# CoreGaze: Core Subgraph-Driven Visual Gaze Diffusion for Training-Free Referring Multimodal Large Language Models

**Xiaoyang Yi[1,3,4], Jing Chen[2,3,4], Yuru Bao[1,3,4], Jian Zhang[1,2,3,4]***
[1]College of Cryptology and Cyber Science, Nankai University
[2]College of Computer Science, Nankai University
[3]Tianjin Key Laboratory of Network and Data Security Technology
[4]Key Laboratory of Data and Intelligent System Security, Ministry of Education
***Correspondence:** zhang.jian@nankai.edu.cn

## Abstract

Referring multimodal large language models enable users to ground queries to specific image regions via spatial prompts, supporting fine-grained referring dialogue. However, existing methods rely on extensive fine-tuning to mitigate attention distraction, which incurs high computational costs and limits adaptability. Without sufficient training data, irrelevant regions in single images easily divert model focus, leading to redundant outputs or hallucinations. To address this, we propose CoreGaze, a training-free framework that simulates human visual gaze diffusion for fine-grained comprehension. First, CoreGaze constructs a sparse semantic graph from visual tokens, modeling region-wise affinities via thresholded similarity. It then maps the user's visual prompt to a core subgraph with amplified initial influence, which drives a degree-normalized diffusion process using restart-equipped random walks to propagate relevance to contextual neighborhoods. This process prunes irrelevant tokens while preserving user-indicated targets and semantically linked context, distilling a focused yet comprehensive subgraph. Finally, CoreGaze fuses this subgraph with prompt tokens in the frozen large language model decoder, facilitating fine-grained referring generation. Experimental results show that CoreGaze achieves outstanding performance in multiple referring dialogue tasks, showcasing its effectiveness.

## 1 Introduction

In recent years, as visual pretraining and large language models (LLMs) matured, multimodal large language models (MLLMs) have become a major focus of both research and practical applications (Fang et al., 2024; Zhang et al., 2025; Lin et al., 2025c). These models split an image into several patches and feed them into a visual encoder (such as ViT (Dosovitskiy et al., 2021)) to obtain visual tokens, then combine those tokens with text tokens in a frozen LLM decoder, achieving deep fusion of image content and linguistic meaning (Ye et al., 2025; Zeng et al., 2025). MLLMs such as LLaVA (Liu et al., 2023) and miniGPT-4 (Zhu et al., 2024) have simplified visual interaction by generating coherent responses based on images and user instructions, demonstrating impressive conversational abilities.

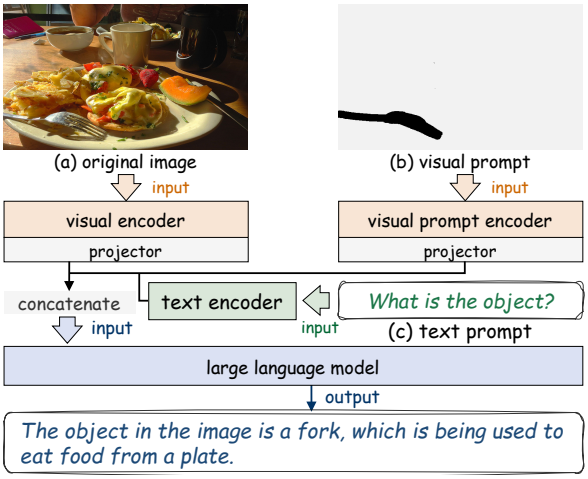



Figure 1: The framework of a referring MLLM, with user inputs of an original image, a visual prompt and a text prompt, requiring the model to generate a fine-grained description of the specified region.

While these MLLMs perform well at global image understanding, they are less effective when users seek to explore specific regions in detail through interactive dialogue. Fine-grained comprehension is critical for advanced vision tasks, and it cannot be achieved simply by asking users to provide more detailed textual descriptions of particular areas (Guo et al., 2024; He et al., 2024). Consequently, there is growing interest in referring dialogue for MLLMs, where a user indicates a region of the image by box, mask, scribble or point prompts, and then receives a precise response focused on that area (You et al., 2024; Zhang et al.,



*Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1297–1315
July 2-7, 2026 

2024b). To address this need, referring MLLMs have emerged (Tian et al., 2024; Zhao et al., 2024; Lin et al., 2025a). As shown in Figure 1, these models need to understand the user's textual prompt while integrating spatial information so that attention is accurately directed at the indicated region.

Some referring MLLMs convert user-provided spatial cues into position tokens or numeric text descriptions (Ma et al., 2024; Yue et al., 2024), but these approaches cannot support free-form region selection. Others rely on extensive training to achieve fine-grained visual-language alignment (Cai et al., 2024; Lin et al., 2025b), requiring fine-tuning on hundreds of thousands of image-text pairs with localization annotations, which incurs high computational cost and struggles to adapt to new data (Wu et al., 2024). However, without sufficient training data, the model cannot quickly focus on the designated area in a single image, and unrelated regions may distract its attention, leading to redundant or hallucinated outputs.

The visual encoding process within an MLLM, where images are first processed by a pre-trained visual encoder, presents new possibilities for addressing the aforementioned attention distraction without training. The resulting visual tokens inherently possess richer representational information than the raw image (Pan et al., 2021), which enables the removal of tokens from irrelevant regions to concentrate the model's attention. Yet in referring dialogue tasks requiring precise understanding, randomly removing too many tokens leads to a sharp performance decline, indicating substantial loss of critical information. Moreover, retaining only tokens corresponding to the user-specified region discards important context from surrounding foreground and background regions that support accurate interpretation (Jiang et al., 2025). Therefore, it becomes a challenge that removes redundant tokens while preserving both the user-indicated region and its relevant context.

Motivated by these observations, we introduce CoreGaze, a training-free referring MLLM framework simulating **Core** subgraph-guided visual **Gaze** diffusion. Specifically, CoreGaze decomposes the image into visual tokens that serve as nodes in a graph, with edges weighted by similarity to model semantic relationships among regions. It then anchors user-specified regions as a core subgraph and initiates controlled gaze diffusion to propagate relevance to contextual neighborhoods. This biologically inspired process dynamically prunes irrelevant tokens while preserving critical cues, distilling a focused yet comprehensive visual subgraph. By leveraging degree-normalized attention scores to eliminate connectivity bias, CoreGaze distills a concise yet expressive visual subgraph. This refined input, which retains both user-focused cues and diffusion-activated context, is seamlessly fused with prompt tokens in the frozen LLM decoder, facilitating fine-grained referring generation.

To summarize, our contributions are as follows:

- We introduce CoreGaze, a training-free referring MLLM framework that provides refined and focused visual inputs for language generation through core divergence, supporting free-form user selection of image regions.
- We simulate human visual focus by constructing a sparse token graph to capture semantic associations among regions, and convert the user's selection into a core subgraph as an accurate starting point for diffusion.
- We emulate the human gaze diffusion process to spread relevance from the core subgraph to related neighborhoods, capturing contextual cues while pruning irrelevant tokens to eliminate noise.
- Experimental results show that CoreGaze achieves outstanding performance in multiple referring dialogue tasks, showcasing both expressiveness and effectiveness.

## 2 Related Work

Referring dialogue allows users to point out regions in an image and enabling the model to locate them precisely. MLLMs that support referring dialogue typically let users mark areas with boxes, points or other cues and then align those regions with text. For example, KOSMOS-2 (Peng et al., 2024) turns continuous coordinates into positional tokens embedded as hyperlinks, ensuring exact text-to-region alignment, while GPT4-RoI (Zhang et al., 2024c) replaces users' bounding box with multi-level feature-pyramid RoI descriptors woven into the language embedding sequence. Shikra (Chen et al., 2023) handles both point and box inputs in a unified framework by expressing all coordinates as natural-language numbers, which removes the need for extra vocabulary or modules. Additionally, Ferret (You et al., 2024) merges discretized

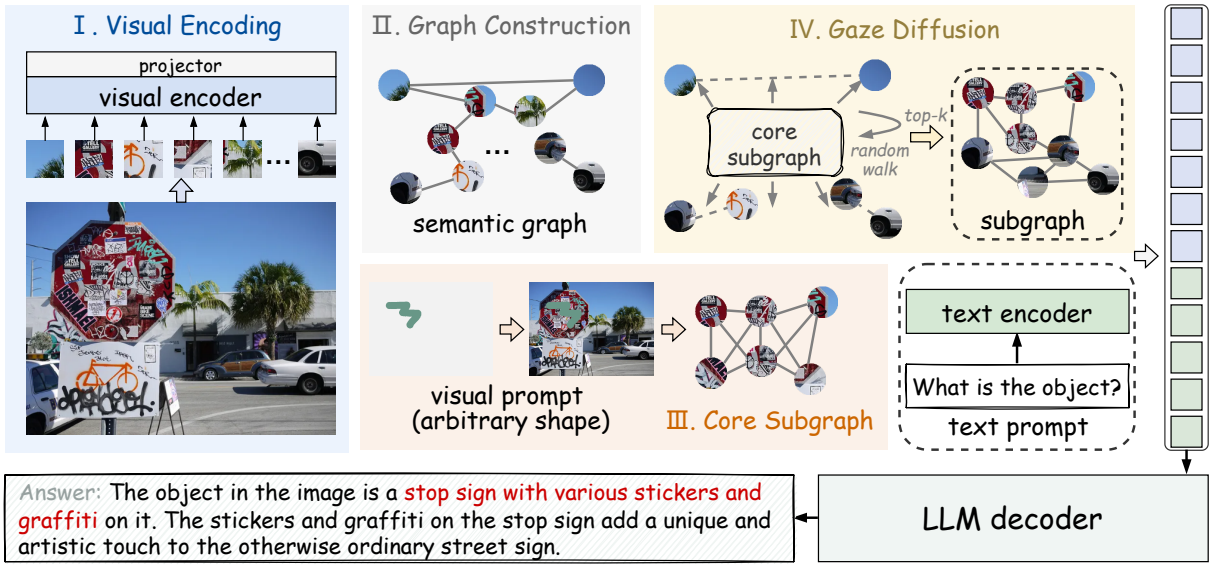



Figure 2: The overall framework of CoreGaze. It first constructs a sparse semantic graph from image tokens, and the user's region prompt is then mapped to a core subgraph, initiating a gaze diffusion process that dynamically prunes irrelevant tokens. This refined input is fused with prompt tokens to the frozen LLM decoder, facilitating fine-grained referring generation.

