Appendix

A. Experimental Details

Experimental Setting. All experiments and evaluations are conducted on a single NVIDIA GeForce RTX A6000 48GB GPU. We only use the inference stage of the models without any fine-tuning or training.

Analysis Setting. In Sec. 4, we identify and analyze localization heads in various LVLMs. We use the RefCOCO training set to prevent validation set leakage. To calculate the selection frequency of individual heads, we randomly select 1,000 image-text pair samples from the RefCOCO training set and average the results over five trials to validate consistency. When analyzing the selection frequency and IoU, we binarize the attention weights by assigning 1 above the mean value and 0 below it and calculate the IoU between the binarized attention weights and the ground-truth mask. We repeat this process for 1,000 image-text pairs and average the IoU scores.

Dataset Details. We evaluate our method on the following datasets:

- RefCOCO, RefCOCO+ [24], and RefCOCOg [19] datasets, sourced from MS-COCO [36], offer a collection of referring expressions and associated images. RefCOCO consists of 19,994 images paired with 142,210 expressions, while RefCOCO+ includes 19,992 images and 141,564 expressions. RefCOCOg, on the other hand, contains 26,771 images and 104,560 expressions. The expressions in RefCOCO and RefCOCO+ are generally concise, with an average of 1.6 nouns and 3.6 words per expression. In contrast, RefCOCOg features more descriptive expressions, averaging 2.8 nouns and 8.4 words.
- ReasonSeg: The dataset and benchmark for reasoning segmentation were first introduced in LISA [30]. The resulting ReasonSeg benchmark consists of 1,218 image-instruction-mask data samples, which are further divided into three splits: training (239 samples), validation (200 samples), and test (779 samples).

Main Experiments Setting. We evaluate our method on the following tasks:

- Referring Expression Comprehension (REC) and Referring Expression Segmentation (RES): The datasets evaluated for the main results in Sec.6.2 include RefCOCO (validation, test-A, test-B), RefCOCO+ (validation, test-A, test-B), and RefCOCOg (validation, test). All evaluations were conducted using the UNC split.
- Reasoning Segmentation (ReasonSeg): Reasoning Segmentation was first introduced in LISA [30]. This task shares a similar formulation with the referring expression segmentation task but is considerably more challenging. The key distinction lies in the complexity of the query text in reasoning segmentation. Rather than simple phrases (e.g., "the blue mug"), the queries involve more nuanced descriptions (e.g., "the container used for drinking, located next to the plate") or longer sentences (e.g., "Find the item on the table that someone would use to hold liquid, often paired with a saucer"). These queries demand advanced reasoning and a deeper understanding of contextual and world knowledge. All reasoning segmentation results were evaluated using the ReasonSeg benchmark, which includes both the validation set and test set. Performance was measured across short queries, long queries, and overall, following the same experimental setup as LISA [30] to ensure consistency in comparisons.

Ablation Studies Setting. In Sec. 6.3, we ablate the effectiveness of each criterion and validate the selection methods. In this section, we provide details of the ablation studies.

For criterion ablation, we consider two approaches: (1) selecting heads based solely on the highest $S_{\text{img}}^{\ell,h}$ values, or (2) selecting heads based solely on the lowest $H(\boldsymbol{A}^{\ell,h})$ values. In approach (1), we select the 10 heads with the highest $S_{\text{img}}^{\ell,h}$ values and calculate their selection frequency. Similarly, in approach (2), we select the 10 heads with the lowest $H(\boldsymbol{A}^{\ell,h})$ values and calculate their selection frequency.

For selection validation, we introduce the 'greedy' selection method, which selects the top-k heads per sample without considering the overall selection frequency. When applying the greedy selection method and criterion (1) simultaneously, we select the top-k heads with the highest $S_{\text{img}}^{\ell,h}$ values for each sample. Criterion (2) is applied in a similar manner, simultaneously selecting the top-k heads with the lowest $H(\boldsymbol{A}^{\ell,h})$ values for each sample.

B. Detailed Description of Algorithms

B.1. Spatial Entropy

Spatial entropy [2] adjusts the probability of attention being focused in a region by factoring in the size of that region, ensuring fair comparison across areas of different sizes. Note that, our spatial entropy calculation is based on the previous work [45] which validated the effectiveness of spatial entropy in image attention maps within vision transformer. We begin by computing the image attention map $\boldsymbol{A}^{\ell,h}$ as follows:

$$\boldsymbol{A}^{\ell,h} = \text{ReLU}\left(\text{reshape}(\boldsymbol{a}^{\ell,h}) - m\right), \tag{4}$$

where the ReLU function is applied after reshaping by $P \times P$, and it retains only those values in $\boldsymbol{a}^{\ell,h}$ that are greater than the mean m. Next, we identify the connected components $C_{\boldsymbol{A}^{\ell,h}} = \{C_1, C_2, \ldots, C_n\}$ from $\boldsymbol{A}^{\ell,h}$:

$$C_{\boldsymbol{A}^{\ell,h}} = \text{ConnectedComponents}(\boldsymbol{A}^{\ell,h}), \tag{5}$$

where the connected components are determined by applying an 8-connectivity relation among the non-zero elements of $\boldsymbol{A}^{\ell,h}$. Each connected component C_n (with

