(MLLM) and demonstrated superior performance. One line is to learn **alignment** between the vision input and the text token for LLM. LLaVA [31] directly feeds visual features to the LLM using only a learnable fully connected layer. BLIP-2 [28] proposed Q-Former to extract a fixed number of query features from visual features, which are aligned with the text in frozen LLM. Another direction is to design **learnable interaction layers** to attend the vision input to the frozen LLM layers. Flamingo [2] adopts a perceiver resampler module to convert visual features into visual tokens and interleave the tokens in plain text at the locations of vision inputs. Besides, Flamingo performed learnable cross attention to attend to the visual tokens of the image that appeared just before it in the interleaved sequence, rather than to all previous images and built a large-scale interleaved image-text dataset. On top of these designs, recent works mainly focus on improving **instruction-following ability**. LLaMA-Adapters [57] aims to adapt LLaMA [49] into an instruction-following model with an additional adapters module and multi-modal prompts. Mini-GPT4 [62], mPLUG-OWL [56], and InstructBLIP [9] adopt the Q-Former on various language models for instruction following capability. Besides, MultiModal-GPT [11] fine-tuned OpenFlamingo [3] using Low-rank Adapter (LoRA) [13] and Otter [27] introduced MultI-Modal In-Context Instruction Tuning (MIMIC-IT) dataset following three heuristics, both of which demonstrate improved instruction-following ability, Simultaneously.

Grounded MLLM Inspired by the success of MLLM, recent works focus on investigating the fine-grain grounding capability between region-text pairs instead of image-text pairs, and further conduct dense regional prediction tasks. One research line is to learn **regional alignment** between the image regions with the corresponding text tokens. RegionCLIP [60] extends CLIP with pseudo image regions and textual concept pairs. Grill [18] proposes to replace the referring words with the corresponding visual patches to align text and image Regions. RegionBLIP [61] takes position-assisted regional objects as soft prompts for LLM on image-region-text data. Another research focus is to unleash the **grounding ability** in a multimodal large language model. VL-T5 [8] converts the visual grounding task into regional *ROI box* feature conditioned text generation to predict the box

https://doi.org/10.52202/079017-0060

id. OFA [52], PEVL [54] and KOSMOS-2 [38] reformulate continuous corner coordinates of the object to *discrete position tokens*. Shikra [5] handles *spatial coordinate* inputs and outputs in natural language without introducing extra vocabulary or position encoders. The works [5, 38, 59] also perform Instruction tuning and convert the position of regional objects into language descriptions. Although the gaze is flexible and interactive, it is easy for humans to understand the gaze's semantic representation but hard for AI agents.

B.2 Region Representation for Large Language Models

The visual region can be represented as **bounding boxes** [63, 32], **points** [34], **traces** [40, 53]. Existing approaches usually leverage Fast-RCNN to detect bounding boxes which limits the predefined or recognized objects in the bounding box and hard to scale out. Points are flexible but are too fine-grained and require a large number of points to represent large regions precisely. Trace is a more natural way to input by using the mouse trace coordinates and is most similar to human's gaze. In AR and VR scenarios, although trace is applicable with gesture, we propose to use gaze more conveniently and interactively. The two works [5, 34] take bounding boxes or points as region input for visual question answering and are the most similar work. Different from them, we take **gaze** as regional inputs.

Region Inputs In order to input the regional information to the model, several methods [59, 4] directly **concatenate cropped image patches** with the original text/image as model input. Another method [29, 30] uses 0/1 **mask or Gaussian map** input with the original image to emphasize the area of user interest. Additionally, other methods [23, 51] first encode points, boxes, or trace to **positional encodings** then add them to intermediate features or learned queries. Specifically for gaze, [41] propose a gaze-directed visual grounding and *fuse* the gaze feature through a multi-modal fusion module. EG-ViT [33] propose a eye-gaze-guided vision transformer which takes the *masked image patches* within the gaze interest.

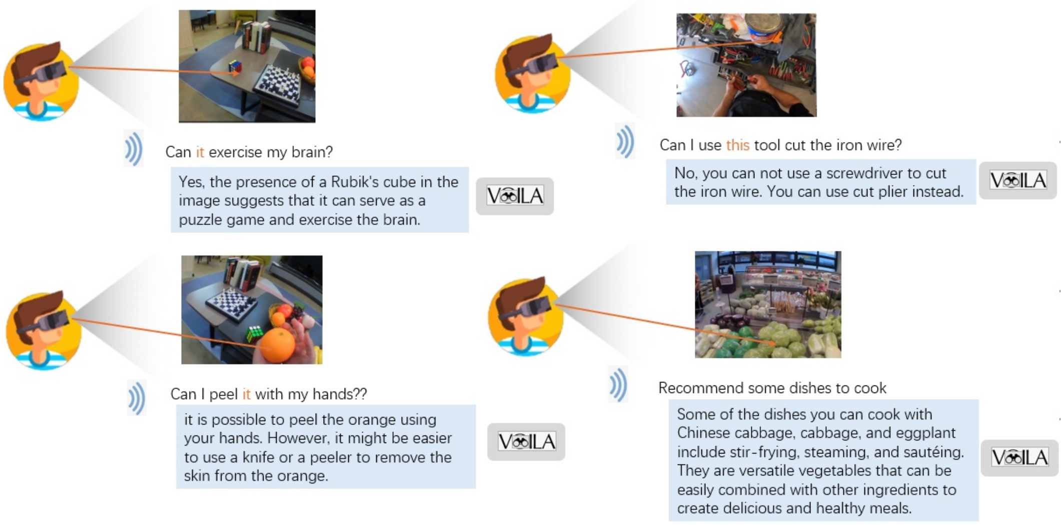
B.3 Gaze and Cursor as a proxy for Attention

