Mitigating Cross-Image Information Leakage in Multi-Image Understanding with Large Vision-Language Models

Yeji Park[1] **Minyoung Lee**[1] **Sanghyuk Chun**[2*] **Junsuk Choe**[1†]
[1]Sogang University [2]Princeton University

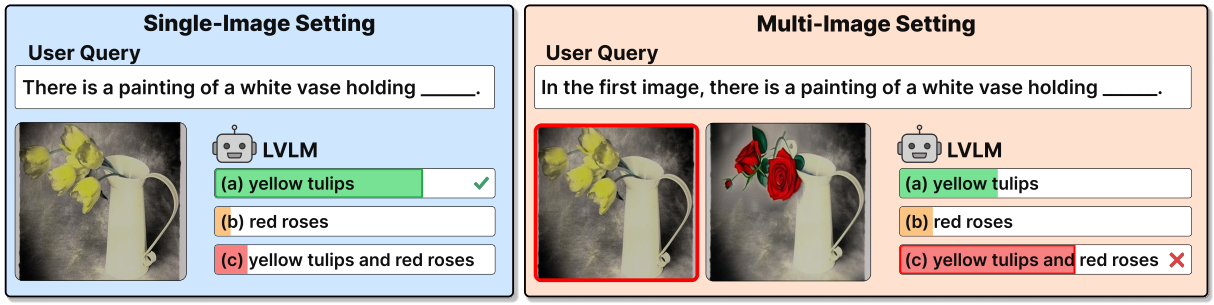


Figure 1: Illustration of *cross-image information leakage* in LVLMs in single-image vs. multi-image settings. Under the multi-image setting, the model entangles visual features across images. As a result, the model incorrectly selects the option that combines objects from both images, rather than the object present only in the target image.

Abstract

Large Vision-Language Models (LVLMs) exhibit strong performance on single-image tasks. However, their performance degrades significantly when handling multi-image inputs. While this degradation has been observed in prior work, its nature remains poorly understood. We empirically observe visual elements from different images become entangled in the model's representations and responses. We refer to this phenomenon as *cross-image information leakage*. To address this issue, we propose **FOCUS**, a training-free and architecture-agnostic method. FOCUS masks all but one image with random noise, guiding the model to focus on the single clean image. This process is applied across the target images to obtain logits under partially masked contexts. These logits are aggregated and then refined using a noise-only reference input, which suppresses the leakage and yields more accurate outputs. FOCUS consistently improves performance on diverse multi-image benchmarks. We further show that FOCUS generalizes to video understanding, extending its applicability beyond static multi-image inputs. This demonstrates that FOCUS offers a general solution for enhancing multi-image reasoning without additional training or architectural modifications. Source code is available at https://github.com/yejipark-m/FOCUS.

*Works done at NAVER AI Lab.
†Corresponding author.

1 Introduction

Large Vision-Language Models (LVLMs) are designed to jointly understand visual and textual information (Chen et al., 2024b; Achiam et al., 2023; Bai et al., 2025b). They excel at vision-language (VL) tasks (Li et al., 2024a; Daxberger et al., 2025), including Visual Question Answering (VQA) (Antol et al., 2015) and Image Captioning (Herdade et al., 2019).

Although LVLMs have shown strong performance on VL tasks with a single image, real-world use cases often go beyond this single-image assumption. Users naturally provide multiple images at once and expect models to integrate information across all of them, such as comparing scenes and resolving cross-image references to produce a coherent response (Jiang et al., 2024). To this end, several studies train models with interleaved image-text sequences to explicitly enhance reasoning across multiple images (Awadalla et al., 2023; Lin et al., 2024; Sun et al., 2024; Jiang et al., 2024).

However, recent studies have revealed that LVLMs still struggle in such settings, showing a notable performance drop when multiple images are provided (Wang et al., 2025b; Zhang et al., 2025). This suggests that the underlying cause of this degradation cannot be fully attributed to insufficient training data. Yet, why this performance degradation occurs remains largely unexplored.

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 645–663
November 6-10, 2026 ©2026 Association for Computational Linguistics

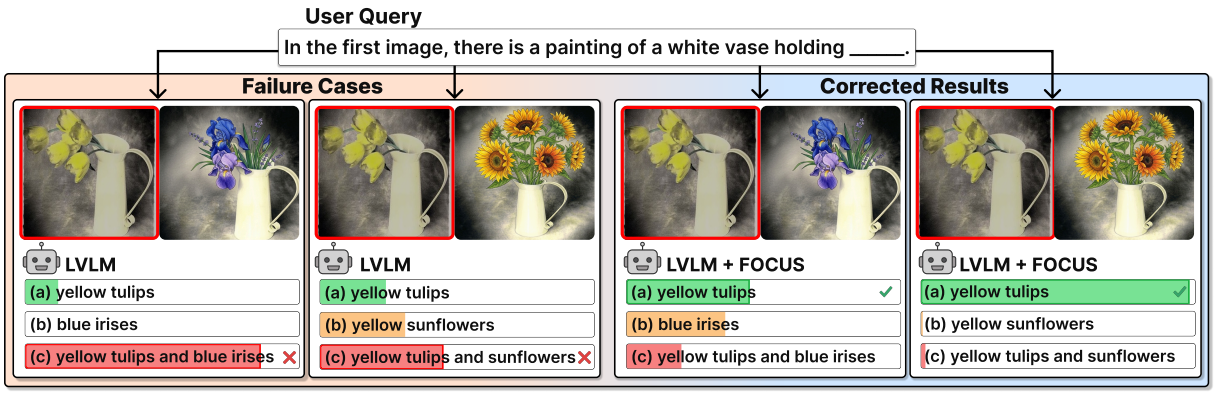


Figure 2: Illustration of cross-image information leakage across different image pairs. Given the same target image (red border), the baseline LVLM consistently selects the incorrect merged caption that combines content from both images (Failure Cases), regardless of which non-target image is present. FOCUS mitigates this confusion and recovers the correct answer in both cases (Corrected Results).

We observe that LVLMs tend to exhibit a phenomenon we term *cross-image information leakage*, where models mix visual cues across inputs rather than interpreting each image independently. As illustrated in Figure 1, a model answers correctly with a single image, but fails when two images are provided simultaneously. Specifically, in the multi-image setting, the probability of "*yellow tulips*" (the correct option) decreases. Instead, the model mistakenly selects "*yellow tulips and red roses*". This answer merges visual content from two separate images into a single, incorrect response. This suggests that current LVLMs cannot reliably isolate and reason over visual information from multiple images independently.

Moreover, this failure is not specific to a particular image pair. In Figure 2, when the non-target image changes from blue irises to yellow sunflowers, the model predicts its incorrect answer from "*yellow tulips and blue irises*" to "*yellow tulips and sunflowers*", consistently incorporating content from whichever non-target image is present.

Our analysis shows that cross-image information leakage is closely related to multi-image understanding degradation. (§2.1). Under multi-image inputs, models show an increased tendency to select captions that merge content from both the target and non-target images, leading to substantial performance degradation. To mitigate this, one naïve solution is to process each image independently, given that models perform well on single images. Yet this fails to capture inter-image relationships that are essential for multi-image reasoning (§2.2).

To this end, we propose FOCUS, a training-free method that mitigates cross-image information leakage while preserving inter-image relationships. Our key idea is to leverage the observation that LVLMs reason well over a single image. We process all images together to preserve inter-image context, while guiding the model to focus on one image at a time. Specifically, we mask all but one image with random noise, prompting the model to concentrate on the single clean image (§3). FOCUS applies this masking to each image in turn to extract the focused logits, and aggregates the resulting output logits while preserving the positional context of the images.

We validate FOCUS using diverse LVLM families across five multi-image benchmarks: Winoground (Thrush et al., 2022), VisMin (Awal et al., 2024), Mantis-Eval (Jiang et al., 2024), MuirBench (Wang et al., 2025b) and MIRB (Zhao et al., 2024). FOCUS achieves consistent improvements, with the best gains of up to +18.8 Image and +16.8 Group score on Winoground, +44.37 Image and +41.9 Group score on VisMin, +5.5%pts accuracy on Mantis-Eval, +3.2%pts on MuirBench, and +1.7%pts on MIRB. Beyond static multi-image settings, FOCUS further generalizes to video understanding (Zhang et al., 2024). All gains are achieved without any additional training or architectural changes, highlighting the effectiveness and generalization ability of our method.

2 Motivation

2.1 Cross-image Information Leakage

We observe that the model exhibits a notable limitation when processing multiple images simultaneously, namely *cross-image information leakage*.

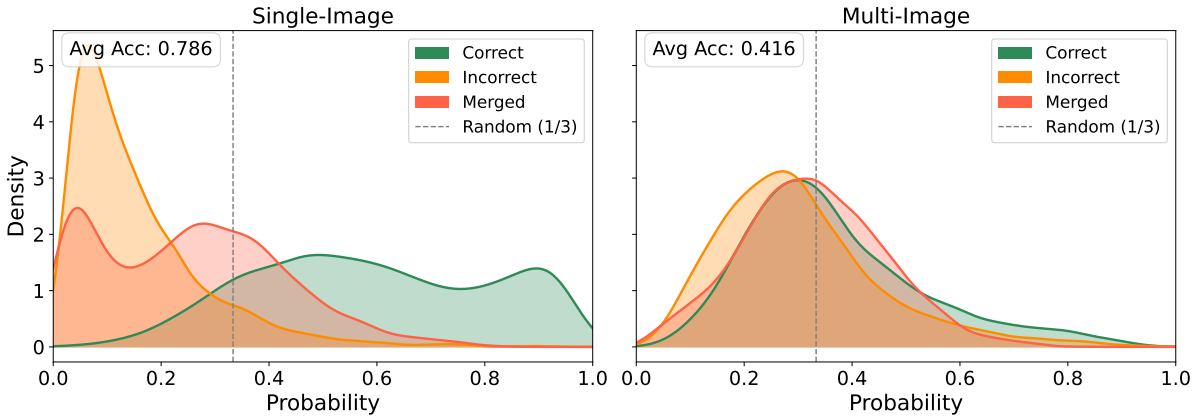


Figure 3: Estimation of probability density function of Qwen2.5-VL 3B predictions over three answer options. We evaluated on 500 image pairs from VisMin (Awal et al., 2024). In the single-image setting, the correct option is favored. In the multi-image setting, all distributions located their center near random chance (the dotted line).