coordinates with continuous features and samples hierarchically to handle arbitrarily shaped areas. Alpha-CLIP (Sun et al., 2024) adds a learnable alpha channel to CLIP so users can specify any focus area without altering the image. RegionGPT (Guo et al., 2024) encodes arbitrary regions into semantic embeddings, aligning them with an LLM for detailed vision-language fusion. PixelLLM (Xu et al., 2024) regresses 2D coordinates through a lightweight MLP head, creating a dynamic semantic trajectory that fuses dual vision backbones with positional prompts. LLaVA-Grounding (Zhang et al., 2024a) bridges Vicuna with a localization model through a three-stage training strategy to support pixel-level grounding. ControlMLLM (Wu et al., 2024) injects visual cues into MLLMs by optimizing learnable latent variables, guiding them focus on the visual content of specified regions.

## 3 Method

The framework of CoreGaze is illustrated in Figure 2. It first constructs a sparse semantic graph from image tokens, where nodes represent visual patches and edges encode region-wise affinities. The user's region prompt is then mapped to a core subgraph, initiating a gaze diffusion process that propagates relevance to contextual neighborhoods while dynamically pruning irrelevant tokens. By leveraging degree-normalized attention scores to eliminate connectivity bias, we distill a concise yet expressive visual subgraph. This refined input, which retains both user-focused cues and diffusion-activated context, is seamlessly fused with prompt tokens to the frozen LLM decoder, facilitating fine-grained referring generation.

### 3.1 Visual Encoding

The MLLM is composed of a visual encoder, an MLP, and an LLM decoder. Given an input image $I$, it is divided into $N$ visual patches, which are then passed through the visual encoder $\mathrm{Enc}_v(\cdot)$ to extract features, yielding a set of raw feature vectors. These feature vectors are then fed into a projection MLP layer to obtain the final sequence of visual token vectors as follows:

$$\mathrm{x}_i = \mathrm{MLP}(\mathrm{Enc}_v(I)_i) \in \mathbb{R}^{N \times d} \quad (1)$$

$$\mathrm{e}_v = [\mathrm{x}_1, \mathrm{x}_2, \ldots, \mathrm{x}_N] \quad (2)$$

where $d$ denotes the dimension of each feature vector, and each feature vector $\mathrm{x}_i$ captures the high-level semantic information of the $i$-th region of the image.

At the same time, the text prompt $p_t = (w_1, w_2, \ldots, w_M)$ for the input image is tokenized and mapped into a continuous embedding space, then passed through a text encoder $\mathrm{Enc}_t(\cdot)$ to obtain text tokens $\mathrm{e}_t$ that typically shares its initial layers with the LLM decoder to maintain consistency

between the visual and linguistic representations:

$$\mathrm{t}_j = \mathrm{Enc}_t(w_j) \in \mathbb{R}^{M \times d} \tag{3}$$

$$\mathrm{e}_t = [\mathrm{t}_1, \mathrm{t}_2, \ldots, \mathrm{t}_M] \tag{4}$$

where $M$ denotes the total number of tokens $t$ in the text prompt after tokenization, and $w_j$ refers to the $j$-th word in that sequence.

The core mechanism of MLLMs lies in integrating visual tokens and text tokens into a unified Transformer decoder, forming the input to the frozen LLM. Following approaches like LLaVA (Liu et al., 2023), MLLMs typically perform cross-modal attention computations within each layer $\mathrm{Att}^{(l)}$ of the standard Transformer architecture:

$$\mathrm{Att}^{(l)} = \mathrm{softmax}(\frac{[\mathrm{e}_v, \mathrm{e}_t]^{(l)} \cdot ([\mathrm{e}_v, \mathrm{e}_t]^{(l)})^\top}{\sqrt{d_k}}) \tag{5}$$

where $\sqrt{d_k}$ is a scaling factor. Specifically, visual tokens and text tokens are treated as query, key, and value, respectively, enabling the progressive integration of multimodal information across successive layers.

### 3.2 Graph Construction

To simulate human visual focus, we construct the image as a sparse graph to model semantic relationships between regions. This occurs after several image patches are processed by the visual encoder and MLP, yielding $N$ visual token representations as described above.

Our objective is to measure the semantic relevance between different regions from a graph-theoretic perspective, thereby selecting the most representative tokens that encompass the key information required for the user-specified region, whether foreground or background. Consequently, we treat these $N$ visual tokens as graph nodes and employ cosine similarity to quantify the semantic proximity between any two tokens $(\mathrm{e}_{v_i}, \mathrm{e}_{v_j})$. This similarity value serves as the edge weight, allowing us to construct a weighted adjacency matrix based on feature affinity as follows:

$$\mathrm{Sim}_{ij} = \frac{\mathrm{e}_{v_i} \mathrm{e}_{v_j}}{\| \mathrm{e}_{v_i} \|_2 \| \mathrm{e}_{v_j} \|_2} \tag{6}$$

Since each token $\mathbf{e}_{v_i}$ encodes global context, the cosine similarity inherently reflects both local feature matching and global semantic alignment.

To eliminate spurious connections arising from noise or weak correlations, we apply a preset threshold $s$, keeping a weighted edge between nodes $\mathrm{e}_i$ and $\mathrm{e}_j$ only if $\mathrm{Sim}_{ij} \geq s$, otherwise setting it to 0, thereby constructing a sparse adjacency matrix $A$:

$$A_{ij} = \begin{cases} \mathrm{Sim}_{ij}, & \mathrm{Sim}_{ij} \geq s \\ 0, & \mathrm{Sim}_{ij} < s \end{cases} \tag{7}$$

This ensures both computational tractability of the graph and guarantees that all retained edges represent strong semantic relationships.

Moreover, to make each node's contribution comparable, we employ a row-wise softmax to the sparsified adjacency matrix $A$, obtaining the normalized adjacency matrix $P$:

$$P_{ij} = \frac{\exp(A_{ij})}{\sum_{k=1}^{N} \exp(A_{ik})} \tag{8}$$

Here, row-wise normalization ensures that each node's outgoing edge weights sum to 1, thereby establishing a valid Markov transition framework.

### 3.3 Core Subgraph

In a referring MLLM, the model incorporates visual region prompts so that text generation concentrates on the user specified area. A common practice is to translate the region prompt $r$ into extra region tokens or positional encodings:

$$\mathrm{e}_r = \mathrm{Enc}_v(I, r) \in \mathbb{R}^{K \times d} \tag{9}$$

where $K$ denotes the number of region tokens extracted to represent the user-specified area.

In a training-free method (Wu et al., 2024), the referring MLLM update visual prompt tokens through attention loss training as follows:

$$\mathcal{L} = \sum_{i=1}^{O} (1 - \frac{\sum_{j=1}^{N} \mathrm{Att}_j \cdot \mathcal{G}(M_{i,j}, \mu, \sigma)}{\sum_{j=1}^{N} \mathrm{Att}_j})^2 \tag{10}$$

where $M_i \in \mathbb{R}^{1 \times N}$ is the mask of token $i$, $\mathcal{G}(\cdot, \mu, \sigma)$ is the Gaussian smoothing function. Then, the model concatenates updated tokens with the image tokens and the text tokens, allowing each attention layer to reinforce the model's focus on the region prompt, thereby conditioning the generation of the target sequence on the region prompt.

Based on these, we propose a core subgraph diffusion mechanism. Specifically, before concatenation, we treat the arbitrarily shaped user annotation $r$ as the focal point of attention. Utilizing the spatial mapping between that region and the visual tokens, we extract a set $\mathcal{C}$ of seed node indices for the core subgraph:

$$\mathcal{C} = \{\mathrm{e}_i | \text{if } i \text{ in } r\} \tag{11}$$

These seed nodes contain the most direct user attention information.

Next, we assign an initial score to each node in the entire graph. Recent research (Kang et al., 2025) has discovered a phenomenon known as visual attention sink in MLLMs. These sink tokens refer to visual tokens that exhibit abnormally high activation in specific hidden state dimensions, persistently attracting the model's attention even when the image regions corresponding to these tokens are irrelevant to the current text content. Therefore, we introduce an importance scoring mechanism based on sink dimension activation. For each visual token $\mathbf{e}_{v_i}$, we compute its activation strength over the predefined sink dimensions $\mathcal{D}_{\text{sink}}$ as follows:

$$\phi(\mathbf{e}_{v_i}) = \max_{d \in \mathcal{D}_{\text{sink}}} \left| \frac{\mathbf{e}_{v_i}[d]}{\text{RMS}(\mathbf{e}_{v_i})} \right| \tag{12}$$

where $\text{RMS}(\mathbf{e}_{v_i}) = \sqrt{\frac{1}{D}\sum_{d=1}^{D}(\mathbf{e}_{v_i}[d])^2}$ is the root mean square value, used for normalization. The set of sink dimensions $\mathcal{D}_{\text{sink}}$ depends on the base language model, and these dimensions develop a tendency to concentrate attention on uninformative tokens during pre-training.

