Mitigating Cross-Image Information Leakage in LVLMs for Multi-Image Tasks

Yeji Park[1], **Minyoung Lee**[1], **Sanghyuk Chun**[2*], **Junsuk Choe**[1†]

[1]Sogang University, [2]Princeton University

Abstract

Large Vision-Language Models (LVLMs) demonstrate strong performance on single-image tasks. However, we observe that their performance degrades significantly when handling multi-image inputs. This occurs because visual cues from different images become entangled in the model's output. We refer to this phenomenon as *cross-image information leakage*. To address this issue, we propose **FOCUS**, a training-free and architecture-agnostic decoding strategy that mitigates cross-image information leakage during inference. FOCUS sequentially masks all but one image with random noise, guiding the model to focus on the single clean image. We repeat this process across all target images to obtain logits under partially masked contexts. These logits are aggregated and then contrastively refined using a noise-only reference input, which suppresses the leakage and yields more accurate outputs. FOCUS consistently improves performance across four multi-image benchmarks and diverse LVLM families. This demonstrates that FOCUS offers a general and practical solution for enhancing multi-image reasoning without additional training or architectural modifications. Source code is available at https://github.com/yejipark-m/FOCUS

Introduction

Large Vision-Language Models (LVLMs) are designed to jointly understand visual and textual information (Li et al. 2023a; Liu et al. 2023; Dai et al. 2023; Chen et al. 2024a; Bai et al. 2025; Li et al. 2024a; Achiam et al. 2023), enabling them to perform a wide range of vision-language tasks, such as Visual Question Answering (VQA) (Antol et al. 2015) and Image Captioning (Herdade et al. 2019).

Although these successes have been largely achieved in single-image settings, challenges still remain when extending these models to multi-image settings, where LVLMs exhibit a notable performance drop (Wang et al. 2025b). We observed that current LVLMs, when given multiple image inputs, often fail to treat each image independently. Instead, they mix visual cues across inputs, leading to a phenomenon we term *cross-image information leakage*.

As illustrated in Figure 1 (a), a model can provide accurate predictions when a single image is given. However, when two images are given simultaneously, as in Figure 1 (b), the model generates incorrect responses that conflate visual elements from both images. Specifically, while image A contains "beer and bananas" and image B contains "banana and coffee container", the model mistakenly describes image A as "beer, banana, and coffee container", incorporating irrelevant content from image B, and vice versa.

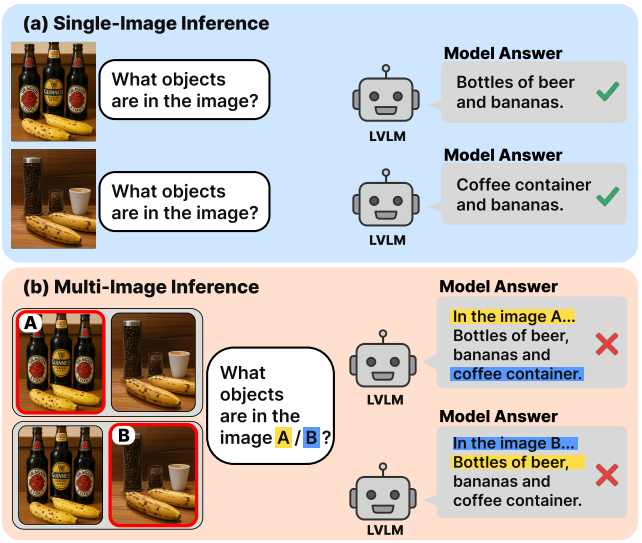

Figure 1: Illustration of cross-image information leakage in LVLMs during single-image vs. multi-image inference.

Why does cross-image information leakage happen? We hypothesize that this problem arises since language models (LMs) lack an explicit mechanism to separate visual semantics across different images. LMs often entangle information across inputs, generating outputs that mistakenly include unrelated content from different images. Our goal is to mitigate this cross-image information leakage and to achieve a better multi-image understanding of LVLMs.