Cursor-based techniques, including approaches like SALICON [17] and BubbleView [21], have emerged as affordable, nonintrusive, and scalable alternatives to traditional eye-tracking methods for collecting human attentional data. Empirical evidence from prior work has established strong connections between cursor-like signals and gaze positions. Studies focusing on web browsing and search tasks have found a high correlation between cursor and gaze locations, with better alignment along the vertical dimension [15, 14, 12, 6]. These findings support the motivation to use cursor-based techniques as a proxy for attention.

Despite their success, existing cursor-based studies have limitations, such as the need for complex post-processing of mouse movement data and evaluations limited to simple aggregate comparisons with eye-tracking data [21]. Furthermore, while prior work serves as a solid foundation from a data-centric perspective, it lacks a demonstration of whether modern applications aiming to assist users using gaze, such as vision language models (VLMs), can be trained from cursor data and later adapt to gaze signals, especially when transitioning from 2D planar images on screens to ego view scenes in head-mounted display (HMD) scenarios. Our work aims to directly tackle this problem, as we believe it is the optimal time to close the entire visionary loop of understanding and utilizing the gaze modality to ultimately achieve smart, in situ personal assistants.

B.4 Saliency models on modeling gaze attention

[55] proposed that tasks could be decoded from fixation patterns, receiving mixed support in subsequent research. Early computational models of visual attention focused on bottom-up approaches, representing pre-attentive selection processes from visual input [24]. Later, the saliency map concept emerged [35]. Initially, models were trained on fixation data from eye-tracking experiments [20, 19], but collecting large datasets proved difficult. The SALICON dataset [17] addressed this challenge by using mouse movements to simulate natural viewing behavior, leading to state-of-the-art performance in saliency models [17, 37, 46]. As deep learning advanced, saliency modeling improved [25], enabling more complex gaze pattern modeling in vision-language tasks [44, 10, 50].



Can it exercise my brain?

Yes, the presence of a Rubik's cube in the image suggests that it can serve as a puzzle game and exercise the brain.

Can I use this tool cut the iron wire?

No, you can not use a screwdriver to cut the iron wire. You can use cut plier instead.

Can I peel it with my hands??

it is possible to peel the orange using your hands. However, it might be easier to use a knife or a peeler to remove the skin from the orange.

Recommend some dishes to cook

Some of the dishes you can cook with Chinese cabbage, cabbage, and eggplant include stir-frying, steaming, and sautéing. They are versatile vegetables that can be easily combined with other ingredients to create delicious and healthy meals.