This issue is closely related to how visual inputs are embedded and processed within an LVLM. Most LVLMs handle multiple images as a sequence of visual tokens, which are concatenated with text tokens and processed through the attention mechanism. During this process, the tokens interact strongly with each other due to *inter-image causal attention* (Tian et al., 2025). This can cause the attention mechanism to mix latent representations across images and the model entangles visual semantics that should remain distinct.

We empirically analyze this using 500 image pairs from Awal et al. (2024). For each pair, we provide three options: a correct caption for the target image, an incorrect caption from the other image, and a merged caption combining both (see Figure 1, options (a)-(c)). We estimate the probability density function of model predictions over these three options using kernel density estimation.

As in Figure 3, in the single-image setting, the correct caption distribution is skewed toward higher probabilities, while the incorrect and merged caption distributions concentrate at lower probabilities. However, in the multi-image setting, the probability of selecting the merged caption increases. For example, in Qwen2.5-VL-3B model, all three distributions center near random chance and become nearly indistinguishable from one another.

These results suggest that the model loses its ability to selectively attend to the target image, making accurate multi-image reasoning substantially less reliable. We further verify that this phenomenon persists across larger model scales, including Qwen2.5-VL-7B and Qwen3-VL-8B, though the degree of degradation vary with model capacity (provided in Appendix §D.1).

2.2 Limitation of Independent Inferences

A straightforward approach to mitigate cross-image information leakage is to perform inference on each image independently and then aggregate the results. However, even an optimal aggregation scheme cannot capture *inter-image relationships*, which are crucial for multi-image reasoning. Instructions such as "*Compare the second image with the first one*" or "*Describe what is common across all images*" inherently require reasoning about relationships between images, and such cross-referential queries are frequent in multi-image tasks.

One natural extension of this approach is to encode inter-image relationships explicitly through textual prompts. For example, one can generate an independent caption for each image and then concatenate these descriptions alongside positional information (*e.g.*, "The first image shows ..., the second image shows ...") as context for a final inference step. However, we experiment with this approach and find it to be ineffective. This limitation is further analyzed in §4.5.

Observations

- Cross-image information leakage is closely related to multi-image understanding degradation, where models increasingly confuse visual content across images and produce incorrect responses.
- Processing each image independently fails to resolve this issue, as it sacrifices inter-image relationships essential for multi-image reasoning.

Motivated by these observations, we aim to design a method that (i) jointly captures the structure and relationships among multiple images within a single, unified reasoning process, and (ii) mitigates cross-image information leakage. We address this challenge with FOCUS, described next.

3 Method

Our method is a training-free approach that enables LVLMs to concentrate on one image at a time while preserving the positional structure of multi-image inputs. Our method consists of three steps: (a) visual masking, (b) image-wise focused inference, and (c) contrastive aggregation. A schematic overview of the proposed framework is shown in Figure 4.

Visual Masking (Figure 4a). First, we prepare masked image inputs via noise injection with scale λ. Let N be the number of input images and $\phi_v(\cdot)$ denote the visual encoder that maps an image to its token embeddings. For each target image I_i, we corrupt all other images I_j ($j \neq i$) into $\tilde{I}_j$, leaving I_i clean. This produces N partially masked inputs and one fully masked reference:

$$\begin{aligned} \mathcal{I}_i &= [v'_1, \ldots, v_i, \ldots, v'_N], \\ \mathcal{I}_{\text{noise}} &= [v'_1, v'_2, \ldots, v'_N], \end{aligned} \tag{1}$$

where $v_i = \phi_v(I_i)$ and $v'_i = \phi_v(\tilde{I}_i)$ denote the clean and noise-corrupted visual embeddings, respectively. Each $\mathcal{I}_i$ allows the model to focus solely on I_i, and $\mathcal{I}_{\text{noise}}$ serves as a noise-only reference.

Image-wise Focused Inference (Figure 4b). For each partially masked image input I_i, only the i-th image remains unmasked. The original image order is preserved to retain positional semantics. We compute the logit distribution ℓ_i for each i, where f denotes the model, $\mathbf{X}$ denotes the text input:

$$\ell_i = f(\mathcal{I}_i, \mathbf{X}). \tag{2}$$

Figure 4(b) depicts the different logit distributions ℓ_i for $i = 1, \ldots, N$ obtained via focused inference. For example, blue distribution (left) corresponds to the model focusing on the first image, while the yellow one (middle) represents focus on the second image (i.e., the clean image in $\mathcal{I}_2$). We also compute a noise reference logit distribution using the fully noise-masked input:

$$\ell_{\text{noise}} = f(\mathcal{I}_{\text{noise}}, \mathbf{X}). \tag{3}$$

The gray logit distribution (right), which contains no information from either the first or the second image, serves as a reference for isolating irrelevant visual content during contrastive aggregation.

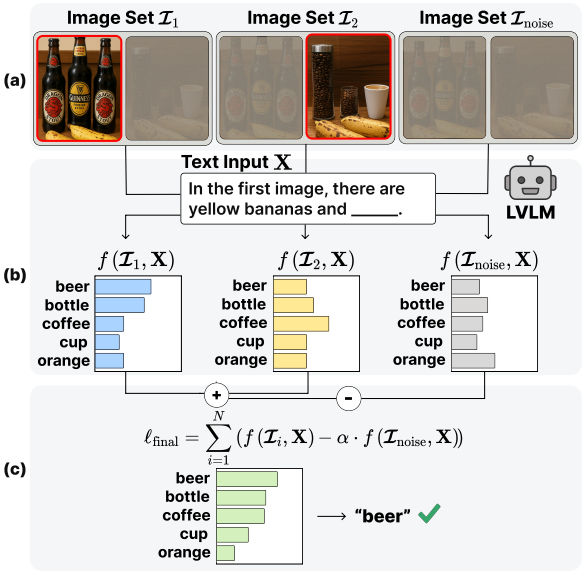


Figure 4: Overview of FOCUS. It consists of three main steps: (a) visual masking, (b) image-wise focused inference, and (c) contrastive aggregation.

Contrastive Aggregation (Figure 4c). Each ℓ_i contains useful signals from the clean image I_i, but also includes residual side-effects. For instance, in Figure 4(b), irrelevant token like *orange* has nonzero probability by the noise masks. To suppress these residuals, we subtract the noise reference logit ℓ_{noise} from ℓ_i and aggregate the results:

$$\ell_{\text{final}} = \sum_{i=1}^{N} (\ell_i - \alpha \cdot \ell_{\text{noise}}), \tag{4}$$

where α is a hyperparameter tuned via validation to appropriately weight the subtraction. The final output token is then sampled from the aggregated logits ℓ_{final}. Figure 4(c) illustrates this step: the final green logit distribution, computed using Equation (4), correctly identifies *beer* as the answer.

The proposed method effectively suppresses cross-image information leakage while preserving positional structure, without requiring any model training or architecture modification. It enables LVLMs to perform multi-image reasoning tasks better in a training-free, generalizable manner.

4 Experiments

4.1 Benchmarks for Multi-image Tasks

Multi-image understanding has been a fundamental challenge in visual reasoning (Suhr et al., 2017, 2018). We employ five evaluation benchmarks:

Model Size	<7B			≥7B		
Family \ Score	T	I	G	T	I	G
InternVL3	47.25	6.25	4.50	**70.75**	40.75	34.25
+FOCUS	47.25	**27.25**	**19.75**	69.25	**42.25**	**35.75**
Qwen2.5-VL	56.25	**36.75**	26.00	74.50	39.75	34.00
+FOCUS	56.25	36.50	26.00	74.50	**58.50**	**50.75**
LLaVA-OV	3.25	**20.00**	0.25	76.75	36.75	33.00
+FOCUS	3.25	19.25	**0.75**	76.75	**48.50**	**42.25**

Table 1: Comparison on Winoground across model families and sizes with and without FOCUS. Metrics: T (Text), I (Image), G (Group). Bold indicates improvement over the baseline.

Model Size	<7B			≥7B		
Family \ Score	T	I	G	T	I	G
InternVL3	93.25	29.96	29.36	94.62	**84.84**	**81.60**
+FOCUS	**93.34**	**74.33**	**71.26**	**94.70**	83.05	80.83
Qwen2.5-VL	**92.14**	47.74	46.20	93.85	**92.65**	88.55
+FOCUS	92.06	**56.92**	**55.82**	93.85	92.57	**88.72**
LLaVA-OV	20.98	9.26	1.55	93.43	62.99	60.68
+FOCUS	20.98	**27.01**	**8.42**	93.43	**64.14**	**62.00**
Qwen3-VL	94.20	90.68	87.35	**94.88**	93.24	90.34
+FOCUS	**94.28**	**92.06**	**88.80**	94.79	**93.50**	**90.51**

Table 2: Comparison on VisMin-Easy across model families and sizes with and without FOCUS. Metrics: T (Text), I (Image), G (Group). Bold indicates improvement over the baseline.