$1 \leq n \leq N$) in $C_{\boldsymbol{A}^{\ell,h}}$ is defined as the set of coordinates $C_n = \{(x_1, y_1), (x_2, y_2), \ldots, (x_{k_n}, y_{k_n})\}$ for the n-th component, where $k_n = |C_n|$ represents the cardinality, or the number of elements, in C_n. Finally, we calculate the spatial entropy $H(\boldsymbol{A}^{\ell,h})$ as follows:

$$H(\boldsymbol{A}^{\ell,h}) = -\sum_{n=1}^{N} P(C_n) \log P(C_n), \tag{6}$$

where this entropy is computed using Shannon's entropy formula. Here, $P(C_n)$ represents the probability of observing each connected component C_n within $\boldsymbol{A}^{\ell,h}$. The probability $P(C_n)$ for each component C_n is defined as:

$$P(C_n) = \frac{|C_n|}{\sum_{n=1}^{N} |C_n|}, \tag{7}$$

where $P(C_n)$ is calculated by dividing the area of C_n by the total area of all components in $\boldsymbol{A}^{\ell,h}$. This provides a normalized measure of spatial focus. The resulting spatial entropy $H(\boldsymbol{A}^{\ell,h})$ ranges from 0 to 1. A value of 0 indicates that attention is completely focused on a single region, while a value of 1 suggests that attention is evenly distributed across the image. This measure thus enables us to evaluate the dispersion of the model's attention across different regions within the image.

B.2. Details of Our Framework

In this section, we provide a detailed description of our framework, described in Sec. 5 of the main paper.

Binarization of the Attention Map. The attention map is binarized by setting values above the mean to 1. This approach effectively highlights the most significant regions of the attention map.

Gaussian Smoothing. Gaussian smoothing is applied using a kernel size of $k = 7$ and a standard deviation of $\sigma = 1.0$. These parameters ensure a balance between smoothing effects and detail preservation.

Convex Hull Algorithm for Bounding Box. To determine the bounding box in an assembled attention map from the localization heads, we employ the convex hull algorithm [13]. In cases where multiple convex hulls are present within the same attention map, we retain only the largest convex hull. Subsequently, we calculate the smallest tight bounding box that encloses the retained convex hull and we use it as the final bounding box.

C. More Analysis on Localization Heads

C.1. Extended Analysis Across More LVLMs

In this section, we extend the analysis of localization heads in Sec. 4 of the main paper to more LVLMs, including InternVL [8], LLaVA [40], Mini-Gemini [34], ShareGPT4V [7], and Yi-VL [73].

Average Attention Sum in More LVLMs. We extend Fig. 3 in the main paper to demonstrate that relatively few attention heads significantly contribute to the model's text-image interaction. As shown in Fig. 11, the trend of the average $S_{\text{img}}^{\ell,h}$ values remains consistent across different LVLMs.

Selection Frequency and IoU in More LVLMs. Similar to the above, we extend Fig. 6 in the main paper to cover additional LVLMs. Fig. 12 presents the selection frequency and a scatter plot of selection frequency rank versus IoU for each attention head across various LVLMs. The results confirm that our observations hold consistently across different LVLMs.

C.2. Robustness of Localization Head Selection

In this section, we validate the robustness of our localization head selection method across different threshold values (τ) and the number of selected heads (N). The experiments below indicate that localization head selection is not sensitive to the choice of τ or N.

Threshold τ. Fig. 3 in the main paper presents the average S_{img} values for each attention head, setting the threshold τ at the point where the maximum curvature is observed. We select maximum curvature as the threshold to reduce the need for manual tuning; however, other τ values can also be considered. Therefore, we further validate that plausible τ values can give consistent results with the maximum curvature. To this end, we calculate the selection frequency of the heads based on different τ values and compare them with the results obtained using the maximum curvature. The results are presented in Fig. 13. We observe that the same localization heads are consistently selected across different τ values, indicating that our analysis results are robust to the choice of τ.

Number of Heads N. In Fig. 6(a) of the main paper, we select the 10 heads with the lowest $H(\boldsymbol{A}^{\ell,h})$ values and repeat the process for 1,000 image-text pairs to calculate the selection frequency. We also investigate the effect of selecting different numbers of heads (N) on the selection frequency. We conduct experiments from $N = 1$ to $N = 14$ and compare the results with the selection frequency obtained using $N = 10$ (default setting). As shown in Fig. 14, we can obtain the same top-3 localization heads consistently across different N values, suggesting that the selection of localization heads is robust to the choice of N.