Based on the finding that sink tokens contribute limitedly to the model's output, we assign different initial values $S^{(0)}$ to seed and non-seed nodes as follows:

$$S_i^{(0)} = \begin{cases} 1, & i \in \mathcal{C} \\ \exp(-\phi(\mathbf{e}_{v_i})), & i \notin \mathcal{C} \end{cases} \tag{13}$$

Here, a higher score indicates greater importance of the node. Consequently, all nodes in the core subgraph are assigned the globally maximum initial score, ensuring that they play a dominant role in the subsequent gaze diffusion process. Meanwhile, other nodes retain their inherent information metric and participate in scattering and competition.

### 3.4 Gaze Diffusion

Although the user requires the model to focus on specified regions, the lack of relevant contextual information still makes it difficult to accurately capture essential cues from the surrounding semantic environment. Therefore, simulating the natural diffusion of human attention from the focal region to its semantic neighborhood, we propagate the initial influence of the core subgraph along the graph structure to capture closely relevant context associated with the user's region of interest.

Specifically, using the row-normalized adjacency matrix as a foundation, we view gaze diffusion as a random walk with a restart mechanism. From any node $\mathrm{e}_i$, at each step we return to the core subgraph's initial distribution $\pi^{(0)}$ with probability $\alpha$, and with probability $1-\alpha$ continue walking to neighboring nodes according to $P$. The score vector $S^{(t)} \in \mathbb{R}^N$ is updated iteratively as follows:

$$S^{(t+1)} - (1-\alpha)P^\top S^{(t)} + \alpha\pi^{(0)},\ S^{(0)} = \pi^{(0)} \tag{14}$$

Here, $(1-\alpha)P^\top S^{(t)}$ describes the current attention diffusing outward according to the transition probabilities, while $\alpha\pi^{(0)}$ continually replenishes the core subgraph's energy, together ensuring that the focal energy neither dissipates too quickly nor fails to explore more distant related nodes.

After several iterations until convergence, we obtain the steady-state scores $\mathrm{s}^*$. And we introduce node degree $D_i$ for posterior normalization:

$$D_i = \sum_{j=1}^{N} A_{ij} \tag{15}$$

$$\widehat{S}_i = \frac{s_i^*}{D_i} \tag{16}$$

This ensures node scores better reflect their genuine connection strength to the core subgraph, rather than relying solely on high connectivity.

Consequently, we can prune irrelevant or redundant visual tokens based on node scores, thereby enhancing downstream generation efficiency and focus. This constitutes a straightforward filtering process, where we sort tokens by their normalized scores $\{\widehat{S}_i\}_{i=1}^N$ in descending order and select the top $k$ tokens to form the final index set as follows:

$$\mathcal{K} = \text{Top}(\{\widehat{S}_i\}_{i=1}^N, k) \tag{17}$$

Thus, the pruned visual token subset is formed:

$$\mathrm{e}_v^{'} = \{\mathrm{x}_i | i \in \mathcal{K}\} \tag{18}$$

This subgraph incorporates both representative tokens from the prompt region and globally contextual tokens activated through propagation, achieving a focused yet comprehensive representation. Finally, we concatenate the refined visual subgraph with updated visual prompt tokens and text prompt tokens, inputting them into the frozen LLM decoder. This enables the model to rapidly focus on user-specified regions, facilitating fine-grained interactive generation.

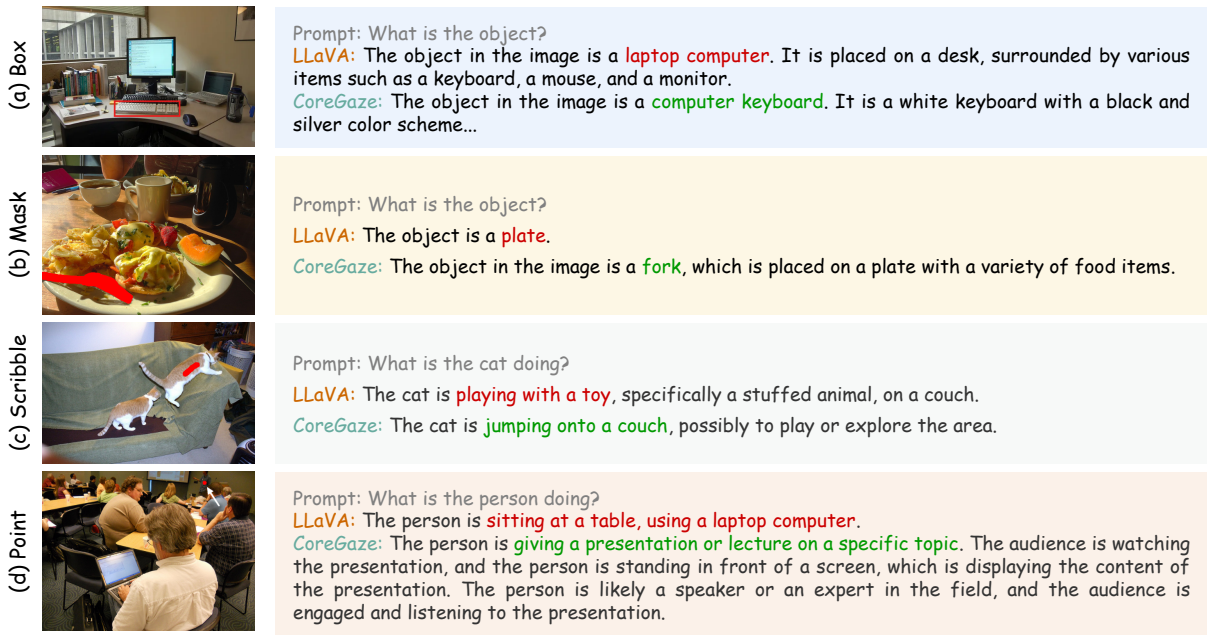



Figure 3: The examples of the referring dialogue, including (a) box, (b) mask, (c) scribble, and (d) point (the latter indicated by white arrows). Correct descriptions are marked in green while incorrect descriptions are marked in red.

# 4 Experiments

## 4.1 Experiment Setup

To demonstrate the effectiveness of CoreGaze, we present comparative responses from CoreGaze using LLaVA-1.5-7B (Liu et al., 2023) as its backbone model and baseline LLaVA in the referring dialogue on LVIS (Gupta et al., 2019) and COCO-Text (Veit et al., 2016). As shown in Figure 3, both models handle four user annotation formats: box, mask, scribble, and point. CoreGaze accurately addresses user queries about specified regions, whereas LLaVA provides irrelevant responses since it processes all visual tokens uniformly, allowing dominant but off-target regions to hijack the attention. CoreGaze effectively mitigates this issue through its core subgraph-driven gaze diffusion.

Additionally, we evaluate CoreGaze on referring object classification (ROC) and referring text classification (RTC) tasks, comparing it against various methods, the original LLaVA, and three LLaVA variants: LLaVA-Blur that blurs backgrounds outside referring regions, LLaVA-Color that highlights referring regions, and LLaVA-Edit Att which modifies attention maps to focus on referring regions. We set the similarity threshold $s = 0.3$, the pruning ratio $k = 0.4$, the restart probability $\alpha = 0.15$, and the iteration $t = 5$ in CoreGaze. All experiments are executed on a Tesla A100 GPU. *Extended experiments are detailed in the Appendixes A - H.*

## 4.2 Main Results

We conduct ROC and RTC on LVIS and COCO-Text respectively, where ROC identifies the object class in the user-specified region, while RTC recognizes text content in the user-specified region and performs binary classification, as shown in Table 1. For training-based methods, they require large amounts of annotated data for training, making it difficult for some methods to support visual prompts with arbitrary shapes (e.g., mask and scribble), while also incurring significant computational overhead. In contrast, CoreGaze surpasses their performance on most tasks without requiring additional fine-tuning, greatly reducing computational overhead. Furthermore, CoreGaze outperforms all other methods, demonstrating its effectiveness.

Simultaneously, we compare LLaVA and its variants with CoreGaze on optical character recognition (OCR) on COCO-Text, as shown in Figure 4. Neither LLaVA nor its variants could accurately recognize the user's referring region, as they struggle to focus on local information. CoreGaze effectively achieves this by leveraging the core subgraph, enabling precise referring dialogue.