Model Size	<7B			≥7B		
Family \ Score	T	I	G	T	I	G
InternVL3	**65.53**	6.84	6.17	**82.86**	47.66	44.80
+FOCUS	64.87	**26.77**	**23.96**	82.54	**52.66**	**48.89**
Qwen2.5-VL	77.38	27.09	22.93	84.37	53.26	49.97
+FOCUS	**78.20**	**28.86**	**25.29**	**84.47**	**61.44**	**56.99**
LLaVA-OV	1.70	**20.12**	0.51	**80.61**	44.42	38.40
+FOCUS	**2.20**	18.41	**0.59**	80.44	**57.89**	**49.40**
Qwen3-VL	83.62	49.62	46.78	**82.86**	56.65	52.95
+FOCUS	**84.11**	**58.31**	**54.32**	82.77	**59.13**	**55.01**

Table 3: Comparison on VisMin-Hard across model families and sizes with and without FOCUS. Metrics: T (Text), I (Image), G (Group). Bold indicates improvement over the baseline.

Winoground (Thrush et al., 2022), VisMin (Awal et al., 2024), Mantis-Eval (Jiang et al., 2024), MuirBench (Wang et al., 2025b) and MIRB (Zhao et al., 2024). We additionally evaluate FOCUS in video understanding and long-context settings using Vinoground (Zhang et al., 2024) and MMLongBench (Wang et al., 2025c), respectively. More details and evaluation setup for each benchmark are provided in Appendix §B and §C.

4.2 Quantitative Results

We evaluate the generalizability of the proposed method on three representative LVLM families: InternVL3 (2B, 8B) (Chen et al., 2024b), Qwen2.5-VL (3B, 7B) (Bai et al., 2025b), and LLaVA-OneVision (0.5B, 7B) (Li et al., 2024a). All hyperparameters are selected on the validation split. We report test split performance on each benchmark. To further validate generalizability across recent architectures, we additionally evaluate on Qwen3-VL (4B, 8B) (Bai et al., 2025a). Appendix §D.2 extends this to larger models, including the MoE variant Qwen3-VL-30B-A3B and Qwen3-VL-32B.

Results on Winoground. Table 1 presents the results on Winoground across three model families and scales, comparing baseline with the proposed FOCUS method. FOCUS consistently improves the Image Score (I) and Group Score (G). In InternVL3-2B, FOCUS increases the Image Score from 6.25 to 27.25. Notably, Qwen2.5-VL-7B achieves a substantial Image Score gain from 39.75 to 58.50. Overall, FOCUS tends to be effective across model families and scales, though the degree of improvement varies.

Results on VisMin-Easy. In Table 2, FOCUS shows substantial improvements across model families. The gains are particularly pronounced in smaller models. InternVL3-2B improves its Image Score from 29.96 to 74.33 and Group Score from 29.36 to 71.26. For larger models, improvements are more modest due to stronger baselines.

Results on VisMin-Hard. In Table 3, baseline performance is comparatively lower, as these categories require more subtle cross-image discrimination. FOCUS still yields consistent improvements across model families. Qwen2.5-VL-7B improves its Image Score from 53.26 to 61.44 and Group Score from 49.97 to 56.99, while Text Score remains stable. Notably, Qwen3-VL-4B achieves an improvement of +8.69 Image Score, confirming that FOCUS remains effective even on the latest models.

Model Size	<7B			≥7B		
Family \ Dataset	Mant.	Muir.	MIRB	Mant.	Muir.	MIRB
InternVL3	49.77	28.04	31.32	64.98	31.62	41.86
+FOCUS	**52.53**	**28.58**	**32.36**	**65.44**	**31.88**	**42.48**
Qwen2.5-VL	55.30	30.38	49.27	70.05	**29.92**	50.53
+FOCUS	**58.99**	**31.31**	**50.00**	70.05	29.73	**52.30**
LLaVA-OV	36.41	21.42	**22.55**	57.14	28.73	48.02
+FOCUS	**41.94**	**24.62**	22.13	**59.91**	**29.81**	**48.85**
Avg	47.16	26.61	34.38	64.06	30.09	46.80
+FOCUS	**51.15**	**28.17**	**34.83**	**65.13**	**30.47**	**47.88**

Table 4: Mantis-Eval, MuirBench and MIRB accuracy across model families and sizes, with and without FOCUS. Bold indicates improvement over the baseline.

Finding. FOCUS improves Image Score (I) and Group Score (G), the metrics that directly reflect multi-image reasoning, while leaving Text Score (T) largely unchanged.

Results on Other Benchmarks. Table 4 summarizes the performance of six model variants across three benchmarks.

On Mantis-Eval, applying FOCUS leads to consistent accuracy gains across all settings. For instance, InternVL3-2B improves by +2.76%pts, while the larger 8B model exhibits a smaller but positive effect. Similarly, Qwen2.5-VL-3B gains +3.69%pts, and LLaVA-OV-0.5B achieves a notable jump from 36.41 to 41.94. Although the absolute improvements are moderate, they are meaningful given the benchmark's difficulty, indicating that FOCUS robustly enhances performance across diverse architectures and scales.

On MuirBench, which emphasizes multi-image reasoning, FOCUS again provides improvements across nearly all model configurations: InternVL3 family shows an average gain of +0.4%pts, Qwen2.5-VL-3B achieves +0.93%pts, and LLaVA-OV family gains +2.14%pts. These results suggest that FOCUS offers incremental benefits even in complex and ambiguous multi-image settings.

Results on MIRB demonstrate similar trends. FOCUS generally improves accuracy across model families and sizes, though smaller models such as LLaVA-OV-0.5B show limited or inconsistent gains. The overall upward trend reinforces FOCUS as a model-agnostic enhancement for multi-image reasoning.

Model Size	<7B			≥7B		
Family \ Score	T	V	G	T	V	G
Qwen3-VL	32.60	17.60	6.20	34.40	35.60	18.20
+FOCUS	**33.00**	**25.80**	**11.60**	**35.60**	**37.00**	**18.60**

Table 5: Results on Vinoground with and without FOCUS. Metrics: T (Text), V (Video), G (Group). Bold indicates improvement over the baseline.

Results on Video Understanding Benchmark. To examine whether FOCUS generalizes beyond multi-image inputs to video understanding, we evaluate on the Vinoground (Zhang et al., 2024) benchmark. As shown in Table 5, FOCUS consistently improves on Qwen3-VL-4B and Qwen3-VL-8B. These results demonstrate that the benefits of image-wise focused decoding also generalize to the video domain.

Extension to Long-Context Setting. We further examine whether FOCUS generalizes to long-context settings (Wang et al., 2025c). Full results and discussion are provided in Appendix §D.3.

4.3 Ablation Study

We conduct ablation studies to assess each component of FOCUS. All experiments are conducted on the Mantis-Eval validation set using Qwen2.5-VL-7B. Unless otherwise specified, we set noise scale $\lambda = 0.3$ and weighting $\alpha = 0.4$.

Table 6 shows that removing noise injection passes all images to the model without masking, which prevents image-wise focused inference. The model cannot isolate individual image signals and suffers from cross-image information leakage, leading to a drop in accuracy from 76.19 to 71.43. When noise reference subtraction is disabled, residual artifacts introduced by noise masking remain unsuppressed during contrastive aggregation. This allows spurious visual signals to contaminate the final prediction. As a result, accuracy drops more significantly, from 76.19 to 61.90.

Both components contribute meaningfully to the final performance, confirming that noise injection and contrastive aggregation are complementary and jointly essential to FOCUS. Appendix §D.4 reports further ablations on our design choices.

4.4 Cross-Image Similarity Analysis

This analysis examines how much visual representations overlap across images in single-image

	w/o Noise (λ)	w/o Reference (α)	FOCUS
Acc.	71.43	61.90	**76.19**

Table 6: Ablation study on each component of FOCUS. We use Qwen2.5-VL-7B on Mantis-Eval validation set.

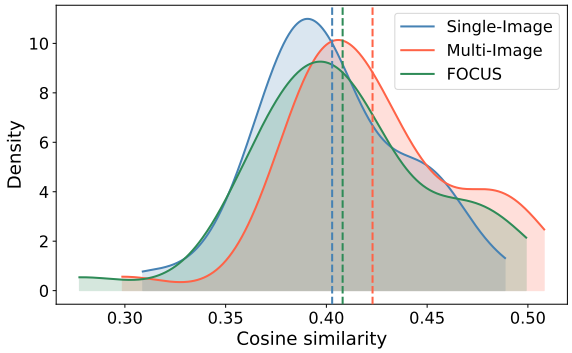


Figure 5: Average pairwise cosine similarity distribution of image token representations after the LLM decoder under three conditions. We use Qwen2.5-VL-3B on samples with 4 images each from Mantis-Eval.

versus multi-image settings, which supports the intuition behind FOCUS. We extract final hidden states at image token positions from the model across samples from Mantis-Eval. For each image, we first average hidden states over its token positions, then compute pairwise cosine similarity across all image combinations per sample. We compare three conditions:

- *Single-Image*: Each image is processed independently via a separate forward pass.
- *Multi-Image*: All images are processed jointly in a single forward pass.
- FOCUS: For each target image, noise is injected into the remaining images and processed jointly.

As shown in Figure 5, the similarity distribution shifts rightward under joint processing. This suggests that causal attention entangles image representations when multiple images are processed together. FOCUS reduces the mean similarity, nearly recovering the single-image baseline. This indicates that injecting noise into non-target images discourages the model from mixing representations across images, encouraging each image to maintain a more distinct representation during decoding. We further verify that this pattern persists across larger model scales, including Qwen2.5-VL-7B and Qwen3-VL-8B (provided in Appendix §D.1).