D. More Experiments

D.1. Comparison with Baseline Models

Most LVLMs, including the LLaVA [40] family, likely encode localization knowledge in their pretrained weights, possibly due to pretraining with bounding box coordinates or visual instruction prompts [40]. In Tab. 6, we compare baseline models and our proposed method, revealing the

Table 6. Comparison performance to baseline models

REC (RefCOCOg)	DeepSeekVL-1.3B	LLaVA-1.5-7B	LLaVA-1.5-13B
Baseline	1.5	2.92	5.28
Ours	65.2	82.3	84.3

Table 7. Performance comparison between F-LMM [68] and our method on the RES task. We note that F-LMM models are trained on the training set of Referring Expression Segmentation datasets.

Method	RefCOCO			RefCOCO+			RefCOCOg	
	val	testA	testB	val	testA	testB	val	test
F-LMM (Fine-tuning on RES)								
DeepSeek-VL-1.3B	75.0	78.1	69.5	62.8	70.8	56.3	68.2	68.5
Mini-Gemini-2B	75.0	78.6	69.3	63.7	71.4	53.3	67.3	67.4
DeepSeek-VL-7B	76.1	78.8	72.0	66.4	73.2	57.6	70.1	70.4
LLaVA-1.5-7B	75.2	79.1	71.9	63.7	71.8	54.7	67.1	68.1
Ours (Training-free)								
DeepSeek-VL-1.3B	56.3	57.0	52.7	51.2	55.5	49.2	52.3	55.8
Mini-Gemini-2B	59.8	60.3	55.5	56.3	59.9	51.8	55.1	60.3
DeepSeek-VL-7B	73.9	76.6	70.7	63.1	67.1	56.5	64.0	68.9
LLaVA-1.5-7B	74.2	76.5	70.4	62.5	65.2	56.0	64.2	68.1

baseline models' poor localization accuracy, likely due to their focus on describing objects rather than precise localization. Moreover, the localization head might provide only indirect support when text generation unfolds in its usual course. As a result, it becomes difficult for the model to directly output accurate object or region coordinates required for visual grounding, unless the information from this head is explicitly scrutinized. Thus, the localization head's practical value can be realized as long as it is integrated with our proposed method.

D.2. Comparision with F-LMM

We compare our method with F-LMM [68], which also leverages the attention weights of frozen LVLMs for visual grounding. The differences between F-LMM and our method are as follows. First, F-LMM still requires fine-tuning its mask decoder modules on visual grounding datasets (i.e., referring expression segmentation datasets). Second, F-LMM uses all attention heads without considering the relative importance of each, leaving the decoder modules to interpret the entire set of attention weights. In contrast, our approach requires no fine-tuning and directly utilizes a few selected attention heads that are particularly useful for localizing objects in the image. Furthermore, our framework provides a transparent understanding of where the model focuses through localization heads, which is not available in F-LMM.

Tab. 7 presents the performance comparison between F-LMM and our method on the RES task. In smaller LVLMs (*e.g.*, DeepSeek-VL-1.3B [42] and Mini-Gemini-2B [34]), F-LMM outperforms our method. However, in relatively larger LVLMs (e.g., DeepSeek-VL-7B [42] and LLaVA-1.5-7B [39]), our method demonstrates performance comparable to F-LMM, with only a slight gap. This result suggests that the localization heads have competitive potential with the specialized mask decoder modules for visual grounding tasks, especially in relatively larger LVLMs.

Table 8. Ablation study on Gaussian smoothing parameters (σ and κ). The performance is evaluated using the RefCOCO validation set (UNC split) with the LLaVA-1.5-13B [39].

Task	σ (standard deviation)					
	0.0	0.4	0.8	1.0	1.4	1.8
REC	85.5	86.8	87.2	87.2	86.8	84.3
RES	74.3	75.2	76.1	76.1	75.2	72.7
Task	κ (kernel size)					
	1	3	5	7	9	11
REC	85.5	86.5	86.5	87.2	87.2	87.2
RES	74.3	75.2	75.2	76.1	76.1	76.1

Table 9. Performance comparison with F-LMM [68] on the PNG [12] benchmark.

PNG (all)	Ours	F-LMM
DeepSeekVL-7B	66.7	65.7

D.3. Gaussian Smoothing Ablation

When assembling the attention map in the localization head (see Sec. 5 of the main paper), we apply Gaussian smoothing to the attention map to minimize potential random noise. In this section, we conduct an ablation study on the parameters of Gaussian smoothing to better understand the robustness of our framework across different values of standard deviation σ and kernel size κ. For the experiments, LLaVA-1.5-13B [39] was evaluated using the RefCOCO validation set (UNC split).