Multi-image reasoning in LVLMs has been approached through both training-based and inference-based methods, yet key challenges remain in both. Training-based methods (Awadalla et al. 2023; Lin et al. 2024; Sun et al. 2024; Jiang et al. 2024) adopt interleaved image-text sequences to explicitly enhance multi-image reasoning capabilities. However, these methods require massive datasets (Laurençon

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

et al. 2023; Li et al. 2024b) and computational resources, often hindering practical use and scalability (Jiang et al. 2024).

To avoid expensive training computation and budget, there have been a number of works to improve image understanding in inference-time. For example, Tian et al. (2025) tackled multi-image tasks by adjusting the causal attention mask to reduce position bias. However, this method requires an intensive architecture-level modification that should be tailored to each model. As another line of work, inference-time decoding strategies in LVLM have been studied by editing logits or hidden states during generation (Leng et al. 2024; Chen et al. 2024b; Park et al. 2025). However, these methods are designed for single-image understanding and are usually not able to be generalized to multi-image tasks.

In this paper, we aim to mitigate the cross-image information leakage problem in a resource-efficient manner solely using inference-time operations. Our key idea is to encourage the model to focus on each image individually rather than processing all images simultaneously. To this end, we propose FOCUS, a novel decoding strategy that leverages a noise-guided image focusing technique. As illustrated in Figure 2, we mask all but one image with random noise, prompting the model to concentrate on the single clean image. We sequentially perform multiple forward passes, each time keeping one image clean, and aggregate the resulting output logits while preserving the positional context of the images. To suppress residual signals from the masked images, we perform an additional forward pass with an input where all images are noise-masked to compute a reference logit, which is subtracted from each output. This process yields a logit distribution that more faithfully reflects the model's independent understanding of each image.

We validate FOCUS on three LVLMs across four multi-image benchmarks: Winoground (Thrush et al. 2022), VisMin-Bench (Awal et al. 2024), Mantis-Eval (Jiang et al. 2024) and MuirBench (Wang et al. 2025b). FOCUS achieves consistent improvements, with the best gains of up to +32.1 Image and +29.9 Group score on VisMin, +18.8 Image and +16.8 Group score on Winoground, +5.5%pts accuracy on Mantis-Eval, and +1.5%pts on MuirBench each observed on different model families. These gains are achieved without any additional training or architectural changes, highlighting the effectiveness and generalizability of our method.

Related Work

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

Prior Work on 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. Training-free approaches that modify model structure have also been explored. For example, Tian et al. (2025) address position bias in multi-image LVLMs, by altering the causal attention mask used during auto-regressive generation. While effective at reducing positional preference, this method still requires model-specific architectural modifications, which may complicate integration into existing model pipelines. In contrast, we focus on a decoding strategy that requires no changes to a model.

Decoding Strategy. Methods focusing on decoding strategy 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. 2023b; Malkin, Wang, and Jojic 2022; Shi et al. 2024). Within LVLMs, several methods (Leng et al. 2024; Huang et al. 2024; Chen et al. 2024b; Wang et al. 2025a; Park et al. 2025; Chen et al. 2025; Suo et al. 2025; Dong et al. 2025) have shown promising results in reducing hallucination, but they have predominantly been developed under single-image assumptions. To the best of our knowledge, no prior work explicitly addresses or mitigates the cross-image information leakage problem in multi-image settings. Our work presents the first LVLM decoding method specifically designed for multi-image understanding.