Figure 7: Different Scenarios in the Future

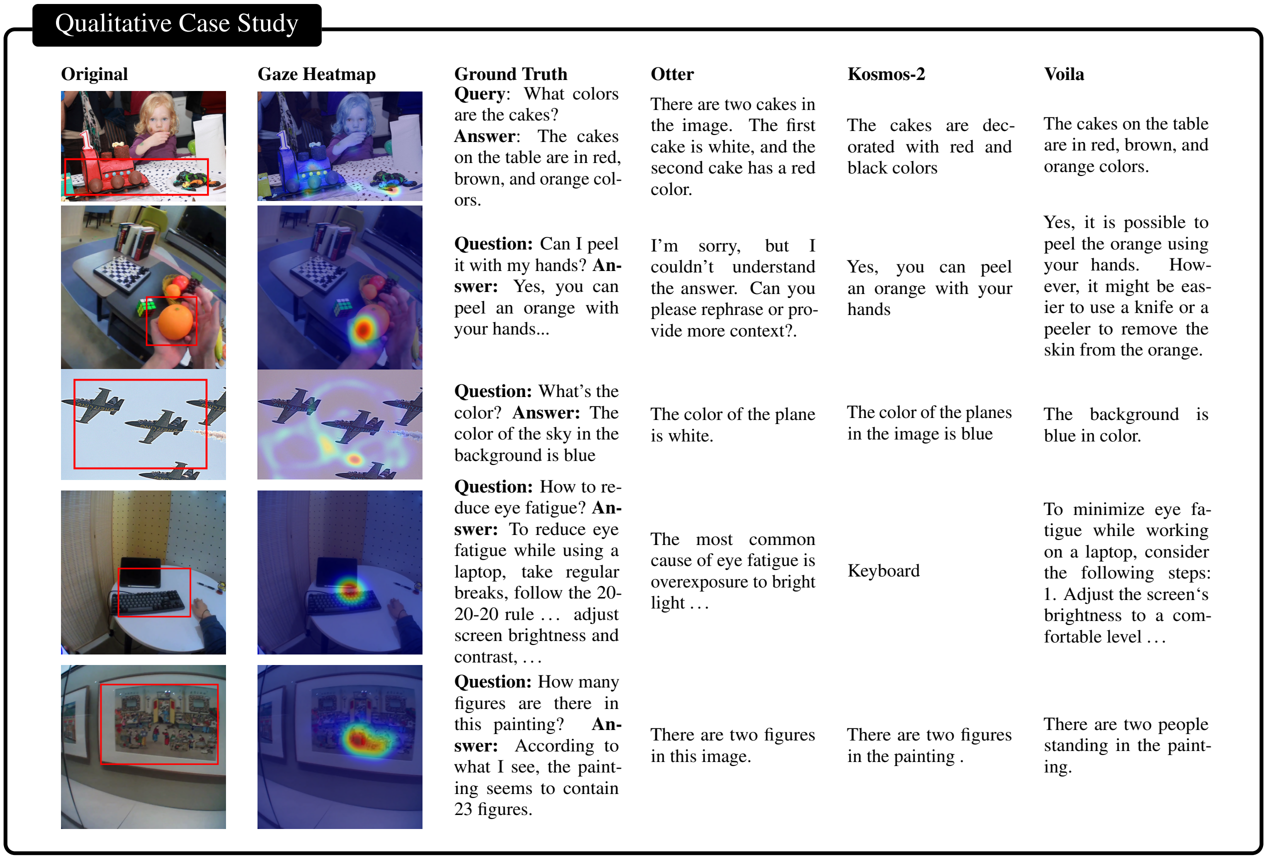
Saliency models aim to approximate the human visual system by predicting eye fixations on images [21]. Unlike traditional saliency models, **our approach takes ground truth gaze data, image, and natural language inputs to generate contextually relevant responses**, presenting a novel challenge. Recent work, such as [44], leverages gaze signals to enhance captioning tasks but does not accommodate dynamic user queries beyond captioning. Additionally, their LSTM-based method falls short compared to contemporary large transformer baselines. [43] introduces gaze data to the visual question-answering (VQA) task, but their analysis remains limited to comparing human and neural attentive strategies learned by VQA models. Besides, a recent attempt [16] leverages gaze for video question-answering in industrial tasks such as assembly and disassembly and proposed a discriminate model for multi-choice answer prediction. With the development of large vision-language models, we believe our work provides a valuable contribution to modern applications by effectively tackling the new challenge.

C Gaze Data Collection

In order to examine the gaze patterns of users, we conducted a preliminary study where participants engaged in one of three daily situations: grocery shopping, visiting a museum, or engaging in domestic activities. Participants were instructed to perform queries specific to each scenario as if the system could provide the desired information. Our research involved 21 participants (8, 5, and 8 for each respective scenario), consisting of 13 males and 8 females, ranging in age from 19 to 30 years old (with a standard deviation of 3.18). The study took each participant between 90 and 150 minutes to finish, and they were compensated at a rate of 15 USD per hour for their involvement. As a result, we get 548 minutes of gaze recording. The Pupil Labs Invisible [48] is a gaze-tracking smart glasses system that has been widely used for research purposes. It is equipped with gaze sensors, an egocentric camera, a microphone, and an inertial measurement unit (which was not used in this work). Participants were asked to wear the Pupil Labs Invisible glasses without any headwear that could obstruct the sensors on the glasses. Since the Invisible glasses require a connection to a mobile phone for operation, we instructed participants to keep the phone in their pockets to minimize potential distractions. Data was continuously recorded as participants engaged in their chosen scenario.

D Voila COCO diversity and quality

We conduct a comprehensive analysis of the VOILA-COCO dataset to evaluate the quality and diversity of the automatically generated QA pairs. Our examination included several aspects:

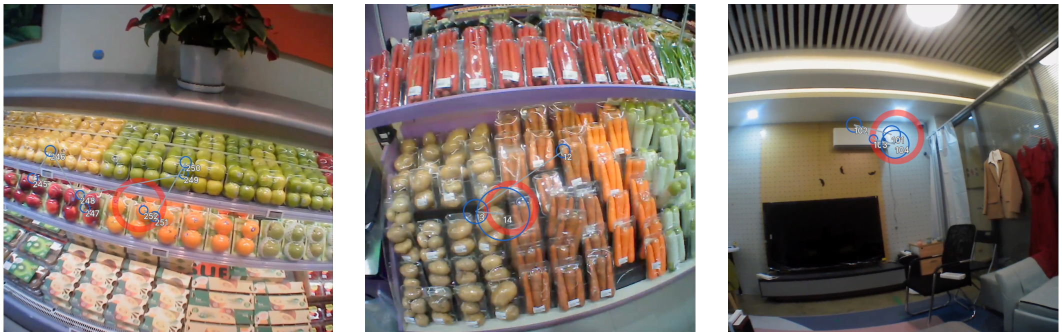


Ground Truth	Otter	Kosmos-2	Voila
Query: What colors are the cakes? **Answer**: The cakes on the table are in red, brown, and orange colors.	There are two cakes in the image. The first cake is white, and the second cake has a red color.	The cakes are decorated with red and black colors	The cakes on the table are in red, brown, and orange colors.
Question: Can I peel it with my hands? **Answer:** Yes, you can peel an orange with your hands...	I'm sorry, but I couldn't understand the answer. Can you please rephrase or provide more context?.	Yes, you can peel an orange with your hands	Yes, it is possible to peel the orange using your hands. However, it might be easier to use a knife or a peeler to remove the skin from the orange.
Question: What's the color? **Answer:** The color of the sky in the background is blue	The color of the plane is white.	The color of the planes in the image is blue	The background is blue in color.
Question: How to reduce eye fatigue? **Answer:** To reduce eye fatigue while using a laptop, take regular breaks, follow the 20-20-20 rule … adjust screen brightness and contrast, …	The most common cause of eye fatigue is overexposure to bright light …	Keyboard	To minimize eye fatigue while working on a laptop, consider the following steps: 1. Adjust the screen‘s brightness to a comfortable level …
Question: How many figures are there in this painting? **Answer:** According to what I see, the painting seems to contain 23 figures.	There are two figures in this image.	There are two figures in the painting .	There are two people standing in the painting.

Figure 8: Qualitative Case Study: Top: We show successful predictions of all models. Middle: We demonstrate the problems of baseline models compared with VOILA including coreference queries, gaze grounding methods, etc. Bottom: We display hard challenges for all models.

Supermarket shopping Task	**Domestic living** Task
Comparison	Appliance Malfunction
Completing Recipe	Activity & Health
Recommend	Snack & Fruits
Knowledge	Dressing Advice
Decision Making	Entertainment
Strengthen Decision	Small Talk

Table 5: Guiding for User in VOILA-GAZE Collection, note this guide aims to facilitate and inspire users to generate questions related to data collection, rather than imposing strict limitations on the scope of their inquiries.



Question: Please recommend some fruits for losing weight.
Answer: To support weight loss, consider incorporating fruits like apples, oranges, and lemons into your diet. Apples are high in fiber and low in calories, which can help you feel full longer. Oranges provide a good source of vitamin C and fiber, while lemons can aid digestion and boost metabolism. Remember to consume these fruits in moderation and as part of a balanced diet for optimal results.

Question: Which one is better for making dumplings?
Answer: Potato is a better option for making dumplings, as it provides a smoother and creamier texture when cooked, making it ideal for dumpling fillings. Carrots can be used as an additional ingredient to add color and flavor, but using potatoes as the main ingredient will result in a more satisfying and traditional dumpling experience.

Question: why is it not working?
Answer: It's possible that your air conditioner is not working due to a variety of reasons, such as a dirty filter, low refrigerant levels, a malfunctioning thermostat, or electrical issues. To determine the exact cause, it's best to consult your user manual or contact a professional HVAC technician for a proper diagnosis and repair.

Figure 9: Data sample from VOILA-GAZE

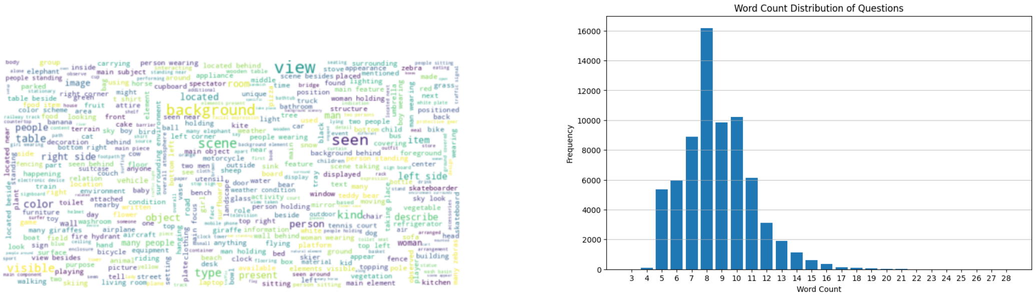
Vocabulary Richness: Utilizing Word Cloud visualizations, we have assessed the range of vocabulary employed within the dataset.

Sentence Length: We have analyzed the length of sentences for direct questions, indirect questions, and answers to ensure a realistic distribution that reflects typical conversational patterns.

Topic Distribution: By applying Latent Dirichlet Allocation (LDA) models, we have investigated the topical variety present within the dataset across the different categories of questions and answers.

D.1 Direct Questions

See Figure 10.