Method	InternVL3 2B	InternVL3 8B	Qwen2.5-VL 3B	Qwen2.5-VL 7B	LLaVA-OV 0.5B	LLaVA-OV 7B
Process Independently						
Per-Img	19.61	35.71	34.57	38.18	10.53	32.82
Process Jointly						
Baseline	31.32	41.86	49.27	50.53	**22.55**	48.02
Conv-CD	31.68	41.07	49.64	51.29	20.43	44.79
SoFA	30.96	40.66	49.43	51.39	12.59	42.87
FOCUS	**32.36**	**42.48**	**50.00**	**52.30**	22.13	**48.85**

Table 7: Comparison of FOCUS with alternative approaches on MIRB. Per-Img: per-image processing. Conv-CD: conventional contrastive decoding. SoFA: Tian et al. (2025).

4.5 Comparison with Alternative Approaches

We compare FOCUS against three alternative approaches to multi-image reasoning.

Per-Image Processing. As discussed in §2.2, the most straightforward way to avoid cross-image information leakage is to process each image independently. However, this approach inherently sacrifices the model's ability to perform holistic reasoning across multiple visual inputs, such as leveraging positional or contextual relationships among them. To mitigate this limitation, one possible strategy is to explicitly provide the image order as part of the textual prompt during decoding.

To examine this idea, we design an experiment where the model generates an independent response for each image. These responses are then concatenated, and the original image order information is added to the prompt to produce a final output.

As shown in Table 7, this approach leads to a significant drop in performance compared to the original multi-image setup. Even when the image order is explicitly provided, the model still fails to generate accurate answers. This result suggests that the model effectively utilizes inter-image relationships when multiple images are processed jointly, and that such relationships cannot be easily replaced by simple textual cues.

Conventional Contrastive Decoding. We further investigate whether contrastive decoding (CD), originally developed for single-image inputs, can be effectively extended to multi-image scenarios. Specifically, we apply VCD (Leng et al., 2024), a representative CD method for LVLMs. We expand VCD by contrasting the original multi-image logits with the logits obtained from the same input with

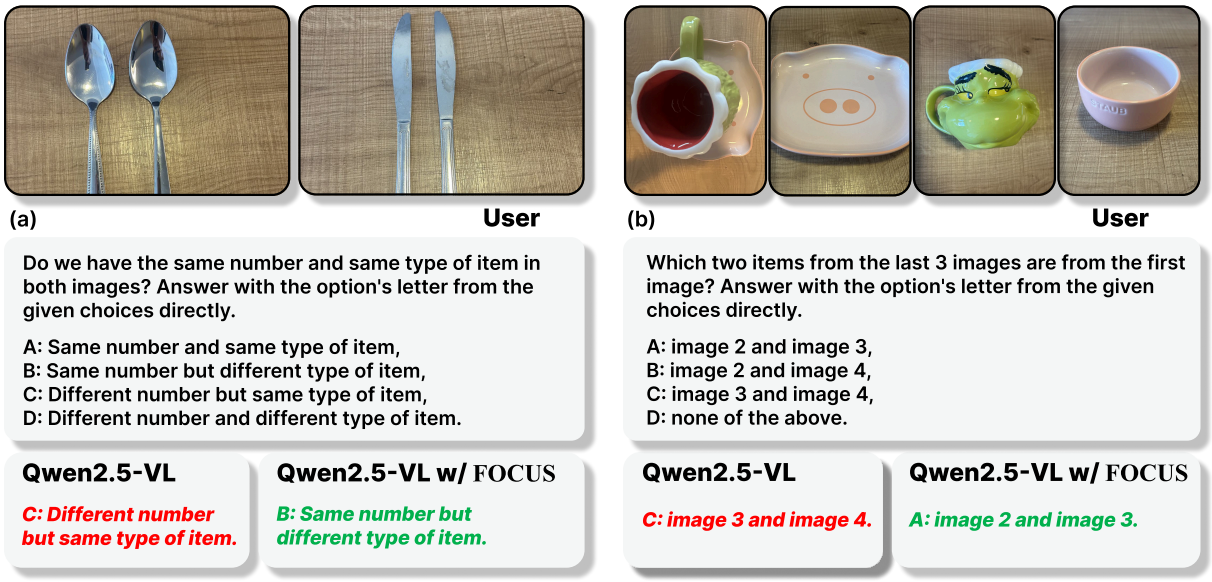


Figure 6: Qualitative results on Mantis-Eval. We compare the Qwen2.5-VL 3B model with and without FOCUS. These samples require the model to interpret each image independently while also integrating information across all images to reach the correct answer. The baseline fails under this setting, however FOCUS preserves per-image understanding and produces the correct responses.

added noise. Unlike our method, which independently masks each image position, VCD contrasts all original inputs with fully noised images. It does not consider cross-image information leakage.

Table 7 shows that conventional CD method in multi-image contexts degrades performance for several model families, even falling below the original model. This demonstrates that simply transferring CD techniques from single-image to multi-image settings not only fails to mitigate information leakage but also further distracts the model's output.

SoFA. Tian et al. (2025) propose SoFA, a training-free method that mitigates position bias of multi-image VL models. As SoFA targets position bias rather than cross-image information leakage, we include it here to examine whether mitigating position bias also alleviates leakage.

In Table 7, SoFA does not consistently improve performance either, and similarly drops below the baseline in certain configurations. This suggests that position bias and cross-image information leakage are distinct problems. In contrast, FOCUS consistently outperforms all alternatives.

4.6 Accuracy-Latency Trade Off

FOCUS introduces moderate latency overhead, which is justified by consistent accuracy gains. As shown in Table 8, FOCUS yields up to +3.22%pt accuracy improvement with manageable latency cost. The overhead stays well below the naive $(N+1)\times$ cost, as we batch the forward passes into a single inference call. Full details are provided in Appendix §C.

N	Qwen3-VL-4B		Qwen3-VL-8B	
	Baseline	+FOCUS	Baseline	+FOCUS
2	1.38 ± 2.82	2.12 ± 2.72	1.62 ± 2.91	2.20 ± 3.05
3	1.00 ± 1.48	2.64 ± 1.22	1.26 ± 1.78	2.67 ± 1.53
4	0.43 ± 0.34	1.35 ± 1.15	1.90 ± 4.26	2.47 ± 3.15
5	0.43 ± 0.00	1.03 ± 0.00	0.48 ± 0.00	1.04 ± 0.00
Acc.	66.82	**68.66**	65.44	**68.66**

Table 8: Latency (mean ± std, seconds) and accuracy on Mantis-Eval across different numbers of input images.

4.7 Qualitative Results

In Figure 6(a), while the baseline conflates visual content across images, FOCUS correctly differentiates objects and maintains accurate counts (Figure 6(b)). Additional results are in Appendix §E.

5 Conclusion

In this work, we empirically characterize an underexplored limitation of LVLMs in multi-image settings: *cross-image information leakage*, where models tend to mix visual information across images rather than reasoning over each independently. To address this, we propose FOCUS, a training-free inference method that masks non-target im-

ages with noise to guide the model to attend to one image at a time, and aggregates the resulting logits through contrastive logit aggregation. We validate FOCUS across diverse LVLM families and benchmarks, and demonstrate that its benefits extend to video understanding. Our results show that FOCUS offers a practical and generalizable inference-time solution for more reliable multi-image reasoning.

Limitations

Baseline Reproducibility. Since LVLM performance is sensitive to generation configuration and instruction prompts, the baseline results reported in this work may differ from those in the official model reports of Bai et al. (2025b,a); Chen et al. (2024b); Li et al. (2024a).

Hyperparameter Sensitivity. The hyperparameters λ and α are selected per model and benchmark via a validation set, and their generalization to unseen settings without validation data has not been verified. Full hyperparameter tables are provided in the Appendix §C.

Inference Overhead. FOCUS requires $N{+}1$ forward passes for N input images, introducing additional inference cost. In practice, this overhead is substantially smaller than the naive $(N{+}1)\times$ cost, since the passes share the same sequence length and are executed as a single batched call, yielding a 2.9 to 3.8$\times$ speedup over sequential execution (Appendix §C). Although the cost grows with the number of images, the batched implementation keeps it well below the naive $(N{+}1)\times$ bound, making FOCUS practical in typical multi-image settings.

Long-Context Applicability. Long-context multi-image understanding is another large body of research with distinct objectives from our work, which focuses on disambiguating visual content across a small set of images. While FOCUS shows promising gains on long-context tasks requiring cross-image comparison, it yields mixed results on retrieval-heavy tasks where the bottleneck lies in locating a single relevant image rather than disambiguating across multiple images.

Acknowledgements

This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants funded by the Korea government (MSIT) (No. RS-2025-25441313, Professional AI Talent Development Program for Multimodal AI Agents, Contribution: 50%; No. IITP-RS-2026-25547954, Artificial Intelligence Innovation Human Resources Development, Contribution: 10%) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00350430, Contribution: 40%). This work was also supported by the Hyundai Motor Chung Mong-Koo Foundation.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Ping Chen, Xiaoqin Zhang, and Shijian Lu. 2025. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29915–29926.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, and 1 others. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

Rabiul Awal, Saba Ahmadi, Le Zhang, and Aishwarya Agrawal. 2024. Vismin: Visual minimal-change understanding. *Advances in Neural Information Processing Systems*, 37:107795–107829.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025b. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Junzhe Chen, Tianshu Zhang, Shiyu Huang, Yuwei Niu, Linfeng Zhang, Lijie Wen, and Xuming Hu. 2025. Ict: Image-object cross-level trusted intervention for mitigating object hallucination in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4209–4221.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. 2024a. HALC: Object hallucination reduction via adaptive focal-contrast decoding. In *Forty-first International Conference on Machine Learning*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267.

Erik Daxberger, Nina Wenzel, David Griffiths, Haiming Gang, Justin Lazarow, Gefen Kohavi, Kai Kang, Marcin Eichner, Yinfei Yang, Afshin Dehghan, and 1 others. 2025. Mm-spatial: Exploring 3d spatial understanding in multimodal llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7395–7408.