The results are presented in Tab. 8. Regardless of the selected σ and κ, Gaussian smoothing consistently enhances performance in almost all cases. The findings highlight that the framework is robust to varying choices of σ and κ. Furthermore, even when using the basic attention map of localization heads without Gaussian smoothing ($\sigma = 0$ or $\kappa = 1$), the performance remains competitive, with only a 1.9% drop compared to the best case. This demonstrates that Gaussian smoothing only serves as an auxiliary post-processing step for refining the attention map from localization heads.

D.4. Multi-Object Grounding Tasks

Beyond single-object tasks, our pipeline also suggests promise for multi-object grounding. We utilize spaCy [10] to extract noun tokens for generating attention maps (see Fig. 10), obtaining comparable results on the PNG benchmark [12], with improvements observed relative to F-LMM (see Tab. 9). Similarly, we believe this approach holds promise for extension to other various tasks [17, 38, 49].

E. More Qualitative Results

We present more qualitative results of our framework, including the performance of 10 LVLMs [7, 8, 34, 39, 40, 42, 73], with parameter numbers ranging from 1.3B to 13B, on visual grounding tasks. Fig. 15, Fig. 16, and Fig. 17 present the qualitative results of our method on the Referring Expression Comprehension (REC), Referring Expression Segmentation (RES), and Reasoning Segmentation tasks, respectively. The results demonstrate that only a few selected localization heads are sufficient to accurately localize objects in the image based on the text query. Our method effectively localizes objects in various scenarios.

F. Applications

F.1. Real World Application

Fig. 18 illustrates that the localization heads effectively capture the region or object of interest in images from the real world, based on the provided expressions. This result demonstrates the robustness of the localization heads across various types of data.

F.2. Image Editing

Fig. 19 presents the results of image inpainting performed by integrating Stable Diffusion XL (SDXL) [47]. The frozen LVLM generates a segmentation mask corresponding to the expression, and this mask, along with an additional text prompt, is used as input to the diffusion model to generate the desired image. These results demonstrate that the segmentation mask corresponding to the referred text, output by a small number of localization heads from the frozen LVLM, can serve as guidance for diffusion models. This compatibility enables its application in image editing tasks.

G. Limitations

We propose a simple yet effective framework for training-free visual grounding, which leverages the localization heads of LVLMs. Our framework successfully localizes objects in images based on text queries without requiring any fine-tuning and achieves superior performance compared to existing training-free methods. However, our method still has some limitations that could be addressed in future work.

First, our work, as illustrated in Fig. 10, reveals the potential for multi-object grounding; however, the establishment of a formalized pipeline or the development of a more streamlined implementation remains limited. The task of rendering the identified localization head more practical, user-friendly, and adaptable across a diverse range of applications continues to pose a significant challenge. This presents a compelling avenue for future research.

Figure 10. Multi-object segmentation results from the localization heads of DeepSeekVL-7B, along with the corresponding raw attention maps.

Second, our method is less suitable for LVLMs or methods that do not preserve spatial information in images (*e.g.*, pooling) [1, 20, 31, 32, 59]. These methods make it challenging to explicitly obtain image attention maps. To collect the attention map, a reverse computation is required to determine the order in which image tokens were input during processing. We leave the application of our framework to these methods for future exploration.

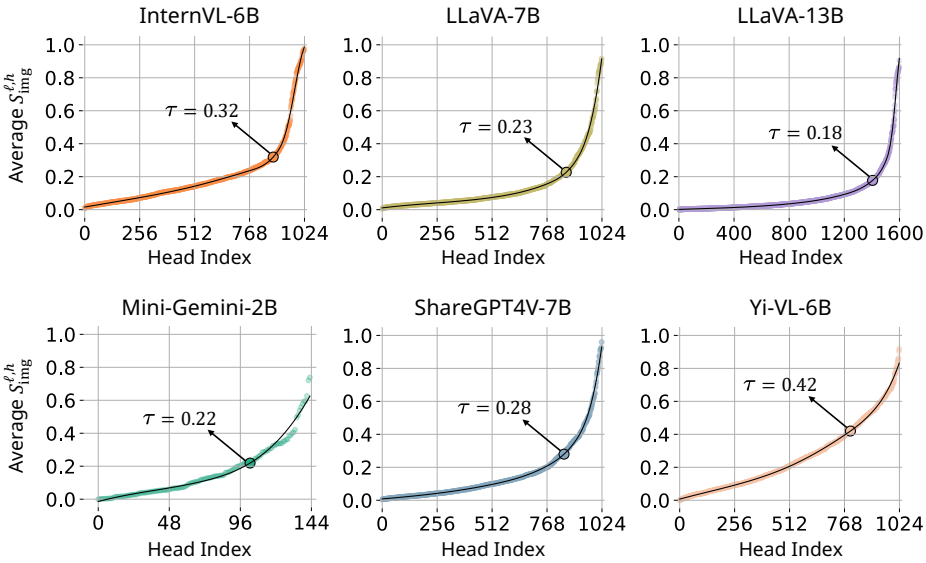


Figure 11. Average $S_{\text{img}}^{\ell,h}$ values for each attention head in more LVLMs. τ is set at the point where the maximum curvature is observed.