## 4.3 Ablation Study

We conduct an ablation study on CoreGaze, evaluating the impact of each component, as shown in Table 2. When semantic graphs undergo random diffusion to form pruned subgraphs, accuracy rises to

| | | ROC | | | | RTC | |
|---|---|---|---|---|---|---|---|
| | | Box | Mask | Scribble | Point | Box | Mask |
| Training Methods | KOSMOS-2 (Peng et al., 2024) | 55.17 | - | - | - | 16.55 | - |
| | GPT4-RoI (Zhang et al., 2024c) | 58.59 | - | - | - | 54.23 | - |
| | Shikra-7B (Chen et al., 2023) | 64.60 | - | - | 56.27 | 50.07 | - |
| | Ferret-7B (You et al., 2024) | 71.71 | 72.39 | 71.58 | 68.54 | 55.47 | 56.34 |
| Training-free Methods | LLaVA-7B (Liu et al., 2023) | 54.13 | 54.13 | 54.13 | 54.13 | 55.03 | 55.03 |
| | LLaVA-Color (Liu et al., 2023) | 55.10 | 56.72 | - | - | 56.34 | 54.23 |
| | LLaVA-Edit Att (Liu et al., 2023) | 36.24 | 37.08 | - | - | 26.09 | 29.16 |
| | FitPrune (Ye et al., 2025) | 57.32 | 57.32 | 57.32 | 57.32 | 56.25 | 56.25 |
| | ControlMLLM (Wu et al., 2024) | 60.59 | 60.79 | 58.33 | 58.85 | 61.22 | 60.28 |
| | CoreGaze (this paper) | **71.22** | **72.22** | **74.14** | **73.04** | **74.23** | **71.43** |

Table 1: Results on ROC and RTC, the best results in training-free methods are in bold.


Figure 4: Results on OCR with different visual prompts.

| Diffusion | Scoring | Core Subgraph | Acc. |
|---|---|---|---|
| ✓ | | | 63.27 |
| ✓ | ✓ | | 67.86 |
| ✓ | ✓ | ✓ | **74.23** |

Table 2: Ablation results on RTC (box).

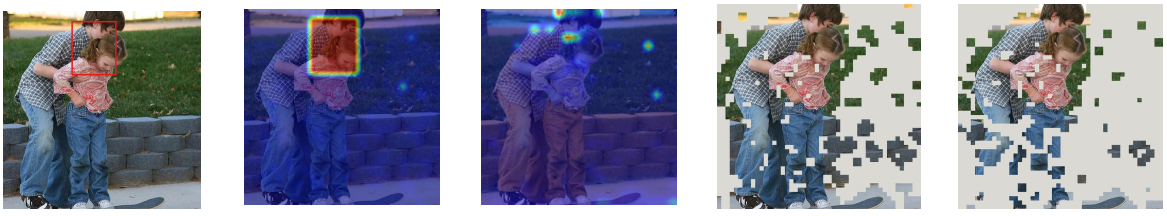
63.27%. This demonstrates that random walk diffusion filters out irrelevant nodes, enhancing regional relevance. Further integrating importance scoring mechanism boosts performance, confirming that constraints effectively mitigate over-diffusion and concentrate the process on structures semantically aligned with the referring region. Finally, after introducing the core subgraph, the substantial overall accuracy improves, showing that propagating importance from the core subgraph effectively preserves critical information while strengthening the model's ability to capture target regions and their contextual surroundings.

| | | Tokens | Speed (s) | GPU |
|---|---|---|---|---|
| Image 1 | LLaVA | 9 | 0.7117 | 8G |
| | Control. | 9 | 1.5524 | 12G |
| | CoreGaze | 9 | 1.4743 | 10G |
| Image 2 | LLaVA | 68 | 5.6641 | 8G |
| | Control. | 70 | 8.6471 | 12G |
| | CoreGaze | 71 | 7.0872 | 10G |
| Image 3 | LLaVA | 291 | 27.3864 | 8G |
| | Control. | 307 | 33.2173 | 12G |
| | CoreGaze | 300 | 29.7048 | 10G |

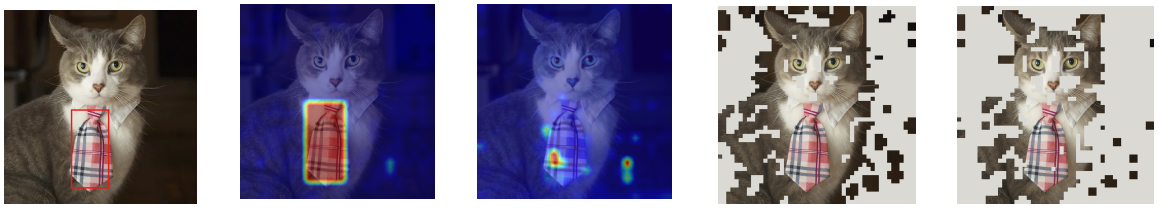
Table 3: The inference cost of LLaVA, ControlMLLM, CoreGaze with different images on a Tesla A100 GPU.

### 4.4 Inference Cost

We select three images of varying complexity on COCO-Text to compare the inference costs between LLaVA-7B and CoreGaze, as shown in Table 3. The number of tokens generated by LLaVA, ControlMLLM and CoreGaze varies across images. Due to the additional iterative computations required for gaze diffusion, CoreGaze exhibits a corresponding increase in inference time, which is still less than the inference time of ControlMLLM. And the percentage of additional overhead decreases in large-image scenarios since the large baseline inference time. Regarding GPU memory usage, CoreGaze incurs an approximately 25% relative increase compared to LLaVA, which remains within acceptable limits for most application scenarios. In summary, while significantly enhancing comprehension and generation capabilities for referring regions, CoreGaze introduces only a moderate increase in inference latency and a slight growth in memory consumption, demonstrating its practicality and scalability.



(a) original image (b) att. before diffusion (c) att. after diffusion (d) 50% pruned image (e) 40% pruned image



(f) original image (g) att. before diffusion (h) att. after diffusion (i) 50% pruned image (j) 40% pruned image

Figure 5: Results of gaze diffusion and visual token pruning visualization, where “att." means “attention".

### 4.5 Visualization Results

In Figure 5, we visualize attention maps before / after core subgraph diffusion and pruning results of two images from LVIS. It can be observed that before diffusion, attention maps focus almost exclusively on nodes within the core subgraph. Through diffusion, the attention distribution significantly expands to surrounding semantic nodes directly connected to the core subgraph, incorporating additional context related to the referring box. Consequently, the pruned subgraph more precisely encompasses the target region and its critical context, preventing over-pruning or retaining extraneous background. This mechanism maintains high focus on the referring region while supplements essential background semantics, ultimately providing a more comprehensive and enriched input foundation for fine-grained comprehension.

### 4.6 Backbone Study

We explore the impact of using MLLMs of varying sizes as backbone models for CoreGaze, as shown in Table 4. Among them, InstructBLIP (Dai et al., 2023) builds upon BLIP with instruction tuning, enabling it to comprehend and execute complex multimodal instructions. InternVL2 (Wang et al., 2024) adopts a dual-tower fusion architecture consisting of a ViT visual encoder, lightweight cross-modal connection modules, and a large language model, enabling efficient feature alignment between vision and language. InternVL3 (Zhu et al., 2025) builds on a native integrated unified multimodal architecture, reconstructs the fundamental visual encoding modules and cross-modal interaction layers, deeply integrates visual perception and language modeling through full-stage joint pre-training. LLaVA-HR (Luo et al., 2025) undergoes supervised fine-tuning for high-resolution images to enhance its capability to capture and generate detailed features. Applying CoreGaze yields substantial gains for both 7B and 13B models. Even instruction-tuned InstructBLIP and high-resolution-optimized LLaVA-HR exhibit improvements comparable to those of general 7B models. This demonstrates that the core subgraph-guided diffusion and pruning strategy delivers consistent performance gains across different parameter scales and model architectures, thereby universally enhancing models' ability to focus on and comprehend referring regions.

| Model | Vanilla | CoreGaze |
|---|---|---|
| LLaVA-1.5-7B | 54.13 | **71.22** |
| LLaVA-1.5-13B | 55.69 | **72.97** |
| InstructBLIP-7B | 49.81 | **68.75** |
| InstructBLIP-13B | 54.33 | **71.01** |
| InternVL2-8B | 58.57 | **74.71** |
| InternVL3-8B | 61.36 | **76.73** |
| InternVL3-14B | 62.86 | **78.57** |
| LLaVA-HR-7B-SFT-1024 | 56.45 | **71.43** |

Table 4: Comparison results of using both different MLLMs and different model sizes on ROC (box).

## 5 Conclusion

In this paper, we introduce CoreGaze, a training-free referring MLLM framework simulating core

subgraph-guided visual gaze diffusion. Specifically, CoreGaze transforms images into sparse semantic graphs, where user-specified regions activate a core subgraph with amplified initial influence. Through restart-equipped random walks, CoreGaze propagates the relevance to contextual neighborhoods, and dynamically eliminate connectivity bias by degree-aware normalization. The resulting subgraph, which preserves both user-indicated targets and diffusion-activated context, enables frozen LLMs to achieve precise referring generation. Extensive validation confirms the effectiveness and expressiveness of CoreGaze.

## Limitations

While CoreGaze demonstrates strong performance in training-free referring multimodal understanding, it has several limitations. First, the method relies on the quality of the visual token representations from the frozen encoder. If the encoder fails to capture essential regional semantics, the graph-based diffusion may propagate noise or miss critical context. Additionally, CoreGaze assumes that the user-provided visual prompt accurately localizes the region of interest. In cases of ambiguous or poorly defined prompts, the core subgraph may not adequately represent the intended target. Finally, while CoreGaze supports arbitrary-shaped visual prompts, its performance may degrade when the prompt covers overly large or semantically heterogeneous regions, as the diffusion process might incorporate irrelevant context.

## Ethics Statement

In conducting this research, we have considered its potential ethical implications. The datasets used in our experiments are publicly available benchmarks widely adopted in the computer vision community. These datasets were curated for academic research purposes. However, as with many web-sourced image collections, we cannot fully rule out the possibility that they contain unintended biases or personal information, despite the efforts of their original creators. Moreover, the core technology presented is intended for positive applications, such as enhancing human-computer interaction and aiding in detailed visual analysis. As with any AI capability, there is a possibility of misuse, for instance, to generate misleading descriptions of images. We strongly discourage any such application and believe that the development of more accurate and controllable models is a step towards more transparent and accountable AI systems.