Xin Dong, Shichao Dong, Jin Wang, Jing Huang, Li Zhou, Zenghui Sun, Lihua Jing, Jingsong Lan, Xiaoyong Zhu, and Bo Zheng. 2025. Inter: Mitigating hallucination in large vision-language models by interaction guidance sampling. *arXiv preprint arXiv:2507.05056*.

Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. 2019. Image captioning: Transforming objects into words. *Advances in neural information processing systems*, 32.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. 2024. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*.

Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, and 1 others. 2023. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36:71683–71702.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and 1 others. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. 2024b. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, Toronto, Canada. Association for Computational Linguistics.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26689–26699.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Nikolay Malkin, Zhen Wang, and Nebojsa Jojic. 2022. Coherence boosting: When your pretrained language model is not paying enough attention. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8214–8236, Dublin, Ireland. Association for Computational Linguistics.

Yeji Park, Deokyeong Lee, Junsuk Choe, and Buru Chang. 2025. Convis: Contrastive decoding with hallucination visualization for mitigating hallucinations in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 6434–6442.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. 2024. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 783–791, Mexico City, Mexico. Association for Computational Linguistics.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A corpus of natural language for visual reasoning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 217–223, Vancouver, Canada. Association for Computational Linguistics.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2018. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. 2024. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409.

Wei Suo, Lijun Zhang, Mengyang Sun, Lin Yuanbo Wu, Peng Wang, and Yanning Zhang. 2025a. Octopus: Alleviating hallucination via dynamic contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 29904–29914.

Wei Suo, Lijun Zhang, Mengyang Sun, Lin Yuanbo Wu, Peng Wang, and Yanning Zhang. 2025b. Octopus: Alleviating hallucination via dynamic contrastive decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29904–29914.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.

Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. 2025. Identifying and mitigating position bias of multi-image vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10599–10609.

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. 2025a. MLLM can see? dynamic correction decoding for hallucination mitigation. In *The Thirteenth International Conference on Learning Representations*.

Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Lorena Yan, Wenjie Jacky Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, and 2 others. 2025b. Muirbench: A comprehensive benchmark for robust multi-image understanding. In *The Thirteenth International Conference on Learning Representations*.

Weiyun Wang, Shuibo Zhang, Yiming Ren, Yuchen Duan, Tiantong Li, Shuo Liu, Mengkang Hu, Zhe Chen, Kaipeng Zhang, Lewei Lu, and 1 others. 2024. Needle in a multimodal haystack. *Advances in Neural Information Processing Systems*, 37:20540–20565.

Zhaowei Wang, Wenhao Yu, Xiyu Ren, Jipeng Zhang, Yu Zhao, Rohit Saxena, Liang Cheng, Ginny Wong, Simon See, Pasquale Minervini, and 1 others. 2025c. Mmlongbench: Benchmarking long-context vision-language models effectively and thoroughly. *Advances in Neural Information Processing Systems*, 38.

Jianrui Zhang, Mu Cai, and Yong Jae Lee. 2024. Vinoground: Scrutinizing lmms over dense temporal reasoning with short videos. *arXiv preprint arXiv:2410.02763*.

Juntian Zhang, Chuanqi Cheng, Yuhan Liu, Wei Liu, Jian Luan, and Rui Yan. 2025. Weaving context across images: Improving vision-language models through focus-centric visual chains. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27782–27798, Vienna, Austria. Association for Computational Linguistics.

Bingchen Zhao, Yongshuo Zong, Letian Zhang, and Timothy Hospedales. 2024. Benchmarking multi-image understanding in vision and language models: Perception, knowledge, reasoning, and multi-hop reasoning. *arXiv preprint arXiv:2406.12742*.

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*.

A Related Work

While LVLMs acheive strong performance on standard single-image tasks (Chen et al., 2024b; Li et al., 2024a), their ability to understand multi-image remains underdeveloped. Many open-source LVLMs (Dai et al., 2023; Liu et al., 2023; Zhu et al., 2024; Liu et al., 2024) are trained under the assumption of single-image inputs, and therefore struggle to generalize when presented with multiple images simultaneously (Wang et al., 2025b).

Multi-Image LVLMs. To address this gap, several recent LVLMs (Awadalla et al., 2023; Sun et al., 2024; Laurençon et al., 2023; Lin et al., 2024; Li et al., 2024b) have been trained using large-scale interleaved image-text datasets (Laurençon et al., 2023; Li et al., 2024b). For example, Laurençon et al. (2023) comprises hundreds of millions of interleaved image-text data that support multiple, contextually related images. While these approaches improve multi-image understanding capabilities, they require extensive training resources (Jiang et al., 2024) and suffer from limited reusability and flexibility due to their reliance on model-specific architectures. However, even recent state-of-the-art models continue to show notable performance gaps in multi-image settings (Wang et al., 2025b), and the underlying causes remain insufficiently studied. Tian et al. (2025) identify position bias in multi-image settings, where predictions vary with image order. In contrast, our work targets cross-image information leakage, where the model confuses visual content across images regardless of their order.

Training-free Methods. Training-free methods guide generation without additional training or architectural changes. These approaches manipulate logits or hidden states during inference to steer outputs toward desired properties (Li et al., 2023; Malkin et al., 2022; Shi et al., 2024). Within LVLMs, a growing body of work (Leng et al., 2024; Huang et al., 2024; Chen et al., 2024a; Wang et al., 2025a; Park et al., 2025; Chen et al., 2025) has shown promising results in reducing hallucination, but they have predominantly been developed under single-image assumptions (Suo et al., 2025a; Dong et al., 2025; Suo et al., 2025b; An et al., 2025). Tian et al. (2025) propose a training-free method that mitigates position bias by modifying inter-image attention masks. However, their method targets position bias rather than cross-image information leakage. In contrast, our method directly addresses cross-image information leakage through logit manipulation.

B Benchmarks

Winoground (Thrush et al., 2022) evaluates whether a model can correctly associate captions with images or vice versa. It contains 400 samples, each consisting of two images and two captions, $\{(I_1, T_1), (I_2, T_2)\}$, with subtle differences. The two captions contain the same words but in a different order. The model must assign the correct caption to each image.

VisMin (Awal et al., 2024) builds on this setup by focusing on 2K image-text pairs with minimal semantic differences. It is organized into four semantic categories, targeting more fine-grained discrimination. We further split these into VisMin-Easy and VisMin-Hard, where VisMin-Easy contains Object and Attribute categories, and VisMin-Hard contains Relation and Counting categories.

VisMin-Easy tests mutually exclusive properties by nature, as an object cannot simultaneously belong to two distinct categories, nor can an attribute hold two contradictory values at once.

VisMin-Hard tests more inherently ambiguous categories, as spatial relations can co-occur across images and numerical counts are susceptible to cross-image information leakage.

Evaluation metrics for Winoground and VisMin. We adopt the generative LVLM evaluation protocol introduced in Awal et al. (2024), which is designed to assess open-ended generation by three metrics.

- *Text Score*: For each input set $\{I_1, T_1, T_2\}$ and $\{I_2, T_1, T_2\}$, the model is prompted to choose the caption that best matches the image. A point is awarded only if the model selects the correct caption for both inputs. This metric evaluates single-image caption grounding.
- *Image Score*: For each input set $\{I_1, I_2, T_1\}$ and $\{I_1, I_2, T_2\}$, the model is prompted to select the image that best matches the caption. A point is awarded only if the correct image is chosen for both inputs. This metric evaluates comparative reasoning and primarily reflects multi-image understanding.
- *Group Score*: It requires both the Text Score and the Image Score to be correct. It reflects the model's ability to reason consistently across both single-image and multi-image contexts.

Recet Multi-image Benchmarks. We also evaluate on three recent multi-image benchmarks for

LVLMs. Mantis-Eval (Jiang et al., 2024) covers diverse topics such as size perception and weight comparison. Each sample is crafted by annotators to require deep cross-image understanding. MuirBench (Wang et al., 2025b) assesses 12 multi-image reasoning skills across 10 inter-image relation types. It pairs each instance with an unanswerable variant to test fine-grained discrimination. MIRB (Zhao et al., 2024) evaluates cross-image comparison and reasoning across four categories: perception, visual knowledge, reasoning, and multi-hop reasoning. All three benchmarks are evaluated by accuracy.

Vinoground (Zhang et al., 2024) is a temporal counterfactual benchmark consisting of 1,000 short video-caption pairs designed to evaluate whether models can distinguish temporal differences between actions and object transformations. Models are evaluated with Text, Video, and Group scores, following the same protocol as Winoground.

MMLongBench (Wang et al., 2025c) is a long-context vision-language benchmark consisting of 13,331 examples spanning five task categories, including Visual RAG and Many-Shot ICL, evaluated at standardized input lengths from 8K to 128K tokens. We evaluate on four categories: ICL, DocVQA, VRAG, and NIAH, using Qwen2.5-VL-7B at 8K context length.

B.1 Data Usage and Licensing

All datasets used in this work are employed strictly for research and evaluation purposes. Winoground (Thrush et al., 2022) is used under the Meta Images Research License, which permits use for non-commercial research and evaluation. Mantis-Eval (Jiang et al., 2024) is used under the Apache 2.0 License. MuirBench (Wang et al., 2025b) and MIRB (Zhao et al., 2024) are used in accordance with their respective terms of use, which permit research use with appropriate attribution. All datasets are used solely for evaluation without any modification or redistribution.