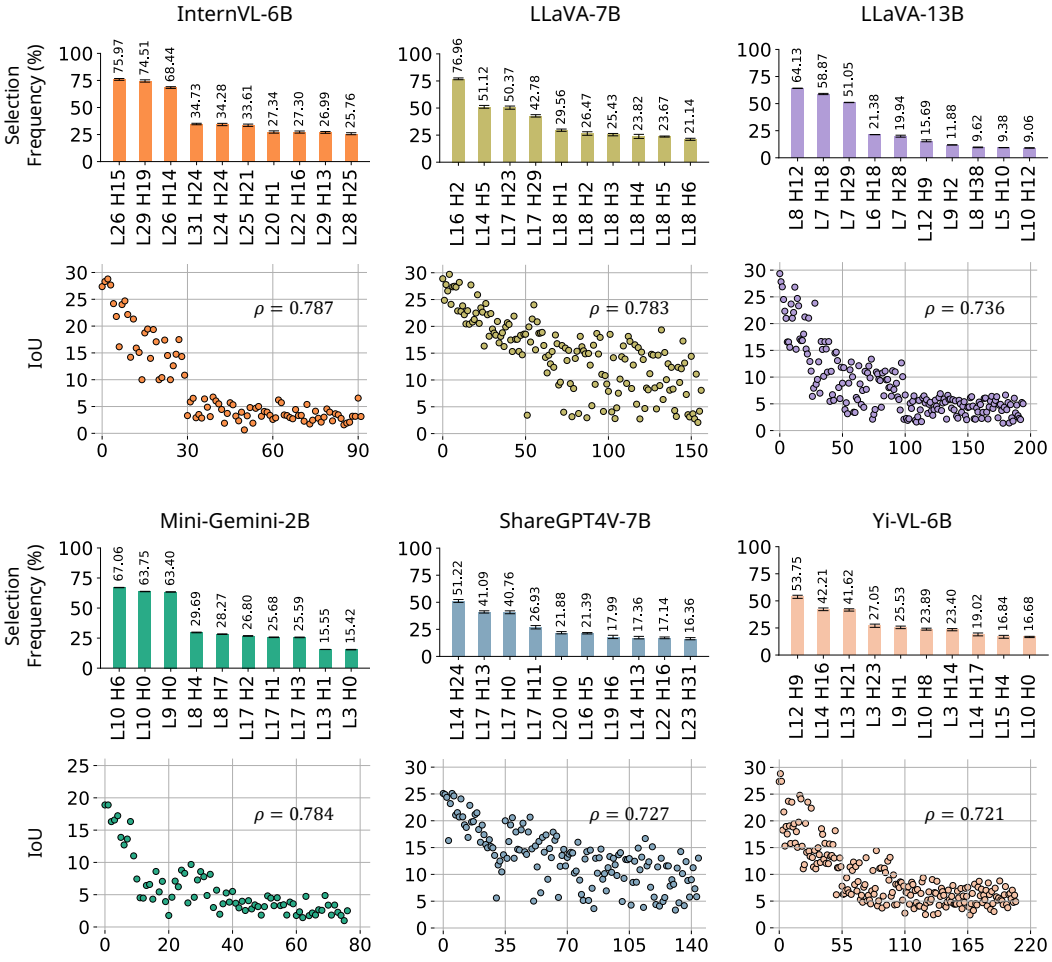


Figure 12. Selection frequency of individual heads and scatter plot of selection frequency rank versus each head's average IoU in more LVLMs. The Spearman correlation coefficient (ρ) between the selection frequency rank and the average IoU is displayed in the top-right corner of each plot. The observed Spearman correlation are statistically significant ($p < 0.001$) for all LVLMs.