Motivation

Preliminaries

LVLMs are composed of a vision encoder ϕ_v and an auto-regressive language model p_θ. Given an input image I, the model encodes it via the vision encoder as $v = \phi_v(I)$, and generates a textual output conditioned on both the image and the text input $\mathbf{X}$, as well as previously generated tokens $y_{<t}$ where output sequence has length L.

$$p(y \mid v, x) = \prod_{t=1}^{L} p_\theta(y_t \mid v, \mathbf{X}, y_{<t}) \tag{1}$$

When extending to N image inputs $\{I_1, I_2, \ldots, I_N\}$, each image I_i is independently encoded into visual tokens $v_i = \phi_v(I_i)$. These tokens are then concatenated as $[v_1, \ldots, v_N]$ and processed by the p_θ along with $\mathbf{X}$. While special tokens may indicate image boundaries, the model still operates over a flattened sequence with positional embeddings.

Cross-image Information Leakage

While concatenated visual tokens $[v_1, v_2, \ldots, v_N]$ allow the model to process multiple images simultaneously, they also introduce a key limitation: cross-image information leakage. This problem is connected with how visual inputs are embedded and processed within LVLMs. LVLMs usually handles multiple images as a sequence of visual tokens. These

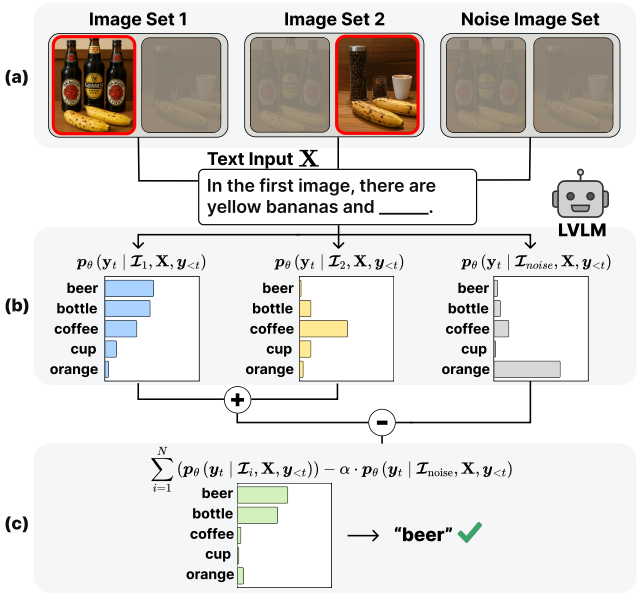

Figure 2: Overview of FOCUS. It consists of three main steps: (a) visual masking, (b) image-wise focused inference, and (c) contrastive aggregation.

visual tokens are then concatenated with text tokens and processed by the attention mechanism. During attention operations, the tokens are highly influenced by inter-image causal attention (Tian et al. 2025), resulting in a mingled representation in the latent space. A naive solution is to perform inference independently for each image (i.e., single-image inference). While this prevents the information leakage, it will discard positional and relational instructions (e.g., "first image", "compare the second image with the first"), that are essential for multi-image understanding tasks.

Method

Our method is a training-free decoding strategy that enables LVLMs to concentrate on one image at a time while preserving positional structure of multi-image inputs. As shown in Figure 2, our method consists of three steps: (a) visual masking, (b) image-wise focused inference, and (c) contrastive aggregation. The full procedure is described in Algorithm 1.

Visual Masking (Figure 2a). First, we prepare the masked image inputs using noise injection. Let N be the number of input images. For each inference $k = 1, \ldots, N$, we corrupt all images I_i into I'_i for $i = 1, \ldots, N$, except for the target image I_k. This allows the model to treat the corrupted images as masked and focus solely on a single clean image, resulting in a partially masked input $\mathcal{I}_i$:

$$\mathcal{I}_k = [v'_1, \ldots, v_k, \ldots, v'_N], \quad \mathcal{I}_{\text{noise}} = [v'_1, v'_2, \ldots, v'_N], \tag{2}$$

where v'_i denotes the noise-corrupted version of the visual embedding $\phi_v(I'_i)$. We also define a fully noise-masked input $\mathcal{I}_{\text{noise}}$ which serves as a noise-only reference.