## Acknowledgments

This work was supported by the National Key R&D Program of China (2022YFB3103202) and the National Science and Technology Major Project of China (2025ZD1501602).

## References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: semantic propositional image caption evaluation. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*, volume 9909 of *Lecture Notes in Computer Science*, pages 382–398. Springer.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization@ACL 2005, Ann Arbor, Michigan, USA, June 29, 2005*, pages 65–72. Association for Computational Linguistics.

Mu Cai, Haotian Liu, Siva Karthik Mustikovela, Gregory P. Meyer, Yuning Chai, Dennis Park, and Yong Jae Lee. 2024. Vip-llava: Making large multimodal models understand arbitrary visual prompts. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 12914–12923. IEEE.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023. Shikra: Unleashing multimodal llm's referential dialogue magic. *CoRR*, abs/2306.15195.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven C. H. Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Junfeng Fang, Zac Bi, Ruipeng Wang, Houcheng Jiang, Yuan Gao, Kun Wang, An Zhang, Jie Shi, Xiang Wang, and Tat-Seng Chua. 2024. Towards neuron attributions in multi-modal large language models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Qiushan Guo, Shalini De Mello, Hongxu Yin, Wonmin Byeon, Ka Chun Cheung, Yizhou Yu, Ping Luo, and Sifei Liu. 2024. Regiongpt: Towards region understanding vision language model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13796–13806. IEEE.

Agrim Gupta, Piotr Dollár, and Ross B. Girshick. 2019. LVIS: A dataset for large vocabulary instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5356–5364. Computer Vision Foundation / IEEE.

Junwen He, Yifan Wang, Lijun Wang, Huchuan Lu, Jun-Yan He, Jin-Peng Lan, Bin Luo, and Xuansong Xie. 2024. Multi-modal instruction tuned llms with fine-grained visual perception. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13980–13990. IEEE.

Yutao Jiang, Qiong Wu, Wenhao Lin, Wei Yu, and Yiyi Zhou. 2025. What kind of visual tokens do we need? training-free visual token pruning for multi-modal large language models from the perspective of graph. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 4075–4083. AAAI Press.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. 2025. See what you are told: Visual attention sink in large multimodal models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 787–798. ACL.

Weifeng Lin, Xinyu Wei, Ruichuan An, Peng Gao, Bocheng Zou, Yulin Luo, Siyuan Huang, Shanghang Zhang, and Hongsheng Li. 2025a. Draw-and-understand: Leveraging visual prompts to enable MLLMs to comprehend what you want. In *The Thirteenth International Conference on Learning Representations*.

Weifeng Lin, Xinyu Wei, Ruichuan An, Peng Gao, Bocheng Zou, Yulin Luo, Siyuan Huang, Shanghang Zhang, and Hongsheng Li. 2025b. Draw-and-understand: Leveraging visual prompts to enable mllms to comprehend what you want. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Zhihang Lin, Mingbao Lin, Luxi Lin, and Rongrong Ji. 2025c. Boosting multimodal large language models with visual tokens withdrawal for rapid inference. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 5334–5342. AAAI Press.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Gen Luo, Yiyi Zhou, Yuxin Zhang, Xiawu Zheng, Xiaoshuai Sun, and Rongrong Ji. 2025. Feast your eyes: Mixture-of-resolution adaptation for multimodal large language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Chuofan Ma, Yi Jiang, Jiannan Wu, Zehuan Yuan, and Xiaojuan Qi. 2024. Groma: Localized visual tokenization for grounding multimodal large language models. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VI*, volume 15064 of *Lecture Notes in Computer Science*, pages 417–435. Springer.

Bowen Pan, Rameswar Panda, Yifan Jiang, Zhangyang Wang, Rogério Feris, and Aude Oliva. 2021. Ia-red$^2$: Interpretability-aware redundancy reduction for vision transformers. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 24898–24911.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. 2024. Grounding multimodal large language models to the world. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and

Marcus Rohrbach. 2019. Towards VQA models that can read. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 8317–8326. Computer Vision Foundation / IEEE.

Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. 2024. Alpha-clip: A CLIP model focusing on wherever you want. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13019–13029. IEEE.

Yunjie Tian, TianRen Ma, Lingxi Xie, Jihao Qiu, Xi Tang, Yuan Zhang, Jianbin Jiao, Qi Tian, and Qixiang Ye. 2024. Chatterbox: Multi-round multimodal referring and grounding. *CoRR*, abs/2401.13307.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 4566–4575. IEEE Computer Society.

Andreas Veit, Tomas Matera, Lukás Neumann, Jiri Matas, and Serge J. Belongie. 2016. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *CoRR*, abs/1601.07140.

Weiyun Wang, Zhe Chen, Wenhai Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Jinguo Zhu, Xizhou Zhu, Lewei Lu, Yu Qiao, and Jifeng Dai. 2024. Enhancing the reasoning ability of multimodal large language models via mixed preference optimization. *CoRR*, abs/2411.10442.

Mingrui Wu, Xinyue Cai, Jiayi Ji, Jiale Li, Oucheng Huang, Gen Luo, Hao Fei, Guannan Jiang, Xiaoshuai Sun, and Rongrong Ji. 2024. Controlmllm: Training-free visual prompt learning for multimodal large language models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Jiarui Xu, Xingyi Zhou, Shen Yan, Xiuye Gu, Anurag Arnab, Chen Sun, Xiaolong Wang, and Cordelia Schmid. 2024. Pixel aligned language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13030–13039. IEEE.

Weihao Ye, Qiong Wu, Wenhao Lin, and Yiyi Zhou. 2025. Fit and prune: Fast and training-free visual token pruning for multi-modal large language models. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 22128–22136. AAAI Press.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. 2024. Ferret: Refer and ground anything anywhere at any granularity. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Tongtian Yue, Jie Cheng, Longteng Guo, Xingyuan Dai, Zijia Zhao, Xingjian He, Gang Xiong, Yisheng Lv, and Jing Liu. 2024. SC- tune: Unleashing self-consistent referential comprehension in large vision language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13073–13083. IEEE.

Xingchen Zeng, Haichuan Lin, Yilin Ye, and Wei Zeng. 2025. Advancing multimodal large language models in chart question answering with visualization-referenced instruction tuning. *IEEE Trans. Vis. Comput. Graph.*, 31(1):525–535.

Hao Zhang, Hongyang Li, Feng Li, Tianhe Ren, Xueyan Zou, Shilong Liu, Shijia Huang, Jianfeng Gao, Leizhang, Chunyuan Li, and Jainwei Yang. 2024a. Llava-grounding: Grounded visual chat with large multimodal models. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLIII*, volume 15101 of *Lecture Notes in Computer Science*, pages 19–35. Springer.

Haotian Zhang, Haoxuan You, Philipp Dufter, Bowen Zhang, Chen Chen, Hong-You Chen, Tsu-Jui Fu, William Yang Wang, Shih-Fu Chang, Zhe Gan, and Yinfei Yang. 2024b. Ferret-v2: An improved baseline for referring and grounding with large language models. *CoRR*, abs/2404.07973.

Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Yu Liu, Kai Chen, and Ping Luo. 2024c. Gpt4roi: Instruction tuning large language model on region-of-interest. In *Computer Vision - ECCV 2024 Workshops - Milan, Italy, September 29-October 4, 2024, Proceedings, Part VIII*, volume 15630 of *Lecture Notes in Computer Science*, pages 52–70. Springer.

Xiaofeng Zhang, Fanshuo Zeng, Yihao Quan, Zheng Hui, and Jiawei Yao. 2025. Enhancing multimodal large language models complex reason via similarity computation. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pages 10203–10211. AAAI Press.

Liang Zhao, En Yu, Zheng Ge, Jinrong Yang, Haoran Wei, Hongyu Zhou, Jianjian Sun, Yuang Peng, Runpei Dong, Chunrui Han, and Xiangyu Zhang. 2024. Chatspot: Bootstrapping multimodal llms via precise referring instruction tuning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024, Jeju, South Korea, August 3-9, 2024*, pages 1743–1752. ijcai.org.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2024. Minigpt-4: Enhancing

vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, and 32 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *CoRR*, abs/2504.10479.

# A Hyper-parameter Study

## A.1 Pruning Ratio

Table 5 reports the effect of varying the pruning ratio $k$ for selecting the most salient tokens in the final index set on RTC. It can be observed that performance peaks at $k = 40\%$ for box prompts and at $k = 30\%$ for mask prompts. Within this range, the pruned subgraph is sufficiently compact to eliminate the majority of distracting, irrelevant regions, thereby forcing the model to concentrate on the user's query. Simultaneously, it preserves enough semantically linked contextual tokens to support accurate and coherent comprehension. As the pruning ratio increases beyond this optimal point, a consistent decline in performance for both prompt types. This degradation occurs because a larger $k$ incorporates an increasing number of tokens with lower relevance scores, which effectively reintroduces noise and distraction that the gaze diffusion process sought to eliminate. The model's attention is once again diluted by semantically weaker regions, leading to less precise grounding and a drop in accuracy.

| | Box | Mask |
|---|---|---|
| $k = 30\%$ | 73.91 | **71.88** |
| $k = 40\%$ | **74.23** | 71.43 |
| $k = 50\%$ | 74.07 | 71.23 |
| $k = 60\%$ | 72.41 | 70.00 |

Table 5: Hyper-parameter results of the pruning ratio $k$ on RTC when selecting tokens to form the pruned set.