(a) Wordcloud

(b) Length Distribution

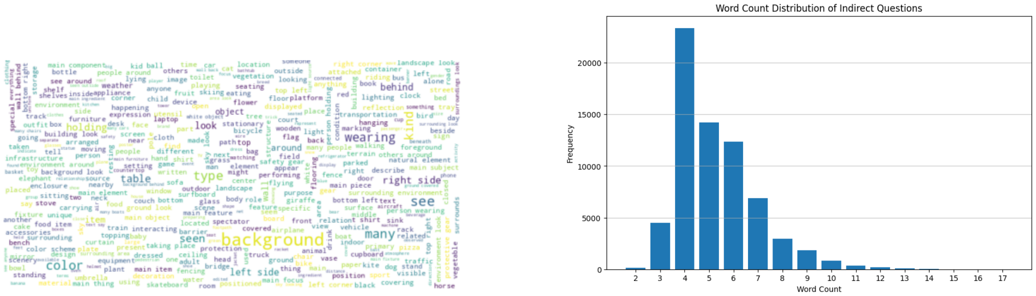
Figure 10: Visualizations of VOILA-COCO: direct questions

D.2 Indirect Questions

See Figure 11.

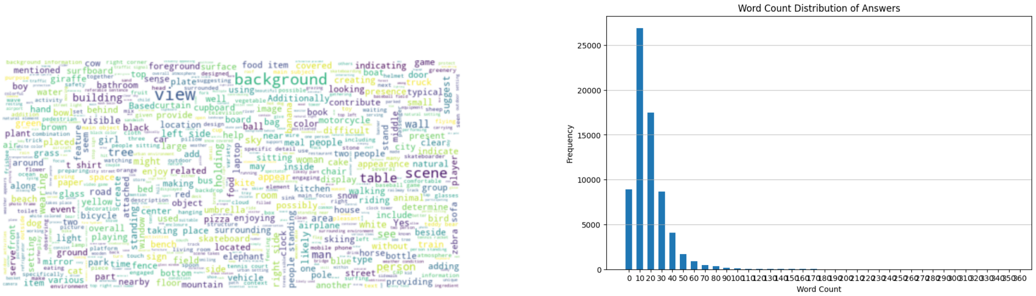
D.3 Answers

See Figure 12.

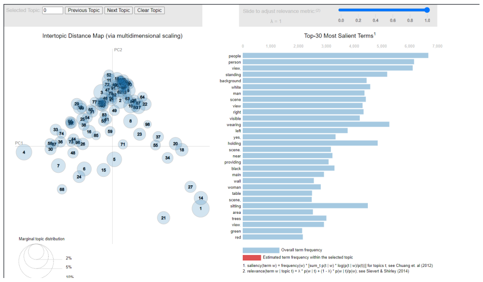


(a) Wordcloud (b) Length Distribution

Figure 11: Visualizations of VOILA-COCO: indirect questions



(a) Wordcloud (b) Length Distribution


Selected Topic:
Previous Topic
Next Topic
Clear Topic
Intertopic Distance Map (via multidimensional scaling)
PC1
PC2
Top-30 Most Salient Terms
Marginal topic distribution
Overall term frequency
Estimated term frequency within the selected topic

(c) LDA Topic Model Distribution

Figure 12: Visualizations of VOILA-COCO: answers

E Details for Automatic Data Annotation

Figure 13 shows the system prompt and in-context examples for Automatic Data Annotation

F Implementation Details

In this section, we describe the implementation details of our model. The model architecture is derived from Otter, which combines a text model and a vision model. The text model is an instance of MPTForCausalLM 7B and the vision model is based on the CLIP ViT-Large model with a patch size of 14.

- **Text Model Configuration:** The text model has 32 layers, each with 32 attention heads. The model has a hidden size and d_{model} of 4096. The attention mechanism uses multi-head attention with torch implementation and alibi attention. The model uses learned position embeddings and low-precision layer normalization. The model is trained with a maximum sequence length of 2048 and a vocabulary size of 50432. The tokenizer used is EleutherAI/gpt-neox-20b. The model's torch data type is set to bfloat16.

https://doi.org/10.52202/079017-0060

Prompt for Automatic Referring Annotation

System Prompt: As an AI visual assistant, your task is to analyze a single view as if you are directly observing it. You will be provided with background and referable sentences describing the view, but your responses should be based on your observations rather than the descriptions. First, annotate the referable sentence using the <Q#></Q#> tag, where # is the tag number, to emphasize the section you want to discuss. Note: DO NOT annotate background sentences. Then, engage in a conversation with a person inquiring about the view, focusing on the tagged elements while considering the background information. Use the <Q#> tag for questions related to specific tags in the sentence, and the <Q> tag for other questions. The conversation should resemble the tone of a visual AI assistant examining the view and answering questions. Include a variety of questions and their corresponding answers. Your questions should cover various aspects of the view's visual content, such as object types, object counts, actions, locations, and relative positions between objects. Ensure your questions have definite answers: (1) the view clearly displays the content in question, allowing for a confident response; or (2) it can be confidently determined from the view that the content is not present. Avoid questions that cannot be answered with certainty. Try your best to incorporate complex questions relevant to the view's content, including background information about the objects or discussions about events occurring within the view. Refrain from asking about uncertain details. When using <Q#>, follow the original question with an indirect question, as if the questioner is looking at the area of interest; use pronouns to refer to the relevant part, and omit some context in the original question if it can be inferred from the user's eye gaze fixation. When answering complex questions, provide detailed responses, including examples or reasoning steps, to create a convincing and well-organized explanation. Use multiple paragraphs if necessary. Make the answer comprehensive and informative. Throughout the conversation, maintain the impression that you are directly observing the view rather than reading descriptions. Do not mention any words like "image", "picture", or "text" in the question and answer, as if you are equipped in an HMD and sharing the same view with the user.