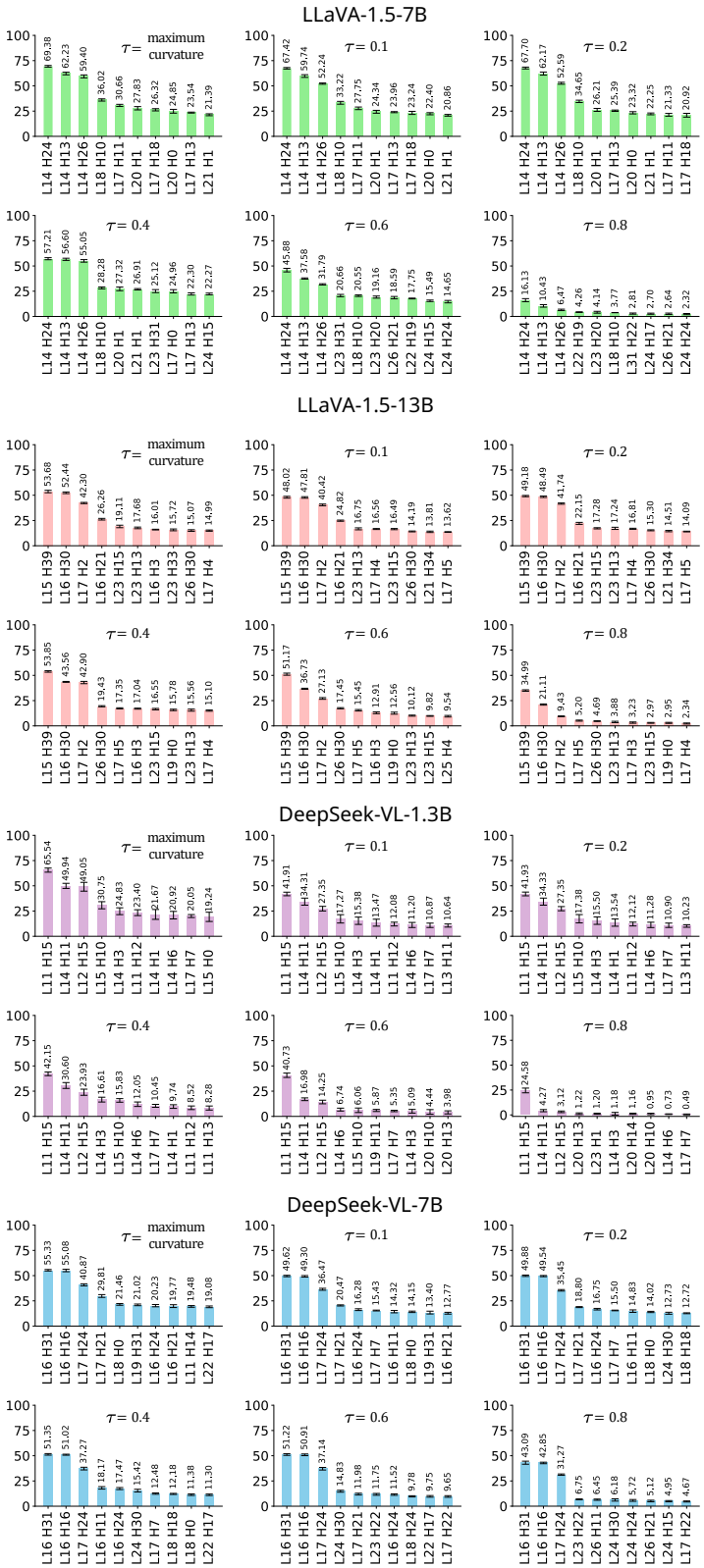


Figure 13. Selection frequency of individual heads across different τ values. τ represents the threshold for the sum of each head's attention map. Our analysis focuses on heads with attention map sums greater than τ, which are selected as targets for selection frequency evaluation. In the main paper, we select the threshold where the maximum curvature is observed. The top-3 localization heads remain consistent across different τ values, demonstrating the robustness of our analysis to variations in τ.