## A.2 Similarity Threshold

Moreover, we explore the impact of the similarity threshold on CoreGaze, with results shown in Table 6. Higher thresholds yield sparser adjacency matrix $A$, consequently affecting the construction of normalized matrix $P$. It can be observed that performance remains comparable when thresholds range between 0.3 and 0.5, whereas a threshold of 0.7 causes performance degradation. This indicates that the high threshold during sparse adjacency matrix construction leads to overly sparse graph structures, which impairs the core subgraph diffusion's ability to propagate contextual importance.

| | Box | Mask |
|---|---|---|
| $s = 0.7$ | 73.13 | 70.48 |
| $s = 0.5$ | 74.07 | 71.13 |
| $s = 0.3$ | **74.23** | **71.43** |

Table 6: Hyper-parameter results of the similarity threshold $s$ on RTC when constructing the sparse adjacency matrix.

## A.3 Restart Probability

Table 7 shows the impact of the restart probability $\alpha$ on CoreGaze. The restart probability controls the trade-off between reinforcing the core subgraph's influence and exploring contextual neighborhoods during the gaze diffusion process. It can be observed that a higher restart probability yields the best performance, achieving 74.70% and 72.15% accuracy for box and mask, respectively. This indicates that a stronger emphasis on the core subgraph helps maintain focus on the user-specified region, preventing the diffusion process from drifting to semantically distant or irrelevant nodes. As $\alpha$ decreases to 0.05, performance gradually declines to 73.47% and 71.00%, suggesting that insufficient reinforcement of the core region leads to a loss of referential precision. These results highlight the importance of balancing core reinforcement and contextual exploration. A moderate-to-high $\alpha$ ensures that the model retains strong grounding in the target region while still leveraging relevant contextual cues, which is crucial for accurate referring comprehension in training-free settings.

| | Box | Mask |
|---|---|---|
| $\alpha = 0.35$ | **74.70** | **72.15** |
| $\alpha = 0.25$ | 74.36 | 71.59 |
| $\alpha = 0.15$ | 74.23 | 71.43 |
| $\alpha = 0.05$ | 73.47 | 71.00 |

Table 7: Hyper-parameter results of the restart probability $\alpha$ on RTC when performing random walk.

### A.4 Iterations

Table 8 is the impact of the iterations $t$ when performing random walk on CoreGaze. The number of iterations $t$ in the random walk process determines how far the gaze diffusion propagates from the core subgraph. As $t$ increases from 1 to 15, we observe a general trend of improving performance, with the best accuracy achieved at $t = 15$. This indicates that more iterations allow the model to capture broader contextual relationships, which contributes to more robust referring comprehension. However, the performance gain beyond $t = 5$ is marginal. Specifically, the accuracy at $t = 5$ is already competitive, and further increasing $t$ to 10 or 15 only brings minor improvements. Considering the computational cost associated with additional iterationsm, we choose $t = 5$ as the default setting for CoreGaze. This offers a favorable balance between model performance and inference efficiency, making it suitable for practical deployment while maintaining strong referential grounding.

| | Box | Mask |
|---|---|---|
| $t = 15$ | **74.73** | **71.84** |
| $t = 10$ | 74.19 | 71.15 |
| $t = 5$ | 74.23 | 71.43 |
| $t = 1$ | 73.68 | 70.97 |

Table 8: Hyper-parameter results of the iteration $t$ on RTC when performing random walk.

## B Screenshot Study

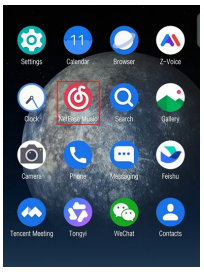
In the screenshot-understanding tasks presented in Figure 6, CoreGaze demonstrates a clear advantage in grounding user queries to the correct UI elements. In the first example, the textual prompt asks “What is this app in the picture used for?", both the vanilla LLaVA and its simple attention editing or blur-based variants produce generic or outright incorrect responses attributing the interface to weather, messaging, or social media functions, since their attention remains diffused across the entire screen. CoreGaze, by contrast, isolates the music-player region specified by the bounding box and accurately identifies its purpose as “music streaming and sharing", showing that the core subgraph diffusion effectively filters out unrelated icons and background elements.

The second example, the textual prompt asks “What is this icon used for?", further highlights CoreGaze’s capacity to resolve fine-grained visual details. Other methods mistakenly label the selected camera icon as a generic cell phone or misinterpret it entirely, reflecting their inability to focus attention on the precise UI token. CoreGaze’s attention visualization confirms that its random-walk diffusion from the core subgraph converges on the camera glyph, enabling the model to correctly describe its function. This elimination of hallucinated attributions underscores how CoreGaze’s pruning and propagation steps translate into more faithful, context-aware descriptions of interface components.

## C Referring Description Task Study

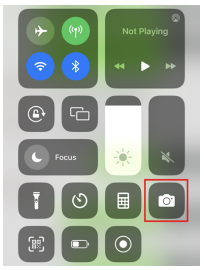
Following prior work configurations (Wu et al., 2024), we compare the performance of LLaVA (Liu et al., 2023) and its variants, ControlMLLM (Wu et al., 2024), and CoreGaze on the referring description task (RDT), with results presented in Table 9. Specifically, we construct a test set based on region-text pairs from the test splits of RefCOCOg, RefCOCO, and RefCOCO+ (Kazemzadeh et al., 2014), using traditional captioning metrics including BLEU@4 (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), CIDEr-D (Vedantam et al., 2015), and SPICE (Anderson et al., 2016), which are metrics for assessing image-text generation quality. BLEU@4 measures phrase-level overlap between generated and reference texts through exact 1-to-4-gram matching, primarily evaluating linguistic accuracy. METEOR extends this foundation by incorporating lemmatization, synonym matching, and word order penalties to better align with human semantic understanding. CIDEr-D employs TF-IDF weighting to score generated content based on consensus among reference texts, emphasizing sentence diversity and semantic relevance. SPICE adopts a semantic-structure perspective by parsing sentences into scene-graph triplets (objects, attributes, relations) to directly assess the correctness of visual-semantic expressions, making it suitable for analyzing model comprehension depth.

Specifically, CoreGaze’s gains on region description go well beyond raw score bumps, which reflect qualitatively more focused and semantically coherent outputs. On RefCOCO, LLaVA-Edit Att essentially collapses (BLEU@4 and METEOR near zero), showing that naive attention editing cannot reliably ground text in the target box. ControlMLLM partially restores fluency and modest ground-



(a) input image 1

Prompt 1: What is this app in the picture used for?
LLaVA: The app in the picture is used for displaying various types of information, such as the time, weather, and other relevant data. The app's interface features a collection of icons representing these different aspects of information...
LLaVA+Att: The app in the picture is used for communication purposes, as it features a messaging icon. This suggests that the app allows users to send and receive text messages, which is a common function of messaging apps.
LLaVA+Blur: The app in the picture is a social media platform, specifically designed for sharing and discovering content related to food and cooking. It allows users to create and share their own recipes, photos, and videos with a community of food enthusiasts...
CoreGaze: The app in the picture is used for music streaming and sharing. It is a music streaming service that allows users to listen to and share music with others...



(b) input image 2

Prompt 2: What is this icon used for?
LLaVA: The icon in the image is a cell phone, which is used for making and receiving phone calls, sending and receiving text messages, and accessing various apps and services on a smartphone...
LLaVA+Att: The icon in the image is a sun, which is often used to represent the sun or sunlight. In the context of the image, it is likely used to represent a bright or sunny day, or to indicate the presence of sunlight in the scene...
LLaVA+Blur: The icon represents the camera function of a digital camera or smartphone. It is commonly associated with photography and image capture. In the context of the image, it is likely that the camera is being used to capture a photo or record a video.
CoreGaze: This icon, which is a square with a camera inside, is commonly used to represent a camera or photography feature. It is often found in various applications, such as photo editing software, camera settings, or even as a shortcut to access the camera function on a smartphone or other devices...

Figure 6: Results on screenshot with different input images, where correct descriptions are marked in green while incorrect descriptions are marked in red, and hallucinated descriptions are marked in another color.

| | RefCOCO | | | | RefCOCO+ | | | | RefCOCOg | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | B@4 | M | C | S | B@4 | M | C | S | B@4 | M | C | S |
| LLaVA | 0.76 | 9.63 | 21.26 | 12.96 | 0.95 | 10.86 | 20.94 | 13.69 | 8.62 | 14.13 | 54.32 | 16.72 |
| LLaVA-Blur | 1.32 | 12.32 | 29.71 | 16.90 | 1.42 | 13.38 | 29.90 | 18.67 | 10.91 | 16.57 | 69.68 | 20.91 |
| LLaVA-Edit Att | 0.00 | 0.08 | 0.17 | 0.12 | 0.00 | 0.23 | 0.34 | 0.27 | 0.10 | 0.23 | 0.28 | 0.16 |
| FitPrune | 0.69 | 10.13 | 21.87 | 13.98 | 0.94 | 11.15 | 22.36 | 14.14 | 7.96 | 15.43 | 56.26 | 18.21 |
| ControlMLLM | 1.00 | 10.67 | 23.36 | 14.31 | 1.20 | 12.16 | 25.09 | 15.61 | 10.26 | 15.53 | 62.39 | 19.35 |
| CoreGaze | 1.13 | 11.58 | 28.85 | 16.27 | 1.22 | 13.20 | 28.38 | 18.16 | 10.53 | 16.25 | 65.35 | 20.10 |

Table 9: Results on RDT, which describes the user-specified region. Gray font represents the lower / upper bounds achieved using LLaVA.

ing, but its CIDEr-D and SPICE remain tethered to background cues. By contrast, CoreGaze not only nearly doubles BLEU@4 over ControlMLLM, it also lifts SPICE by over one full point, indicating that the generated descriptions better capture the scene graph relationships around the referent rather than generic object mentions.