Image-wise Focused Inference (Figure 2b). We run N inference passes, each using a partially masked image input $\mathcal{I}_k$, where only the k-th image remains unmasked. The original image order is preserved to retain positional semantics. For each i, we compute the logit distribution:

$$f_i = p_\theta(y_t \mid \mathcal{I}_i, \mathbf{X}, y_{<t}). \tag{3}$$

Figure 2 (b) depicts the different logit distributions f_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:

$$f_{\text{noise}} = p_\theta(y_t \mid \mathcal{I}_{\text{noise}}, \mathbf{X}, y_{<t}). \tag{4}$$

The gray logit distribution (right), containing no information from either the first or the second image, serves as a reference for isolating irrelevant visual content during contrastive aggregation. As a result, FOCUS performs $N + 1$ forward passes in total.

Contrastive Aggregation (Figure 2c). Each f_k contains useful signals from the clean image I_k, but also includes residual side-effects—e.g., nonzero values for irrelevant tokens like *orange* in Figure 2(b)—induced by the noise masks. To suppress these residuals, we subtract the noise reference logit f_{noise} from f_i and aggregate the results:

$$f_{\text{final}} = \sum_{k=1}^{N} (f_k - \alpha \cdot f_{\text{noise}}), \tag{5}$$

where the hyperparameter α is a scaling hyperparameter tuned via validation to appropriately weight the subtraction. The final output token is then sampled from the aggregated logits f_{final}. Figure 2(c) illustrates this step: the final green logit distribution, computed using Equation (5), correctly identifies *beer* as the answer.

This method effectively suppresses cross-image information leakage while preserving positional structure, without

---

Algorithm 1: FOCUS

Input: Images $I_1, \ldots, I_N$, Text input $\mathbf{X}$
Parameters: Noise scale λ, aggregation weight α
Models: Visual encoder ϕ_v, Language model p_θ
Functions: $\texttt{noise}(\cdot, \lambda)$
Output: Sampled token y_t

Initialize logits list $\mathcal{F} \leftarrow [\,]$
for $k = 1$ to N **do**
  $v'_j \leftarrow \phi_v(\texttt{noise}(I_j, \lambda))$ for all $j \neq k$
  $v_k \leftarrow \phi_v(I_k)$
  $\mathcal{I}_i \leftarrow [v'_1, \ldots, v_k, \ldots, v'_N]$
  $f_k \leftarrow p_\theta(y_t \mid \mathcal{I}_k, \mathbf{X}, y_{<t})$
  Append f_k to $\mathcal{F}$
end for
$v'_i \leftarrow \phi_v(\texttt{noise}(I_i, \lambda))$ for all $i = 1$ to N
$\mathcal{I}_{\text{noise}} \leftarrow [v'_1, \ldots, v'_N]$
$f_{\text{noise}} \leftarrow p_\theta(y_t \mid \mathcal{I}_{\text{noise}}, \mathbf{X}, y_{<t})$
$f_{\text{final}} \leftarrow \sum_{k=1}^{N} (f_i - \alpha \cdot f_{\text{noise}})$
return $y_t \sim \text{Sample}(f_{\text{final}})$

requiring any model training or architecture modification. It enables LVLMs to better perform multi-image reasoning tasks in a training-free, generalizable manner.

Cross-image Information Leakage Analysis

In this section, we empirically verify the severity of cross-image information leakage in LVLMs. We follow the protocols of Winoground (Thrush et al. 2022) and VisMin (Awal et al. 2024) object category validation sets to test whether the model can independently reason about each image.

Input Setup. For each image pair (I_1, I_2), we define two evaluation scenarios: (1) the model answers based only on I_1, and (2) the model answers based only on I_2. Each scenario is tested under two conditions: a single-image setting (only the target image is provided) and a multi-image setting (both I_1 and I_2 are given as input).