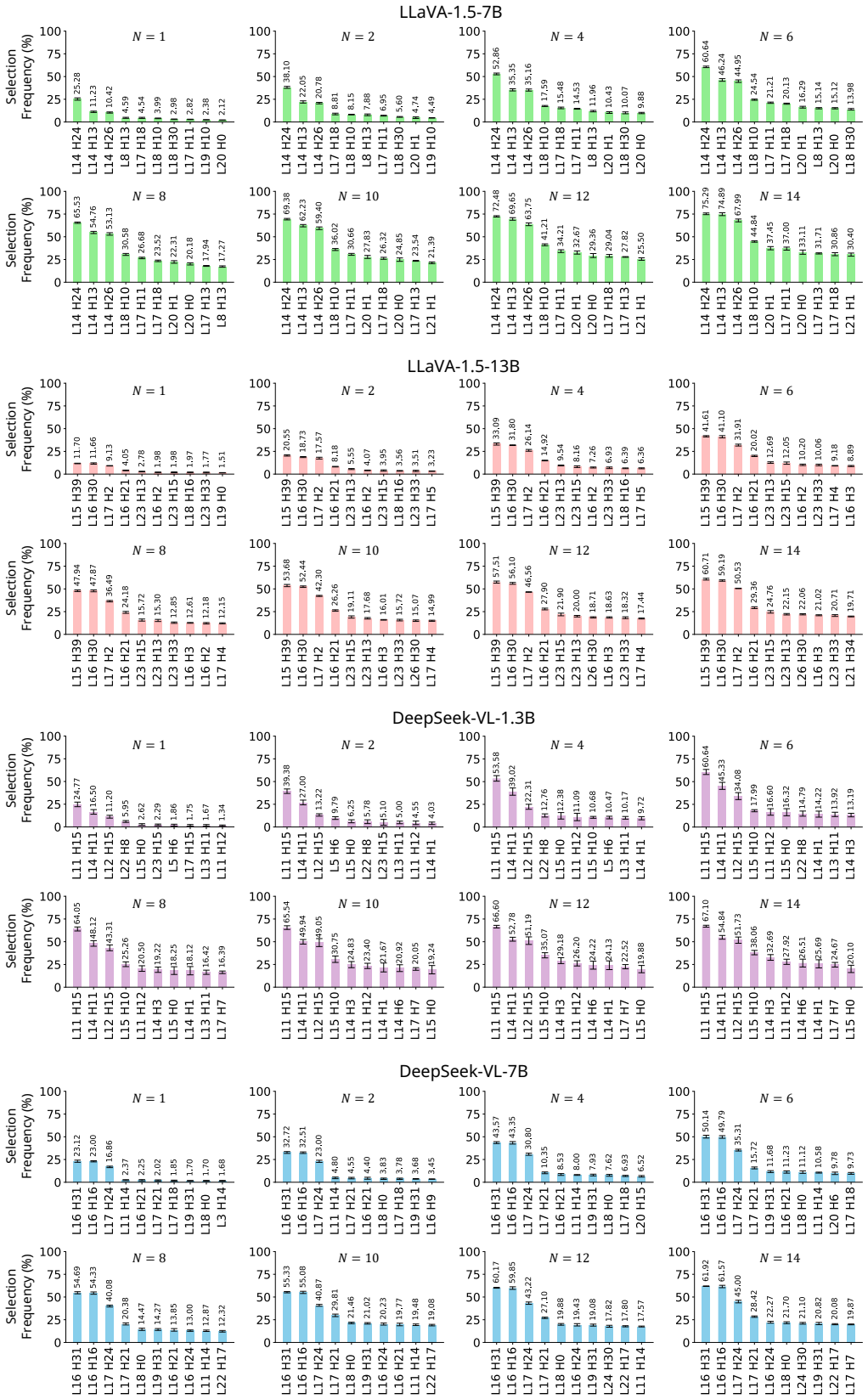
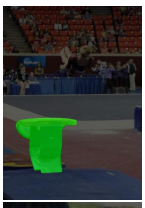


Figure 14. Selection frequency of individual heads across different N values. N refers to the number of selected heads based on the lowest $H(\boldsymbol{A}^{\ell,h})$ values. Default setting is $N = 10$. The top-3 localization heads are consistent across different N values, indicating the robustness of localization head selection to the choice of N.

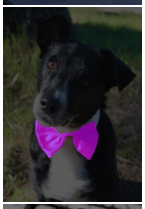
Figure 15. Qualitative results of Referring Expression Comprehension.

Figure 16. Qualitative results of Referring Expression Segmentation.

LLaVA-1.5-13B with only three attention heads (L15 H39, L16 H30, L7 H2)



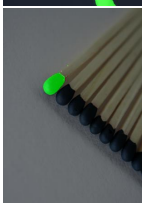
Expression:
In gymnastics competitions, athletes perform a variety of acrobatic movements using different apparatus. Regarding the picture, what equipment could be utilized by athletes to accomplish challenging and impressive movements such as flips and vaults?



Expression:
What item in the picture could be utilized as the accessory that people typically wear around their neck for elegant formal attire?

Expression:
Something used for playing music.

Expression:
During an air show, pilots perform various aerial maneuvers to entertain the audience. What spectacle in the picture indicates that a pilot is performing a spectacular and dazzling aerial stunt?



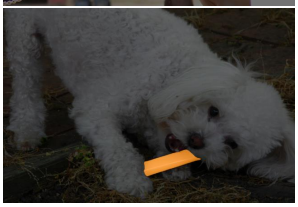
Expression:
The frictional part used for igniting.

Expression:
If we were attending a car show and desired to sit down and observe the displayed cars, where might we locate a suitable place to relax?

Expression:
In modern cuisine, food decoration plays an important role in enhancing the dining experience. Based on the picture, which food item is most likely to be used for decoration purposes?

Expression:
What is the item in this picture that could be utilized to hit the ball during a game of tennis?

Expression:
In horse riding, it is crucial to have control and direction over the horse. What object in the picture is typically used for guiding and controlling the movements of a horse?