In context Example:

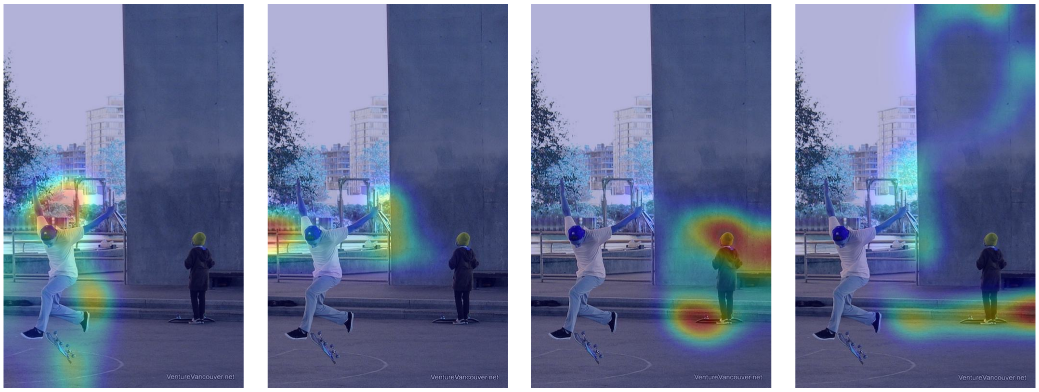
Role: user
Background: There is a movie theater that displays the show times above the doors.A hotel for dogs in the snow in winter.
Referable: A red fire hydrant is deep in the snow. The fire hydrant is in the snow near a recently plowed sidewalk. This city has had a very hard winter with snow.

Role: assistant
Refer:<Q1>A red fire hydrant</Q1> is <Q2>deep in the snow</Q2>.<Q3>The fire hydrant is in the snow near a recently plowed sidewalk.</Q3><Q4>This city has had a very hard winter with snow.</Q4>

===
Question:
<Q1>What is the red thing?
===
Indirect Question:
What is it?
===
Answer:
It is a red fire hydrant.
===
Question:
<Q2>Is the fire hydrant fully covered by snow or partially visible?
===
Indirect Question:
Is it fully covered by snow or partially visible?
===
Answer:
The fire hydrant is partially visible, with a significant portion of it covered by the deep snow.
===
Question:
<Q3>What is the condition of the sidewalk near the fire hydrant?
===
Indirect Question:
How's the condition of the sidewalk there?
===
Answer:
The sidewalk near the fire hydrant has been recently plowed, so it is relatively clear of snow compared to the surrounding area.
===
Question:
<Q4>What does this scene suggests about the winter season in the city?
===
Indirect Question:
How's the winter here?
===
Answer:
It suggests that the winter season in the city has been quite harsh and challenging. The significant snow accumulation seen in the view indicates that the city has experienced heavy snowfall, which can lead to various difficulties and disruptions for its residents. For instance, the deep snow covering the fire hydrant could potentially hinder its accessibility in case of emergencies, which poses safety concerns. The recently plowed sidewalk implies that the city's maintenance crews have been actively working to keep the walkways clear and safe for pedestrians, but the sheer amount of snowfall might make it difficult to keep up with the continuous snow removal efforts.
Furthermore, such extreme winter conditions can impact transportation, leading to traffic issues, delays in public transit, and increased risks of accidents due to slippery roads. It can also cause problems for businesses and schools, as people might face challenges commuting to work or attending classes.
In conclusion, the red fire hydrant deep in the snow and the recently plowed sidewalk suggest that the city has faced a particularly severe winter season, with substantial snowfall that has likely caused various challenges and disruptions for its residents and infrastructure."

Figure 13: Prompt for Automatic Referring Annotation

https://doi.org/10.52202/079017-0060

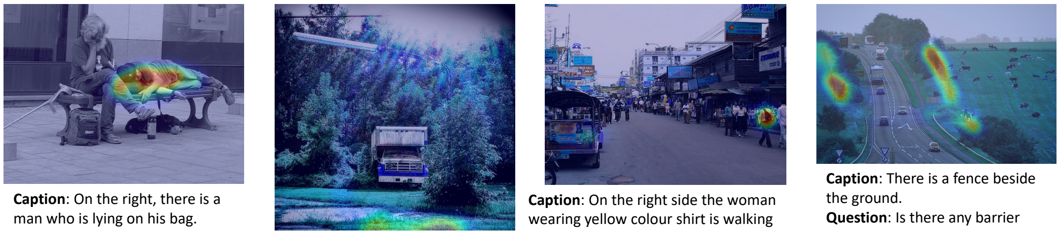


Caption: Bottom left side of the image a man is doing skateboarding and jumping.
Question: What is the skateboarder in the bottom left doing?
Indirect Question: What is he doing?
Answer: The skateboarder in the bottom left is performing a trick, jumping into the air with his skateboard.