C Implementation Details

Models. We evaluate on four representative LVLM families: InternVL3 (2B, 8B) (Chen et al., 2024b), Qwen2.5-VL (3B, 7B) (Bai et al., 2025b), Qwen3-VL (4B, 8B, 32B, 30B-A3B) (Bai et al., 2025a), and LLaVA-OneVision (0.5B, 7B) (Li et al., 2024a). All models are evaluated in a frozen state without any fine-tuning; only the decoding phase is modified. Model checkpoints are publicly available on Hugging Face.[1]

Dataset	Var.	InternVL3		Qwen2.5-VL		LLaVA-OV	
		2B	8B	3B	7B	0.5B	7B
Winog.	λ	1	1	0.6	1	1	1
	α	0.6	0.8	0.15	0.2	0.1	0.3
VisMin	λ	1	0.7	0.5	0.7	1	0.9
	α	0.4	0.7	0.5	0.15	0.1	0.2
Mantis	λ	0.8	0.3	0.8	0.3	0.1	0.2
	α	0.1	0.1	0.5	0.4	0.2	0.5
Muir.	λ	0.2	0.5	0.8	0.6	0.1	1
	α	0.1	0.1	0.1	0.1	0.1	0.1
MIRB	λ	0.3	0.2	0.1	0.3	0.1	0.1
	α	0.1	-0.15	-0.2	0.1	-0.15	0.1

Table C1: Hyperparameters used for evaluation. Search space: $\lambda \in (0, 1)$ and $\alpha \in (-1, 1)$.

Decoding Setup. We use multinomial sampling with temperature $T = 0.2$.

Noise Masking. Each image v_i is masked using additive uniform noise as $v_i' = (1-\lambda)\cdot v_i + \lambda \cdot \mathcal{U}(0, 1)$, where $\lambda \in [0, 1]$ is a hyperparameter controlling the degree of corruption. The masking is applied in the normalized image space using ImageNet statistics ($\mu = [0.485, 0.456, 0.406]$, $\sigma = [0.229, 0.224, 0.225]$), then denormalized back to pixel space and clipped to $[0, 1]$. The optimal value of λ is selected via validation per model and benchmark.

Hyperparameter Tuning. All hyperparameters including noise type, noise scale λ, and aggregation weight α are selected per model and benchmark via validation. For VisMin, we randomly sample a subset of the training set to construct a validation split. For benchmarks without official validation sets, we randomly choose 10% of the test data for this purpose.

Hyperparameter Settings To ensure reproducibility, we provide the list of hyperparameters used in our experiments in Table C1. The two primary hyperparameters are the noise scale λ and

[1] https://huggingface.co/OpenGVLab/InternVL3-2B, https://huggingface.co/OpenGVLab/InternVL3-8B, https://huggingface.co/Qwen/Qwen2.5-VL-3B-Instruct, https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct, https://huggingface.co/Qwen/Qwen3-VL-4B-Instruct, https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct, https://huggingface.co/Qwen/Qwen3-VL-32B-Instruct https://huggingface.co/Qwen/Qwen3-VL-30B-A3B-Instruct, https://huggingface.co/llava-hf/llava-onevision-qwen2-0.5b-ov-hf, https://huggingface.co/llava-hf/llava-onevision-qwen2-7b-ov-hf

the contrastive weight α. Intuitively, λ controls the strength of visual noise injection, while α balances the contrastive aggregation weight. We perform a random search over $\lambda \in (0, 1)$ and $\alpha \in (-1, 1)$ using validation set. Across datasets and model sizes, we observe two consistent trends: (i) the noise scale λ generally converges to relatively large values, and (ii) the contrastive weight α remains stable around a small magnitude. These patterns hold regardless of the underlying architecture or parameter count. More specifically, we observe that λ tends to saturate at 1.0 for most cases in the Winoground dataset, while α is consistently selected as 0.1 for MuirBench.

More Details. Experiments are conducted on NVIDIA A100 (80GB) and RTX 6000 Pro GPUs. For memory efficiency, images are resized to 512×512 for MuirBench and MIRB evaluations.

SoFA Reproduction. As the official implementation of SoFA (Tian et al., 2025) is unavailable, we reproduce it based on the original paper. We set $\sigma = 0.5$ and apply the interpolation every two layers, following the experimental setup described in the original work.

Accuracy-Latency Trade-off Details. We conduct latency experiments on Mantis-Eval (Jiang et al., 2024) using Qwen3-VL-4B and Qwen3-VL-8B across $N \in \{2, 3, 4, 5\}$ input images, reporting wall-clock end-to-end latency per sample. Since Mantis-Eval has an uneven distribution of samples across image counts, where the $N = 4$ and $N = 5$ subsets contain significantly fewer samples than $N = 2$ and $N = 3$, we report average latency per N and average accuracy over the full test split separately. For the same reason, latency should be compared between the baseline and FOCUS within each N rather than across different N, as each subset has its own answer-length distribution.

Batched Implementation. Rather than running the $N{+}1$ forward passes sequentially, we execute them as a single batched inference call. All $N{+}1$ passes share the same sequence length, so batching is efficient on the LLM side. This keeps the practical overhead far below the naive $(N{+}1)\times$ cost.

Inference Cost We run a controlled synthetic experiment to isolate how the number of input images affects latency. For each N, we generate a fixed set of N random image tensors of shape $(512, 512, 3)$ and repeat each run three times. We fix the maximum number of new tokens to 64. We report the average per-sample latency (s/sample). As shown in Table C2, batched inference yields a $2.9\times$ to $3.8\times$ speedup over the naive $(N{+}1)\times$ cost across all tested image counts. The overhead is modest in the small-N, and our batched implementation keeps it well below the naive $(N{+}1)\times$ cost as N grows.

N	Latency		Speedup
	Baseline	**FOCUS**	**vs.** $(N{+}1)\times$
2	1.87	1.96	2.86
4	1.93	2.58	3.73
8	2.11	4.97	3.81
16	2.47	14.26	2.94

Table C2: Inference cost of FOCUS versus the number of input images N. Per-sample latency in seconds, measured on synthetic $(512, 512, 3)$ tensors and averaged over three runs. The last column is the speedup of our batched implementation over the naive $(N{+}1)\times$ cost.

D Additional Experiments

D.1 Cross-Image Information Leakage in Larger Models

We replicate the analyses from Section §2.1 and Section §4.4 on Qwen2.5-VL-7B and Qwen3-VL-8B to verify that cross-image information leakage is not an artifact of smaller models.

Behavioral Analysis. Following the same setup as Section 2.1, the model selects among a correct caption, an incorrect caption from another image, and a merged caption combining both. As shown in Figure D1, accuracy consistently degrades under multi-image conditions across both scales. For Qwen2.5-VL-7B, merged-caption selection increases under multi-image input, a direct behavioral signature of cross-image confusion. For Qwen3-VL-8B, merged-caption selection decreases, yet accuracy still drops substantially (85.30 → 76.66). This suggests that stronger models may exhibit leakage.

Representational Analysis. As shown in Figure D2, for Qwen3-VL-8B, multi-image input increases hidden state similarity (0.340 → 0.366), and FOCUS reduces it back toward the single-image baseline (0.353), consistent with cross-image entanglement at the representational level. For Qwen2.5-VL-7B, the difference between single- and multi-image conditions is minimal (0.336 vs. 0.335). Nevertheless, FOCUS further reduces similarity (0.330), suggesting active suppression of cross-image leakage even when the effect is not fully captured by this metric. These re-

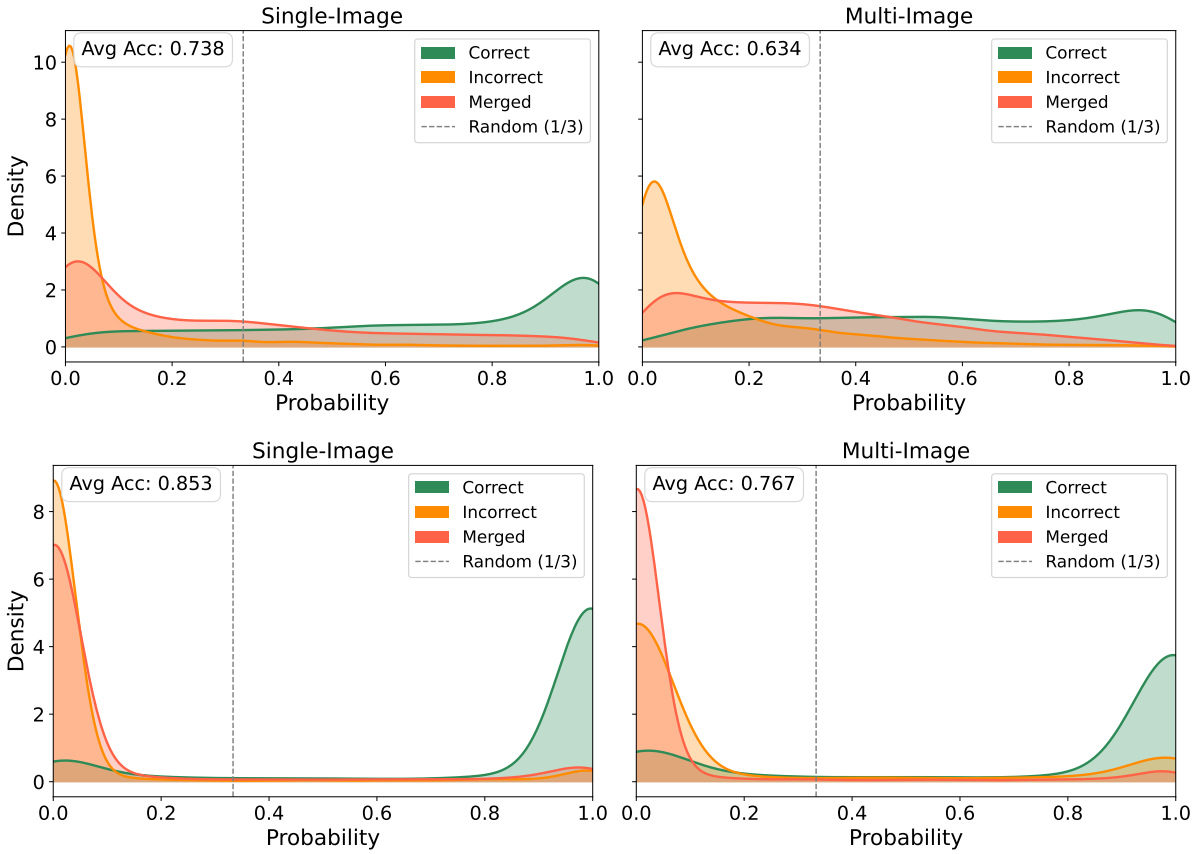


Figure D1: Estimation of probability density function of model predictions over three answer options, replicating the analysis of Figure 3 at larger model scales: Qwen2.5-VL-7B (top) and Qwen3-VL-8B (bottom). Accuracy consistently degrades under multi-image conditions across both scales, confirming that cross-image information leakage is not an artifact of smaller models.