This pattern persists on RefCOCO+ and RefCOCOg, despite their longer expressions and more varied contexts. On RefCOCO+, where referring expressions often include comparative or attribute-based modifiers, CoreGaze's precise subgraph pruning yields sharper CIDEr-D gains, demonstrating its ability to incorporate those modifiers into fluent yet focused sentences. Even on the sprawling RefCOCOg split, CoreGaze sustains consistent upticks in BLEU@4 and CIDEr-D, showing robustness to increased linguistic complexity. Moreover, CoreGaze's SPICE improvements across all three splits confirm that the model is better at encoding relations. These results illustrate that core-subgraph-guided diffusion and token selection transform noisy attention maps into semantically rich inputs, yielding region descriptions that are both accurate and contextually precise.

## D High Resolution Image Study

Figure 7 compares the outputs of LLaVA, ControlMLLM, and CoreGaze on a single high-resolution TextVQA (Singh et al., 2019) image (about 1K resolution). LLaVA's response is overly generic, describing the scene rather than answering the question. ControlMLLM improves fluency but still hallucinates details unrelated to the query, indicating that mere visual prompt tuning cannot fully resolve the grounding challenge. CoreGaze, however, focuses its generated text squarely on the objects within the user-specified region and produces the correct answer, demonstrating that its subgraph-driven diffusion and pruning pipeline ef-

(a) original image

(b) visual prompt

Prompt: Describe the object.

LLaVA: The image features a bar with a large collection of liquor bottles and glasses. The bottles are arranged in various positions, with some placed on the shelves and others scattered throughout the scene. The glasses are also placed in different areas, creating a visually appealing display. The bar appears to be well-stocked, offering a wide selection of drinks for patrons to choose from.

ControlMLLM: The object in the image is a hanging light fixture with a sign attached to it. The sign is written in Spanish and reads "Illegal to sell tobacco." The light fixture is hanging from the ceiling, and it is surrounded by a variety of bottles, possibly containing different types of liquor.

CoreGaze: The object in the image is a light fixture hanging from the ceiling. It is a lamp with a shade, which is illuminating a bar filled with various liquor bottles. The bottles are arranged in a cabinet, and the light shines on them, creating a warm and inviting atmosphere in the bar. The light fixture adds a touch of elegance and sophistication to the space.

Figure 7: Results on a high resolution image with different methods, where correct descriptions are marked in green while incorrect descriptions are marked in red, and hallucinated descriptions are marked in orange.

(a) original image

(b) 50% pruned image

(c) 40% pruned image

(d) 30% pruned image

Figure 8: Visualization results of visual token pruning on a high resolution image.

| | Tokens | Speed (s) | GPU |
|---|---|---|---|
| LLaVA | 85 | 8.2051 | 8G |
| ControlMLLM | 92 | 10.7882 | 12G |
| CoreGaze | 97 | 9.1969 | 10G |

Table 10: The inference cost of LLaVA, LLaVA-based ControlMLLM, and LLaVA-based CoreGaze in a high resolution image on a Tesla A100 GPU.

fectively filters out background noise and steers the model toward the relevant tokens.

Table 10 reports the inference cost of each method when applied to the same high-resolution image in Figure 7 on a Tesla A100 GPU. LLaVA processes 85 visual tokens in 8.2051s with 8 GB of GPU memory, while ControlMLLM actually reduces token count to 92 and slightly speeds up inference at the expense of a jump to 12 GB of memory. CoreGaze increases the token count to 97 due to its subgraph expansion, resulting in a 9.1969s runtime and 10 GB peak memory usage. These measurements show that although CoreGaze incurs a modest overhead in both computation time and memory (relative to LLaVA), it remains within practical bounds and delivers substantially improved grounding performance.

Figure 8 visualizes how varying the pruning ratio affects the remaining visual tokens on an image from HR-Bench (Jiang et al., 2025) (about 4K resolution). At a 50% pruning rate, the retained tokens still include many peripheral elements. When reducing to 40%, it begins to concentrate on the true region of interest while preserving enough context for interpretation. Further tightening to 30% yields a compact yet semantically coherent subgraph that aligns closely with the target area. This sequence illustrates that CoreGaze's top-$k$ token selection can be tuned to balance context retention against noise suppression, enabling robust performance across

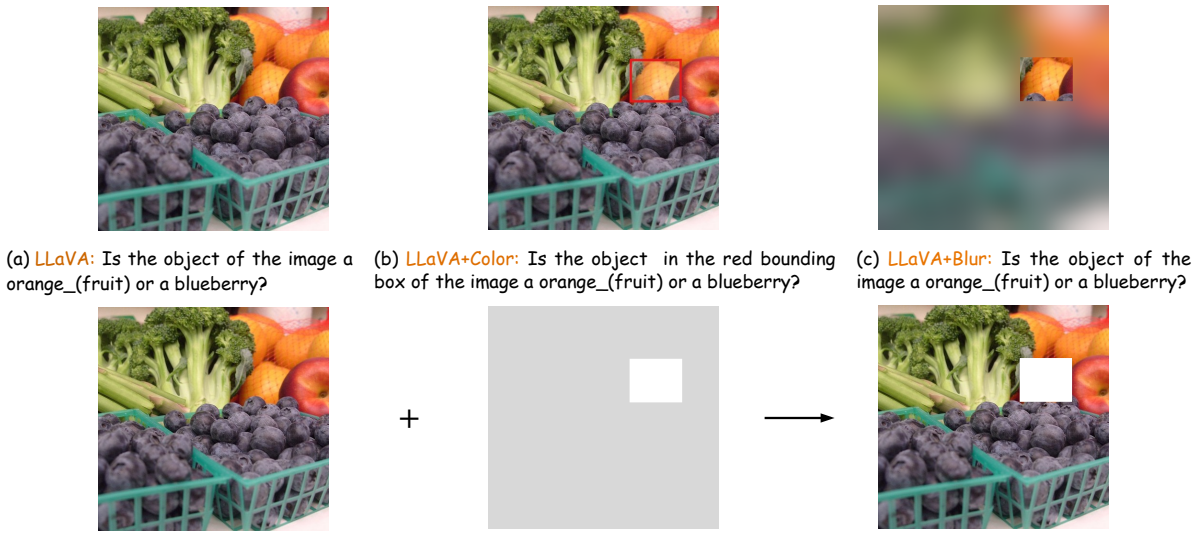



Figure 9: The input example of the referring dialogue with box.

varying image resolutions and content complexities.

## E Input Image Details

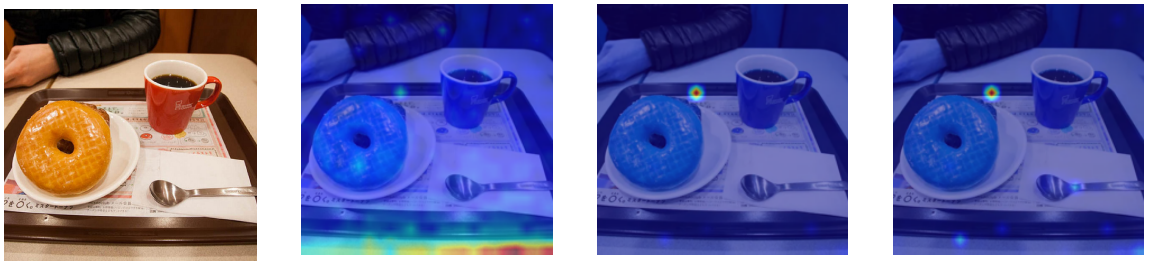
Figure 9 illustrates how different methods incorporate the visual prompt into the input image from LVIS (Gupta et al., 2019). Specifically, the user-specified region is indicated by a red bounding box in the referring dialogue. LLaVA receives only the original image without any visual modification, lacking any explicit spatial grounding signal. LLaVA-Color overlays a red box directly onto the image, making the target region visible but without removing surrounding distractions. LLaVA-Blur attempts to reduce noise by blurring the background while retaining the area inside the box. However, both color-overlay and blur-based variants still operate on the full image tensor, and as shown in earlier results, often fail to sufficiently suppress irrelevant context.

In contrast, CoreGaze isolates the original image along with only the user-specified region inside the red box, leveraging a structured visual prompt that discards uninformative areas during subgraph construction. This focused input allows the model to concentrate on semantically meaningful regions while preserving features within the prompt area. Compared to LLaVA-Blur or LLaVA-Color, which rely on superficial visual editing, CoreGaze incorporates the prompt into a graph-based attention mechanism, enabling deeper structural integration and more precise grounding. This difference in how visual prompts are operationalized is key to CoreGaze's superior performance in both referential understanding and multimodal reasoning.