Target task. For both single-image and multi-image settings, we let an LVLM select the answer from the following three options. (1) *Target-specific caption* (c_T): A caption accurately describes the target image. (2) *Distractor caption* (c_D): A caption describes the non-target image. (3) *Merged caption* (c_M): A caption contains combined content from both images, i.e., a merged description.

Ideally, the model should select c_T depending on which image is being queried (I_1 or I_2). If the model selects c_M (which contains both I_1 and I_2 information), this indicates that both images influence the model and the model fails to isolate the target image content. For example, suppose c_T is "There is a table below someone." and c_D is "There is someone below a table." Then, c_M becomes "There is a table below someone, and there is someone below a table." This merged caption implies that there are two people, one on the table and one under the table, which is factually incorrect if each image depicts only one of the two situations.

Quantify Information Leakage. We quantify the amount of cross-image information leakage based on the model's tendency to select the merged captions. Let $\mathcal{D}_\text{s}$ and $\mathcal{D}_\text{m}$ denote the sets of model predictions under the single-image and multi-image settings, respectively. We define the selection ratio of c_M in the single-image setting (R_s) and the multi-image setting (R_m) as:

$$R_\text{s} = \frac{1}{|\mathcal{D}_\text{s}|}\sum_{i=1}^{|\mathcal{D}_\text{s}|}\mathbf{1}\left[y_i^\text{s} = c_M\right], \quad R_\text{m} = \frac{1}{|\mathcal{D}_\text{m}|}\sum_{i=1}^{|\mathcal{D}_\text{m}|}\mathbf{1}\left[y_i^\text{m} = c_M\right], \tag{6}$$

where y_i^s and y_i^m denote the model's selected option for instance i under the single-image and multi-image conditions, respectively, and $\mathbf{1}[\cdot]$ is the indicator function. In practice, we use multiple-choice questions where each option corresponds to "A", "B", and "C", respectively.

Finally, the information leakage score C is defined as:

$$C = R_\text{m} - R_\text{s}. \tag{7}$$

A high information leakage score C indicates that the model is more likely to select the merged caption c_M in the multi-image setting, implying that it struggles to disentangle image-specific visual content and instead generates responses influenced by merged information of I_1 and I_2.

Dataset	sim_CLIP	Acc_s.	Acc_m	R_s	R_m	C
Winoground	0.74	71.0	51.5	9.9	**19.3**	**9.4**
VisMin	0.91	89.9	45.2	7.7	**29.9**	**22.2**

Table 1: Cross-image information leakage analysis for Qwen2.5-VL-3B on Winoground and VisMin. We report accuracy in single-image (Acc_s) vs. multi-image (Acc_m) settings (higher is better) as well as CLIP similarity (sim_CLIP) between two given images (I_1 and I_2) (higher denotes more similar images). We also report the ratio of the merged caption (c_M) selection for single-image and multi-image defined in Equation (6) (R_s and R_m, respectively). Finally, the information leakage score C (Equation (7)) is shown (higher indicates a more severe leakage).

As shown in Table 1, model accuracy consistently declines when single-image to the multi-image setting, while the frequency of selecting the merged caption (c_M) increases. In both datasets, R_m exceeds R_s, indicating that the model increasingly favors merged descriptions when multiple images are presented—suggesting a tendency to confuse visual information across inputs.

This information leakage appears more prominent in the VisMin dataset (showing larger C), which shows a higher average CLIP similarity between paired images (0.91 vs. 0.74 in Winoground). Namely, when the images share more semantic overlap, the information leakage becomes more severe. These observations indicate that cross-image information leakage tends to increase as the visual or semantic similarity between input images grows.

Together, these findings empirically demonstrate that LVLMs are susceptible to cross-image information leakage in multi-image settings, and that the severity of this interference correlates with image similarity.

Experiments

Benchmarks for Multi-image Understanding

Multi-image understanding has been a fundamental challenge in visual reasoning (Suhr et al. 2017, 2018). We employ four benchmarks to evaluate our method.