Expression:
Please identify which object in the picture could serve as a toy for a dog, as dogs relish playing with a variety of different toys that are specifically designed for biting and chewing.

Figure 17. Qualitative results of Reasoning Segmentation.

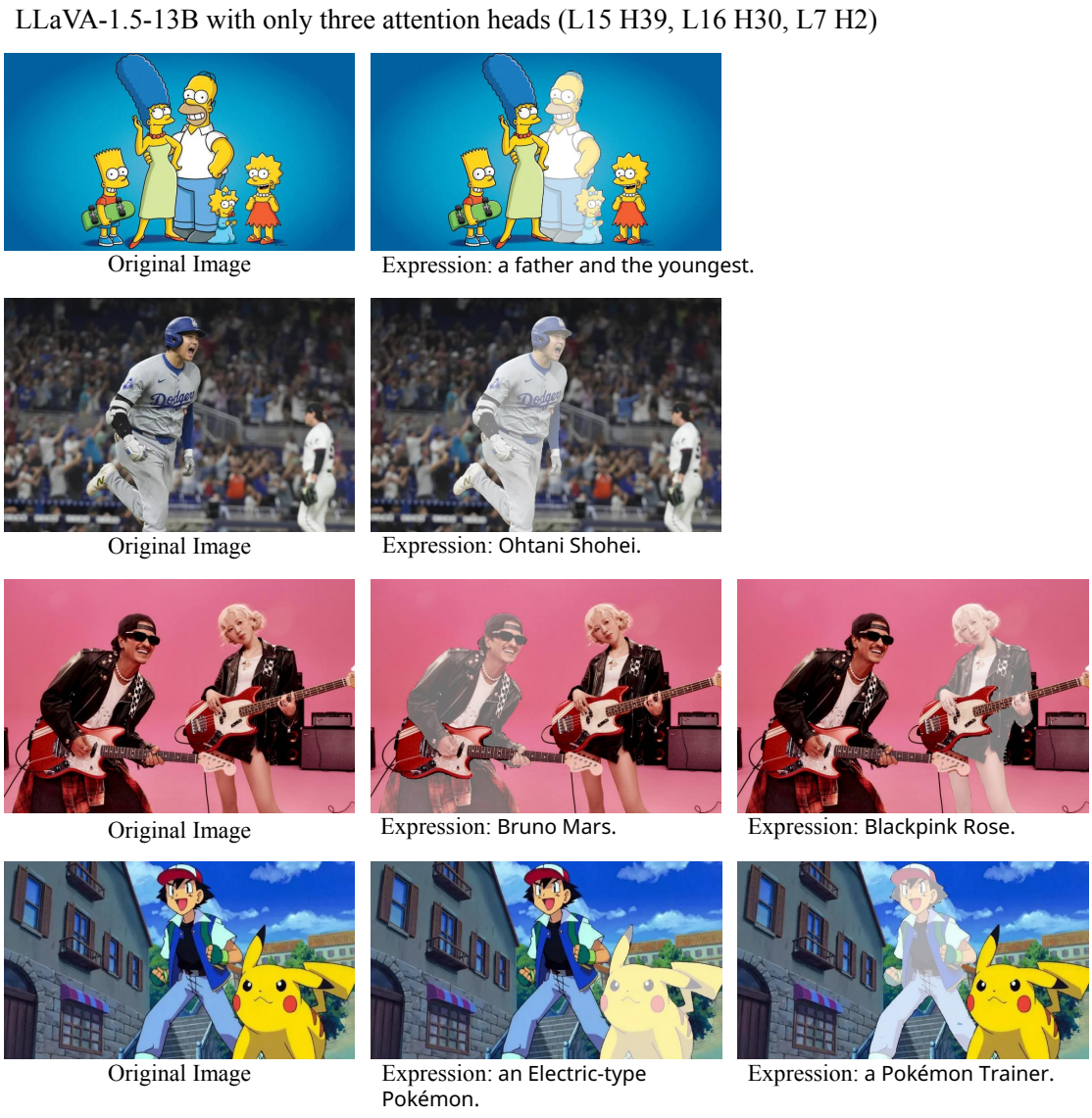


Figure 18. Qualitative results of real-world image segmentation. LLaVA-1.5-13B [39] uses only three attention heads (L15 H39, L16 H30, L7 H2) as localization heads to produce a precise segmentation masks related to the text expressions. The whitened regions in the images represent the segmentation mask output by the model.

Figure 19. Qualitative results of generating the desired image through integration with a diffusion model [47]. Given an original image, our method generates a mask from the LVLM based on the text describing the desired modifications. This mask is then used as guidance for a diffusion model to perform image editing. Using the segmentation mask obtained through the localization head of the frozen LVLM [39], it is possible to generate semantic objects that align with the prompt at the specified mask locations.