Caption: Behind him there is a fencing. Behind the fencing there is water.
Question: What is behind the jumping skateboarder?
Indirect Question: What's behind him?
Answer: Behind the jumping skateboarder, there is a fencing , and behind the fencing, there is water.

Caption: Bottom right side od the image a person is standing and there is a skateboard.
Question: What can you see on the bottom right side of the view?
Indirect Question: What's there on the bottom right side?
Answer: On the bottom right side of the view, there is a person standing, and a skateboard is present near them.

Caption: Behind him there is a wall.
Question: What is behind the person standing on the bottom right?
Indirect Question: What's behind that person?
Answer: Behind the person standing on the bottom right, there is a wall.



Caption: On the right, there is a man who is lying on his bag.
Question: What is the man doing on the bench?
Indirect Question: What is he doing?
Answer: The man is lying on his bag, possibly taking a nap or resting.

Caption: At the bottom of the image I can see the grass in green color.
Question: What color is the grass?
Indirect Question: What color is it?
Answer: The grass is green in color.

Caption: On the right side the woman wearing yellow colour shirt is walking wearing a black colour bag along with her.
Question: Can you describe the woman walking on the right side?
Indirect Question: What does she look like?
Answer: The woman walking on the right side is wearing a yellow shirt and carrying a black bag with her. She seems to be going about her day, possibly shopping or commuting.

Caption: There is a fence beside the ground.
Question: Is there any barrier between the road and the cattle's grazing area?
Indirect Question: Is there any separation between them?
Answer: Yes, there is a fence beside the ground, providing a barrier between the road and the cattle's grazing area.

Figure 14: Annotated Example of VOILA-COCO

- **Vision Model Configuration:** The vision model has 24 hidden layers and a hidden size of 1024. It uses 16 attention heads and an intermediate size of 4096. The activation function is quick_gelu. The model uses an image size of 224 and 3 input channels. The patch size is 14, and the projection dimension is 512. The layer normalization epsilon is set to 1×10^{-5}.
- **Voila Configuration:** The architecture uses cross attention every 4 layers and only attends to previous layers. The model's torch data type is set to float32. Media placement augmentation is enabled during training.

For initialization, we use the Kaiming normal method with fan-in mode, ReLU nonlinearity, and a standard deviation of 0.02.

For optimization, we employ the AdamW optimizer [22] with a starting learning rate of 1e-5 and a batch size of 4. We train Voila for three epochs, scheduling the learning rate using a cosine annealing scheduler. To prevent exploding gradients, we apply gradient clipping with a threshold of 1.0.

G GPT-RANK

Figure 15 presents our prompt and evaluation procedure for GPT-RANKING

GPT Ranking

Overall System Prompt: Given a question along with the ground truth description and answer of an image, evaluate the two provided candidate answers. Determine which answer is factually accurate, logical, and helpful to the user. if you think answer 1 is better, respond with -1, if answer 2 is better respond with 1, if you think the result is tie, output 0. Only respond with either '-1' or '0' or '1' to indicate your choice.

Helpfulness System Prompt: Given a question along with the ground truth description and answer of an image, evaluate the two provided candidate answers. Determine which answer is solving the user's problem and is more helpful to the user. if you think answer 1 is better, respond with -1, if answer 2 is better respond with 1, if you think the result is tie, output 0. Only respond with either '-1' or '0' or '1' to indicate your choice.

Fact Grounding System Prompt: Given a question along with the ground truth description and answer of an image, evaluate the two provided candidate answers. Determine which answer is factually grounded to the Fact provided. if you think answer 1 is better, respond with -1, if answer 2 is better respond with 1, if you think the result is tie, output 0. Only respond with either '-1' or '0' or '1' to indicate your choice.

```
for all key in keys do
    answer1 ← model1[key]["response"]
    answer2 ← model2[key]["response"]
    fact, gt_answer, question ← dataset[key]
    prompt1 ← CreatePrompt(question, fact, gt_answer, answer1, answer2)
    response_1 ← CallGPT(prompt1)
    prompt2 ← CreatePrompt(question, fact, gt_answer, answer2, answer1)
    response_2 ← CallGPT(prompt2)
    score ← ComputeScore(response_1, -1 * response_2)
    Append scores with score
end for
```

Figure 15: GPT-RANKING Procedure

H Ablations on how to incorporate gaze

Figure 16 shows different approaches exist for incorporating gaze data.

I Generalizability of VOILA

We conduct an evaluation using the VizWiz dataset to validate the generalizability of our model. Given that VizWiz is a widely recognized benchmark in the Visual Question Answering (VQA) domain, we were able to directly apply our model in this context. However, it is important to note that the standard VizWiz VQA dataset does not provide positional information, which is critical for evaluating the distinction between conventional VLMs and those enhanced with gaze input. To address this gap, we evaluated our model on the validation split of the VizWiz grounding dataset, employing a metric analogous to that used in VizWiz VQA but incorporating grounding points as a proxy for gaze input. Our model achieved an accuracy of 49.6, outperforming the baseline model, which scored 33.2. We believe these zero-shot results convincingly demonstrate both the effectiveness and generalizability of our method.