Winoground (Thrush et al. 2022) evaluates whether a model can correctly associate captions with images or vice versa. Each instance presents two images and two captions $(I_1, T_1), (I_2, T_2)$ with subtle differences, and the model must assign the correct caption to each image.

VisMin (Awal et al. 2024) builds on this setup by focusing on image-text pairs with minimal semantic differences. It adopts the same evaluation format as Winoground, but targets finer-grained discrimination.

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. (1) **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. (2) **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. (3) **Group Score**: A stricter metric that 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.

Mantis-Eval benchmark is released with the Mantis model, covering 217 diverse topics such as size perception and weight comparison. Each sample is carefully crafted by annotators to require deep cross-image understanding. It includes both multiple-choice and short-answer formats.

MuirBench is a challenging benchmark that assesses 12 multi-image reasoning skills, including geographic understanding, diagram interpretation, and visual retrieval. It covers 10 types of inter-image relations (e.g., narrative, complementary) and pairs each instance with a minimally altered, unanswerable variant to test fine-grained discrimination. Mantis-Eval and MuirBench are evaluated by accuracy.

Implementation Details

Models. We evaluate the generalizability of our method on three representative LVLM families: InternVL3 (2B, 8B) (Chen et al. 2024a), Qwen2.5-VL (3B, 7B) (Bai et al. 2025), 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.

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 λ is a hyperparameter that controls the degree of corruption. The optimal noise type and scale are selected via validation.

Hyperparameter Tuning. All hyperparameters including the 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 reserve 10% of the test data for this purpose.

More Details. All experiments are conducted using NVIDIA A100 GPUs. Most evaluations are performed on a single A100. For memory efficiency, we resize all MuirBench images to 512×512.

Quantitative Results

We report test split performance on each benchmark.

Method	InternVL3-2B			InternVL3-8B		
	T	I	G	T	I	G
Baseline	47.25	6.25	4.50	**70.75**	40.75	34.25
+*FOCUS*	47.25	**<u>27.25</u>**	**<u>19.75</u>**	69.25	**42.25**	**35.75**
Method	**Qwen2.5-VL-3B**			**Qwen2.5-VL-7B**		
	T	I	G	T	I	G
Baseline	56.25	36.75	26.00	74.50	39.75	34.00
+*FOCUS*	56.25	36.50	26.00	74.50	**<u>58.50</u>**	**<u>50.75</u>**
Method	**LLaVA-OV-0.5B**			**LLaVA-OV-7B**		
	T	I	G	T	I	G
Baseline	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 2: Winoground performance comparison across model families and sizes with and without **FOCUS**. T: Text Score, I: Image Score, G: Group Score. Bold: improved over baseline. Underline: largest gain within model group.

Method	InternVL3-2B			InternVL3-8B		
	T	I	G	T	I	G
Baseline	**79.39**	18.40	17.76	**88.74**	66.25	63.20
+*FOCUS*	79.10	**<u>50.55</u>**	**<u>47.61</u>**	88.62	**67.85**	**64.86**
Method	**Qwen2.5-VL-3B**			**Qwen2.5-VL-7B**		
	T	I	G	T	I	G
Baseline	84.76	37.41	34.57	89.11	72.95	69.26
+*FOCUS*	**85.13**	**42.89**	**40.55**	**89.16**	**77.01**	**72.85**
Method	**LLaVA-OV-0.5B**			**LLaVA-OV-7B**		
	T	I	G	T	I	G
Baseline	11.34	14.69	1.03	**87.02**	53.70	49.54
+*FOCUS*	**11.59**	**22.71**	**4.51**	86.93	**61.01**	**55.70**

Table 3: Performance comparison across model families and sizes with and without **FOCUS** on VisMin benchmark. T: Text Score, I: Image Score, G: Group Score. Bold indicates improvement over baseline. Underline highlights the largest gain within each model group.