Benchmark	Metric	Baseline	+FOCUS
VisMin	T	91.06	91.06
	I	70.12	**73.51**
	G	67.88	**70.98**
Mantis-Eval	Acc.	64.06	**66.36**

Table D1: Results on Qwen3-VL-30B-A3B (MoE) on VisMin and Mantis-Eval. FOCUS consistently improves Image Score (+3.39 pp), Group Score (+3.10 pp), and Mantis-Eval accuracy (+2.30 pp), while Text Score remains unchanged.

Benchmark	Metric	Baseline	+FOCUS
Winoground	T	81.50	**82.50**
	I	54.75	**70.50**
	G	49.75	**65.50**
VisMin	T	**91.42**	**91.42**
	I	88.08	**90.12**
	G	83.89	**85.47**
Mantis-Eval	Acc.	74.65	**75.58**

Table D2: Results with Qwen3-VL-32B. T, I, and G denote Text, Image, and Group score. VisMin results are averaged over the Object, Attribute, and Counting categories.

sults indicate that cross-image information leakage persists across model scales, though the degree and type of degradation vary with model capacity.

D.2 Results on Larger Models

To verify that FOCUS generalizes beyond the model scales and architectures in our main experiments, we evaluate two additional models. We first evaluate Qwen3-VL-30B-A3B, a MoE architecture. As shown in Table D1, FOCUS consistently improves multi-image relevant metrics across both benchmarks, confirming that its benefits are not limited to dense architectures.

We further evaluate Qwen3-VL-32B on Winoground, VisMin, and Mantis-Eval. As shown in Table D2, FOCUS yields substantial gains on

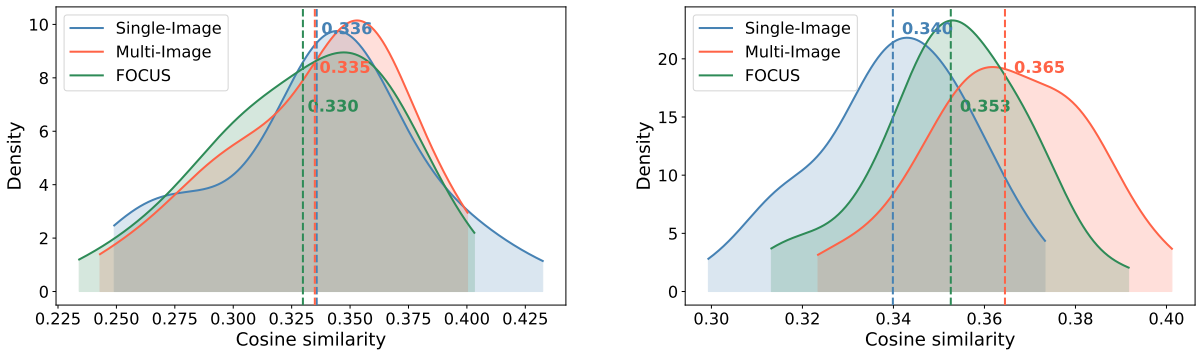


Figure D2: Average pairwise cosine similarity distribution of image token representations under three conditions (single-image, multi-image, and +FOCUS), replicating the analysis of Figure 5 at larger model scales: Qwen2.5-VL-7B (left) and Qwen3-VL-8B (right).

Winoground, improving the Image Score from 54.75 to 70.50 and the Group Score from 49.75 to 65.50. These results indicate that FOCUS remains effective at larger scale, and that its gains are not an artifact of smaller architectures.

D.3 Extension to Long-Context Setting.

Category	Dataset	Metric	Baseline	+FOCUS
ICL	Cars	Acc	20.75	**26.42**
	Food	Acc	24.00	**34.00**
	iNat	Acc	18.00	**22.00**
DocVQA	LongDocURL	doc_qa	58.94	**60.60**
	MMLongDoc	doc_qa	51.05	**53.29**
	SlideVQA	doc_qa	65.75	**68.68**
VRAG	InfoSeek	SubEM	**60.76**	59.78
	ViQuAE	SubEM	**55.68**	54.19
NIAH	VH-Single	Acc	55.57	**55.85**
	VH-Multi	Acc	54.37	**54.77**
	MM-NIAH Count	Soft Acc.	36.88	**38.95**
	MM-NIAH Retriev.	Acc.	**25.56**	25.06
	MM-NIAH Reason.	Acc.	**51.87**	51.52
Average			44.55	**46.54**

Table D3: Results on MMLongBench (Wang et al., 2025c) using Qwen2.5-VL-7B. FOCUS shows consistent gains on ICL and DocVQA tasks requiring cross-image comparison.

We examine whether FOCUS generalizes to long-context settings by evaluating on MMLongBench (Wang et al., 2025c) across four task categories (ICL, DocVQA, VRAG, and NIAH) using Qwen2.5-VL-7B at 8K context length. Hyperparameters are set to $\alpha = 0.1$ and $\lambda = 0.1$ for most datasets, with per-dataset selection for a subset of tasks. Since long-context multi-image understanding is a large body of research (Wang et al., 2024) with distinct objectives from our work, we do not claim full applicability to this setting. Nevertheless, as shown in Table D3, FOCUS shows promising gains on tasks that align with our design objective of cross-image disambiguation, particularly on ICL and DocVQA tasks. Results are mixed on VRAG and NIAH tasks, where the bottleneck lies in locating a single relevant image within a long context rather than disambiguating across multiple images simultaneously.

D.4 Ablation Study

We conduct ablation studies to assess each component in FOCUS. All experiments are conducted on the Mantis validation set using Qwen2.5-VL 7B. If not specified, we set noise scale $\lambda = 0.3$ and weighting $\alpha = 0.4$.

Effect of Noise Type. In Table D4 (a), we compare three types of noise: Gaussian, Impulse, and Uniform for masking non-target images during FOCUS inference. Although all noise variants serve to suppress irrelevant image cues, their effectiveness varies. Uniform noise yields the highest accuracy at 76.19%. In contrast, Gaussian and Impulse noise achieves comparatively lower accuracy.

Effect of λ (Noise Scale). Table D4 (b) reports the effect of the noise strength λ, which controls how much noise is applied to non-target images. A low value $\lambda = 0.1$ moderately improves performance to 71.43%, but is insufficient to fully suppress leakage. A high value $\lambda = 1.0$ degrades performance to 52.38% by overly corrupting the image. Our default setting $\lambda = 0.3$ yields the highest performance at 76.19%, demonstrating that balancing semantic masking and structural retention is key.

(a) Noise Type		(b) Noise Scale		(c) Weight	
Variant	**Acc.**	λ	**Acc.**	α	**Acc.**
Gaussian	71.43	0.1	71.43	0.1	66.67
Impulse	66.67	**0.3**	**76.19**	**0.4**	**76.19**
Uniform	**76.19**	1.0	52.38	1.0	61.90

Table D4: Impact of FOCUS design choices: noise type, masking strength λ, and contrastive weight α. Mantis validation accuracies using Qwen2.5-VL-7B. Bold indicates best per group.

Reference Type	**Mantis-Eval (Test)**
w/o Reference	69.12
Clean Image as Reference	67.28
FOCUS (Noise Reference)	**70.05**

Table D5: Ablation on the choice of reference type, using Qwen2.5-VL-7B on the Mantis-Eval test set. The per-image noise term is held fixed and only the reference type varies.

Effect of α (Weight). Table D4 (c) shows how the contrastive weight α used to subtract noise reference f_{noise} affects performance. A small $\alpha = 0.1$ achieves 66.67%, indicating a benefit from incorporating the noise-only reference. Increasing α to 0.4 (our choice) leads to the best result at 76.19%. However, setting $\alpha = 1.0$ reduces performance to 61.90%, likely due to over-suppressing the clean image logits. This suggests that contrastive suppression should be applied carefully to preserve essential signal while reducing information leakage.

Effect of Reference Type Table D5 compares three choices of reference type on the Mantis-Eval test set. We hold the per-image noise term fixed and vary only the reference. is designed to cancel the spurious signal that masking introduces (i.e., the added noise), not genuine image content. A clean image reference removes real visual information along with the artifact. It reduces accuracy to 67.28%, worse than using no reference at all (69.12%). Pure noise instead improves over that baseline, reaching 70.05%. This confirms that the clean-image reference removes useful visual evidence, while the noise reference cancels only the masking artifact.

D.5 Detailed VisMin Results

VisMin(Awal et al., 2024) contains four semantic categories: Object, Attribute, Relation, and Counting. Image–text pairs differ only minimally within each category, which makes the benchmark highly sensitive to multi-image reasoning. In Table D6, we present results with and without FOCUS.

Object. FOCUS shows clear gains in both the Image Score and the Group Score. These metrics are essential for evaluating cross-image understanding. Improvements appear in almost every model group except InternVL3-8B.

Attribute. In InternVL3-2B, FOCUS achieves nearly three times higher Group Score compared to the baseline. This result highlights its strengthened capability in distinguishing fine-grained visual properties across images.