J Impact Statements

Our experiments have been reviewed and approved by the Institutional Review Board (IRB) at our institute. All participants of the VOILA-GAZE collection process have provided informed consent, agreeing that their data will be used as part of a publicly available dataset. To ensure privacy, all personal information has been properly anonymized. Additionally, we will conduct a thorough review of any potential privacy issues before releasing the dataset to the public. We are committed to upholding ethical standards in research and protecting the privacy of our study participants.

Voila-A can be implemented in HMD AR/VR devices as an egoview copilot, benefiting a wide range of users, including visually impaired individuals who rely on their gaze to communicate their intent.

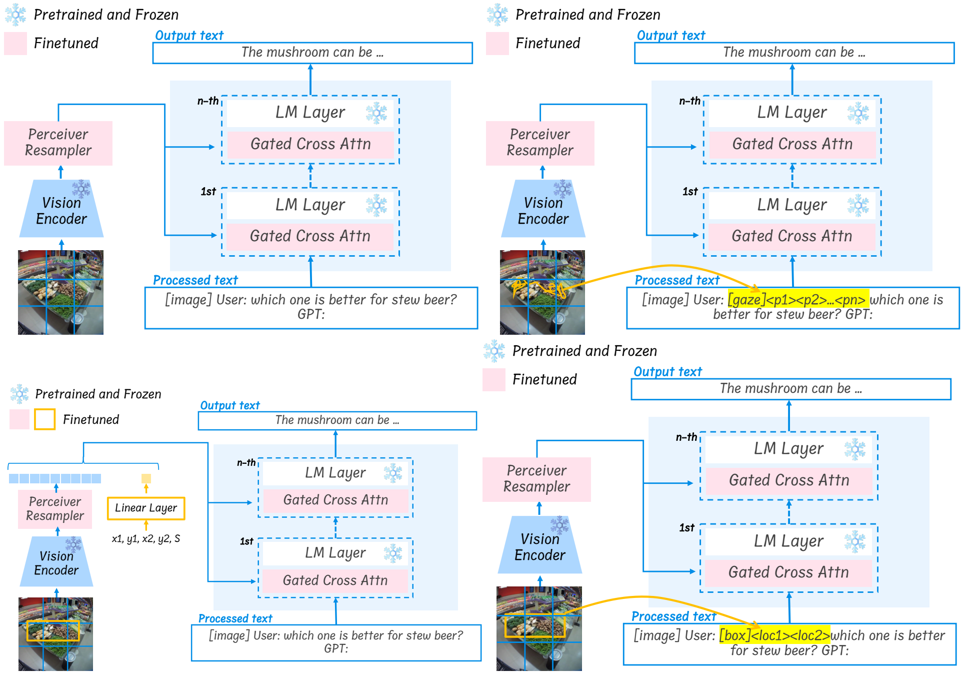


Figure 16: Different model design on how to incorporate gaze.Top-left: Otter-base, Top-right: Gaze as discrete position tokens. Bottom-left: Gaze *bounding box* as image patch Bottom-right:Gaze *bounding box* as discrete position tokens

This method surpasses the capabilities of similar mobile apps that necessitate users to lift their phones for scene capture.

K Reproducibility Statement

In the interest of promoting transparency and facilitating further research within the community, we are committed to providing comprehensive resources alongside the publication of our work. To this end, we will release the VOILA-COCO and VOILA-GAZE datasets, as well as the annotation pipeline, training code, and model weights. By making these materials publicly available, we aim to ensure that our methods can be easily replicated and built upon by other researchers. Our commitment to reproducibility aligns with the broader scientific goal of fostering collaboration and the development of novel ideas and techniques in the field.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Each contribution is discussed in our paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please find in sec A

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

https://doi.org/10.52202/079017-0060

Justification: This paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Please see sec K

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

https://doi.org/10.52202/079017-0060

Answer: [Yes]

Justification: Please see sec K

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Please see sec F

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: Error bar is not applicable to our metric.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

https://doi.org/10.52202/079017-0060

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please see sec F

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We fully adhere the NeurIPS Code Of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We disscussed in section J.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate

https://doi.org/10.52202/079017-0060

deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [Yes]

Justification: Please find in sec K

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All assets are properly cited and licensed.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Please find in our method section.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: Please find in sec C

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [Yes]

Justification: Our experiments have been reviewed and approved by the Institutional Review Board (IRB) at our institute. All participants have provided informed consent, agreeing that their data will be used as part of a publicly available dataset. To ensure privacy, all personal information has been properly anonymized. Additionally, we will conduct a thorough review of any potential privacy issues before releasing the dataset to the public. We are commit to upholding ethical standards in research and protecting the privacy of our study participants.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-0060