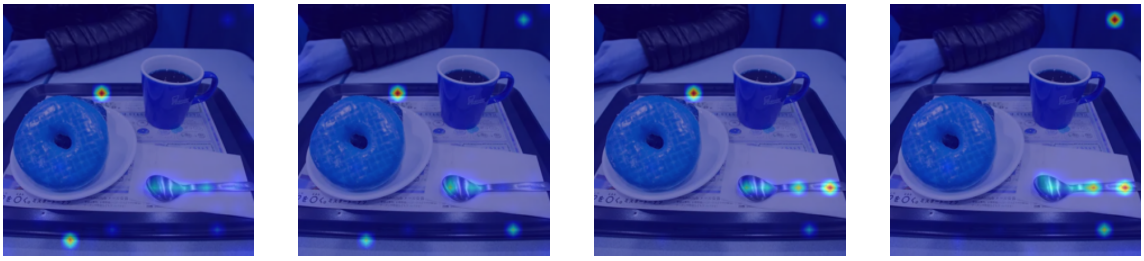
## F Attention Map Visualization

Figure 10 presents a layer-wise attention visualization from the LLM decoder for a sample image from LVIS, where the user-specified region corresponds to a spoon on the table. By examining attention maps at layers 0, 4, 8, 13, 18, 23, and 31, we observe how attention evolves throughout the decoding process. In the lower layers, attention is highly diffuse, with the model distributing focus broadly across the entire scene, including irrelevant background objects. This suggests that initial layers are still dominated by general visual feature alignment and lack fine-grained localization.

As the decoding progresses into the middle layers (layers 8 to 18), attention begins to narrow, gradually concentrating more on the user-highlighted region. By layers 23 and above, the model's attention becomes tightly focused on the spoon, indicating that semantic alignment and referential grounding have become more precise. This progression reflects the benefit of CoreGaze's structured input. By explicitly encoding the user-specified region and propagating contextual cues via subgraph diffusion, the model is guided toward the correct object over the course of decoding. The visualizations demonstrate that CoreGaze not only provides an initial localization cue, but also sustains and reinforces that cue through deeper layers, leading to



(a) original image (b) layer 0 (c) layer 4 (d) layer 8



(e) layer 13 (f) layer 18 (g) layer 23 (h) layer 31

Figure 10: Visualization results of attention maps in different layers.

more accurate and grounded generation.

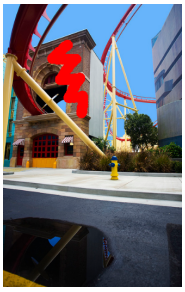
## G Textual Prompt Study

Figure 11 illustrates how the phrasing of textual prompts directly affects the model's ability to generate semantically grounded and functionally accurate descriptions. In the first example, the image depicts a roller coaster station, a highly specific type of structure that serves as both the starting and stopping point of a ride. When given a generic prompt such as "What is this?", LLaVA returns vague or surface-level descriptions, such as "a building" or "a theme park area", failing to capture the functional role of the structure. Only the final prompt, which explicitly asks about the purpose of the region, enables the model to correctly identify it as "a roller coaster station, which serves as the starting point and stopping point for the ride". This demonstrates that CoreGaze, when guided by well-structured prompts, can go beyond object recognition and infer context-sensitive functionality that is often missed by standard attention-based models.

In the second example, the target region is a close-up of a piece of fabric. All models eventually recognize the cut or opening in the material, but only the most specific prompt, which is combined with CoreGaze's structured grounding, leads to a complete description that includes not just the incision, but also the presence of folds and signs of wear. These subtle visual cues are critical for accurate material interpretation but are often overlooked in generic answers. The result highlights that CoreGaze is capable of attending to fine-grained visual attributes such as texture, damage, and structural deformation, especially when the textual prompt explicitly emphasizes those aspects. As a result, these examples confirm that precise language grounding is essential for maximizing CoreGaze's reasoning potential, enabling it to combine visual and linguistic information in a contextually rich and semantically complete manner.

## H Visual Prompt Study

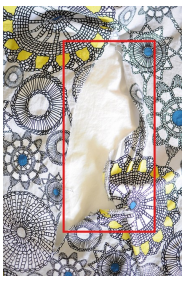
Figure 12 investigates the impact of visual prompt size on the model's ability to generate accurate and functionally grounded descriptions. In this case, the target is a roller coaster station that serves as both the starting and stopping point of the ride. The visual prompts are provided in arbitrary shapes, allowing flexible selection of the relevant region. When the prompt is too small and includes only a limited portion of the structure, such as a part of the roof or a single section of the platform, the model struggles to recognize the function of the building. The generated responses in these cases are vague or incorrect, often referring to the scene as a general building or theme park area, which reflects the lack of sufficient visual context for reasoning about its functional identity.



(a) input image 1

Prompt 1: What is the building?
LLaVA: The building in the image is a roller coaster, which is a type of amusement park ride.
CoreGaze: The building in the image is a large brick structure, possibly a part of a roller coaster.

Prompt 2: Describe the spatial location and function of this building.
LLaVA: The building is located in a city, possibly near a theme park or an amusement park, as it is surrounded by a roller coaster and a ferris wheel. The building's function is not clear from the image, but it could be a part of the theme park or an adjacent establishment, such as a restaurant or a shop...
CoreGaze: The building in the image is a roller coaster station, which **serves as a starting point and a stopping point for the roller coaster ride**. It is located on a street corner, and the roller coaster ride passes by it. The presence of a fire hydrant nearby indicates that the station is situated in an urban environment...



(b) input image 2

Prompt 1: Describe this area.
LLaVA: The image features a close-up of a black and white patterned fabric. The fabric has a mix of black and white colors, as well as some yellow accents...
CoreGaze: The image features a piece of fabric with a black and white pattern. The fabric is cut in half, revealing the pattern on both sides...

Prompt 2: Describe the form of this fabric.
LLaVA: The fabric is a black and white, possibly blue, patterned piece with a design resembling a lace or a net. The fabric appears to be a part of a larger piece, possibly a tablecloth or a curtain, with a cut or tear in it...
CoreGaze: The fabric is a black and white, possibly blue, patterned piece with a design resembling a flower. It appears to be a piece of clothing, possibly a dress or a shirt, with a cutout in the middle. The fabric is **folded and has a somewhat worn appearance**, suggesting that it has been used or washed multiple times.

Figure 11: Results on input images with different textual prompts, where correct descriptions are marked in green while incorrect descriptions are marked in red, and the most correct descriptions are in bold.

As the prompt expands to include a larger and more semantically coherent portion of the scene, such as the track entry and exit points, surrounding equipment, and signage, the model begins to generate more accurate descriptions. When the prompt effectively captures the operational layout of the station, the model is able to recognize not just the structure but its role in the roller coaster system. The most complete and correct answer is produced when the prompt covers a broad yet focused region that contains the key visual evidence needed to infer both the type of building and its specific purpose. These findings suggest that arbitrary-shaped visual prompts, when sufficiently informative, allow CoreGaze to ground language in complex functional contexts with greater precision.

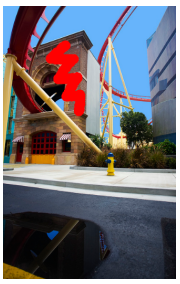
However, when the visual prompt becomes too large and extends beyond the immediate structure into unrelated parts of the scene, the model's output begins to lose specificity. In these cases, the prompt includes not only the roller coaster station but also large portions of the background, such as nearby pathways, other buildings, or scenery that is not directly relevant to the query. As a result, the model's attention becomes diluted across a broader area, leading to descriptions that are less focused or even incorrect. Instead of identifying the station and its function, the model may shift toward describing the general environment or mention objects unrelated to the user's intent.

This observation highlights the importance of carefully balancing the scope of visual prompting. While a prompt that is too small lacks the necessary context, one that is too large introduces semantic noise that can mislead the model during generation. The ideal visual prompt is one that captures the key components related to the query while minimizing irrelevant content. CoreGaze's support for arbitrary-shaped regions offers the flexibility to achieve this balance, enabling prompts that are neither too narrow nor overly expansive, and thus ensuring accurate, context-aware responses in visually complex scenes.

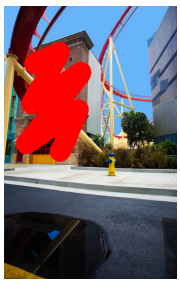
(a) original image



(b) visual prompt 1



(c) visual prompt 2



(d) visual prompt 3

Prompt: Describe the spatial location and function of this building.

CoreGaze 1: The building in the image is a large, multi-story structure with a mix of architectural styles. It features a combination of red brick and yellow elements, giving it a unique and eye-catching appearance. The building is located on a street corner, and it appears to be a restaurant or a similar type of establishment. The presence of a fire hydrant nearby suggests that the building is situated in an urban environment. The combination of the building's design, its location, and the presence of a fire hydrant indicate that it serves as a dining or entertainment venue for the local community.

CoreGaze 2: The building in the image is a roller coaster station, which serves as a starting point and a stopping point for the roller coaster ride. It is located on a street corner, and the roller coaster ride passes by it. The station is designed to accommodate the roller coaster's movement and ensure the safety of the riders. The presence of a fire hydrant nearby indicates that the station is situated in an urban environment.

CoreGaze 3: The building in the image is a large, multi-story structure with a mix of architectural styles. It features a fire hydrant on the sidewalk, a yellow pole, and a red roller coaster. The building's function is not entirely clear, but it could be a combination of a restaurant, a shopping center, or a theme park. The presence of the roller coaster suggests that it might be a theme park or an amusement park, where people can enjoy various rides and attractions.

Figure 12: Results on an image with different visual prompts, where correct descriptions are marked in green while incorrect descriptions are marked in red, and approximately correct descriptions are marked in another color.