Relation. Llava-OV-7B baseline obtains an Image Score of 35.53, whereas FOCUS reaches 45.82. The Group Score also increases from 32.15 to 42.93. These improvements demonstrate significantly enhanced relational reasoning.

Counting. FOCUS shows consistent improvements across all model groups in Counting. The InternVL3-2B model exhibits the largest performance increase.

Summary. Overall, FOCUS consistently and robustly outperforms the baselines across all VisMin semantic categories. The method shows strong generalization in understanding each image while effectively mitigating cross-image information leakage.

E Additional Qualitative Results

Figure D3 compares baseline decoding with FOCUS using examples from the VisMin benchmark. In (a), the baseline incorrectly mentions both bananas and oranges regardless of which image is queried. In contrast, FOCUS describes only bananas for the first image and only oranges for the second, showing a clear separation of visual content. In (b), the baseline mixes elements from both images, referring to doughnuts and sandwiches even when asked about one. FOCUS avoids this confusion and generates precise, image-specific descriptions. These examples show that baseline decoding fails to disentangle visual signals across multiple images, whereas FOCUS effectively suppresses cross-image information leakage.

Figure D4 presents additional examples from the Mantis benchmark. In (a), the baseline misunderstands that the only difference between the two images is the number of items. FOCUS correctly identifies this distinction and provides the right answer. In (b), the two images depict bacteria and viruses.

Category	Object			Attribute			Relation			Counting		
	T	I	G	T	I	G	T	I	G	T	I	G
InternVL3-2B	95.68	36.10	35.92	90.82	23.81	22.79	**60.77**	1.29	0.96	**70.29**	12.39	11.38
+ FOCUS	**95.85**	**78.58**	**76.86**	90.82	**70.07**	**65.65**	60.13	**10.93**	**8.36**	69.61	**42.61**	**39.56**
InternVL3-8B	96.03	**86.01**	**83.94**	93.20	**83.67**	**79.25**	90.68	47.43	46.14	75.04	47.88	43.46
+ FOCUS	**95.85**	85.49	83.77	**93.54**	80.61	77.89	89.87	**50.64**	**48.71**	**75.21**	**54.67**	**49.07**
QwenVL2.5-3B	**94.82**	48.19	47.84	89.46	47.28	44.56	83.44	24.12	20.90	71.31	30.05	24.96
+ FOCUS	94.65	**60.10**	**59.93**	89.46	**53.74**	**51.70**	**84.24**	**24.60**	**21.38**	**72.16**	**33.11**	**29.20**
QwenVL2.5-7B	96.20	94.47	92.40	91.50	**90.82**	**84.69**	**91.32**	50.48	49.52	77.42	56.03	50.42
+ FOCUS	96.20	**95.34**	**93.09**	91.50	89.80	84.35	91.00	**58.36**	**56.75**	**77.93**	**64.52**	**57.22**
LLaVA-OneVision-0.5B	29.02	10.36	2.76	12.93	8.16	0.34	0.00	**21.22**	0.00	3.40	19.02	1.02
+ FOCUS	**29.36**	**28.84**	**11.74**	**12.59**	**25.17**	**5.10**	**0.32**	15.59	**0.16**	**4.07**	**21.22**	**1.02**
LLaVA-OneVision-7B	96.03	63.04	61.83	90.82	**62.93**	**59.52**	86.17	35.53	32.15	75.04	53.31	44.65
+ FOCUS	96.03	**69.43**	**67.88**	90.82	58.84	56.12	86.17	**45.82**	**42.93**	**74.70**	**69.95**	**55.86**

Table D6: Performance across categories (Object, Attribute, Relation, Counting) with and without the proposed **FOCUS** decoding method. T: Text Score, I: Image Score, G: Group Score. Bold indicates improvement over the baseline. Underline highlights the largest gain within each model group for a given category.

The model should respond accurately about the queried image, yet the baseline fails, while FOCUS answers correctly. These cases further highlight FOCUS's ability to maintain image-specific reasoning.

F Use of AI Assistants

We used AI assistants for polishing the manuscript and partially for code debugging. AI-generated images were used as input images in qualitative experiments to illustrate cross-image information leakage.

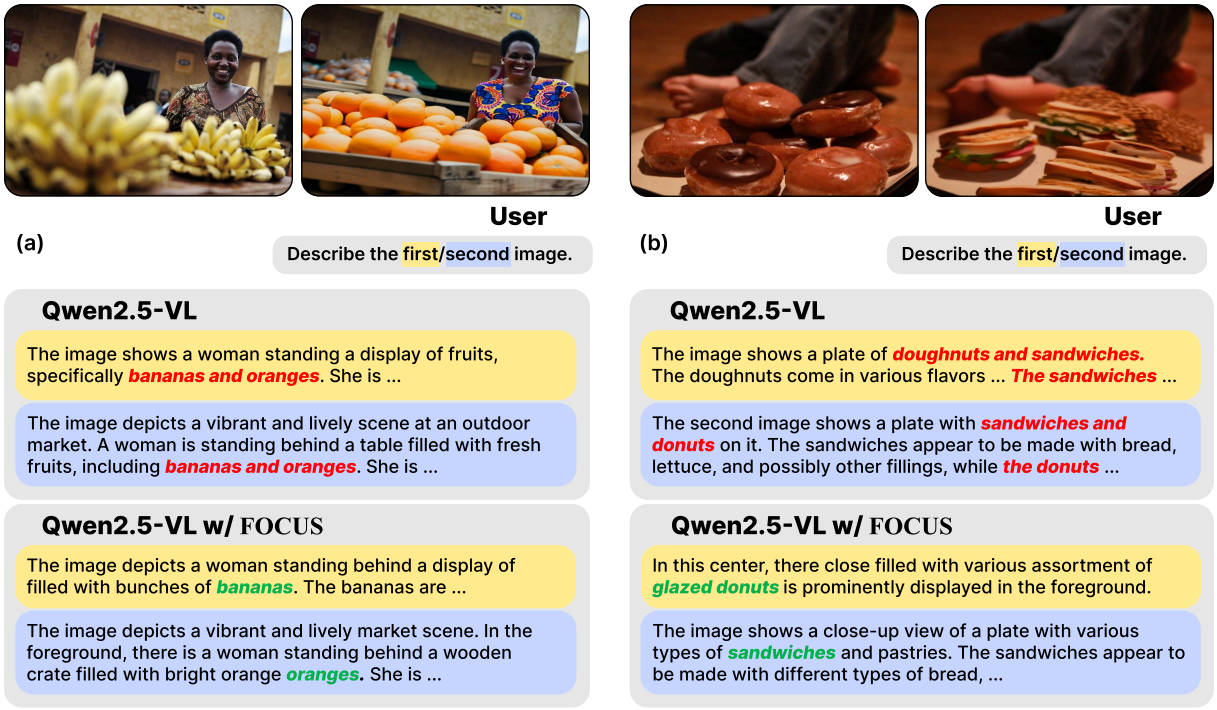


Figure D3: Qualitative samples using the Qwen2.5-VL 3B model. In both multi-image settings, the baseline decoding strategy often produces a mixed information of the other image not indicated by the question. On the other hand, FOCUS disentangles cross-image information effectively, resulting in better multi-image understanding. These examples illustrate that FOCUS helps generate image-specific responses while suprressing cross-image information leakage.

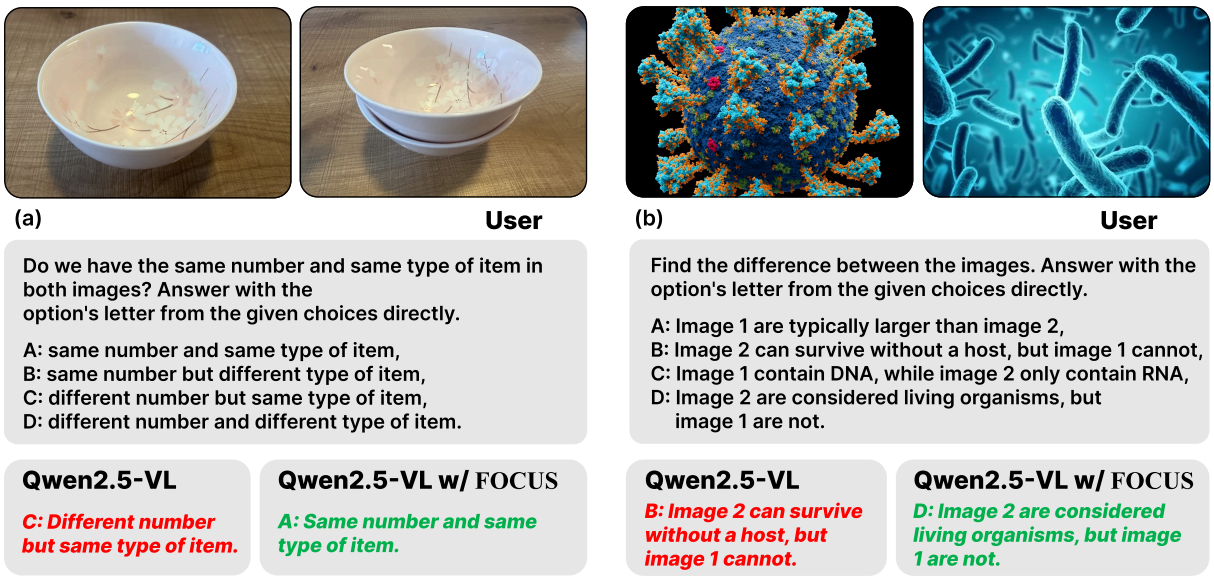


Figure D4: More qualitative samples using the Qwen2.5-VL 3B model. Examples are from the Mantis benchmark. Figure shows that baseline decoding confuses cross-image differences, while FOCUS correctly identifies image-specific information.