Result on Winoground. Table 2 reports Winoground performance across three model famillies and scales, comparing baseline decoding with FOCUS. Results are measured identically with VisMin, reflecting the model's ability to correctly match image and the text in multi-image setting. In InternVL3-2B, FOCUS achieves a substantial gain in Image Score from 6.25 to 27.25. For LLaVA-OV-7B, the Group Score improves from 33.00 to 42.25, and the Image Score increases by nearly 12%pts, confirming that FOCUS also could enhance multi-image understanding performance in models with larger parameter counts.

Result on VisMin. Table 3 shows the performance of FOCUS compared to baseline decoding using various models.

Model Family	InternVL		Qwen2.5-VL		LLaVA-OV	
Size	2B	8B	3B	7B	0.5B	7B
Baseline	49.77	64.98	55.30	70.05	36.41	57.14
+ *FOCUS*	**52.53**	**65.44**	**58.99**	70.05	**41.94**	**59.91**

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

Model Family	InternVL		Qwen2.5-VL		LLaVA-OV	
Size	2B	8B	3B	7B	0.5B	7B
Baseline	28.42	31.62	30.38	29.92	22.85	28.88
+ *FOCUS*	27.46	**31.88**	**31.31**	29.73	**24.38**	**29.73**

Table 5: MuirBench accuracy across model families and sizes, with and without **FOCUS**. MuirBench includes fine-grained and unanswerable multi-image tasks. Bold indicates improvement over the baseline.

Across all configurations, FOCUS consistently improves the Image Score (I) and Group Score (G), which directly measure multi-image reasoning and balanced performance for single and multi-image reasoning. The improvement is especially pronounced in smaller models, even in the same LVLM family. For instance, as shown in the InternVL3-2B case, FOCUS boosts the Image Score from 18.40 to 50.55 and the Group Score from 17.76 to 47.61, representing the largest gains across all settings. These results suggest that smaller models benefit significantly from inference-time focused decoding of visual inputs.

Result on Mantis-Eval. Table 4 shows Mantis-Eval accuracy for six model variants. Across all settings, applying FOCUS yields modest but consistent improvements.

For InternVL, the 2B model sees a +2.76%pts gain, while the 8B model shows a smaller improvement. Qwen2.5-VL-3B gains +3.69%pts. In the case of LLaVA-OV, the 0.5B variant benefits notably, improving from 36.41 to 41.94. These results indicate that while the overall accuracy improvements are modest in scale, they are meaningful given the challenging nature of the benchmark. FOCUS consistently improves or maintains performance across diverse model architectures and sizes.

Result on MuirBench. Table 5 shows the performance of models on MuirBench, a challenging benchmark for multi-image reasoning. FOCUS yields modest improvements across several configurations: Qwen2.5-VL-3B +0.93%pts, LLaVA-OV-0.5B +1.53%pts, and LLaVA-OV-7B +0.85%pts. While overall gains are limited, the results suggest that FOCUS offers incremental benefits even in complex and ambiguous multi-image settings.

Qualitative Results

Figure 3 compares baseline decoding with FOCUS. In (a), the baseline incorrectly mentions both bananas and oranges, regardless of which image is queried. FOCUS, in contrast, correctly describes only bananas for the first image and only oranges for the second, showing a clear separation of visual content. In (b), the baseline confuses elements from both images, referring to both doughnuts and sandwiches even when asked about just one. FOCUS avoids this confusion and generates precise, image-specific descriptions. These examples highlight how baseline decoding struggles to disentangle visual signals across multiple images, whereas FOCUS effectively suppresses cross-image information leakage.

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

Table 6: Impact of FOCUS design choice: noise type, masking strength λ, and contrastive weight α. The Mantis validation accuracies using Qwen2.5-VL-7B are shown.

Ablation Study

We conduct an ablation study 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 6 (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 6 (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.

Effect of α (Weight). Table 6 (c) shows how varying the contrastive weight α subtracting 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.

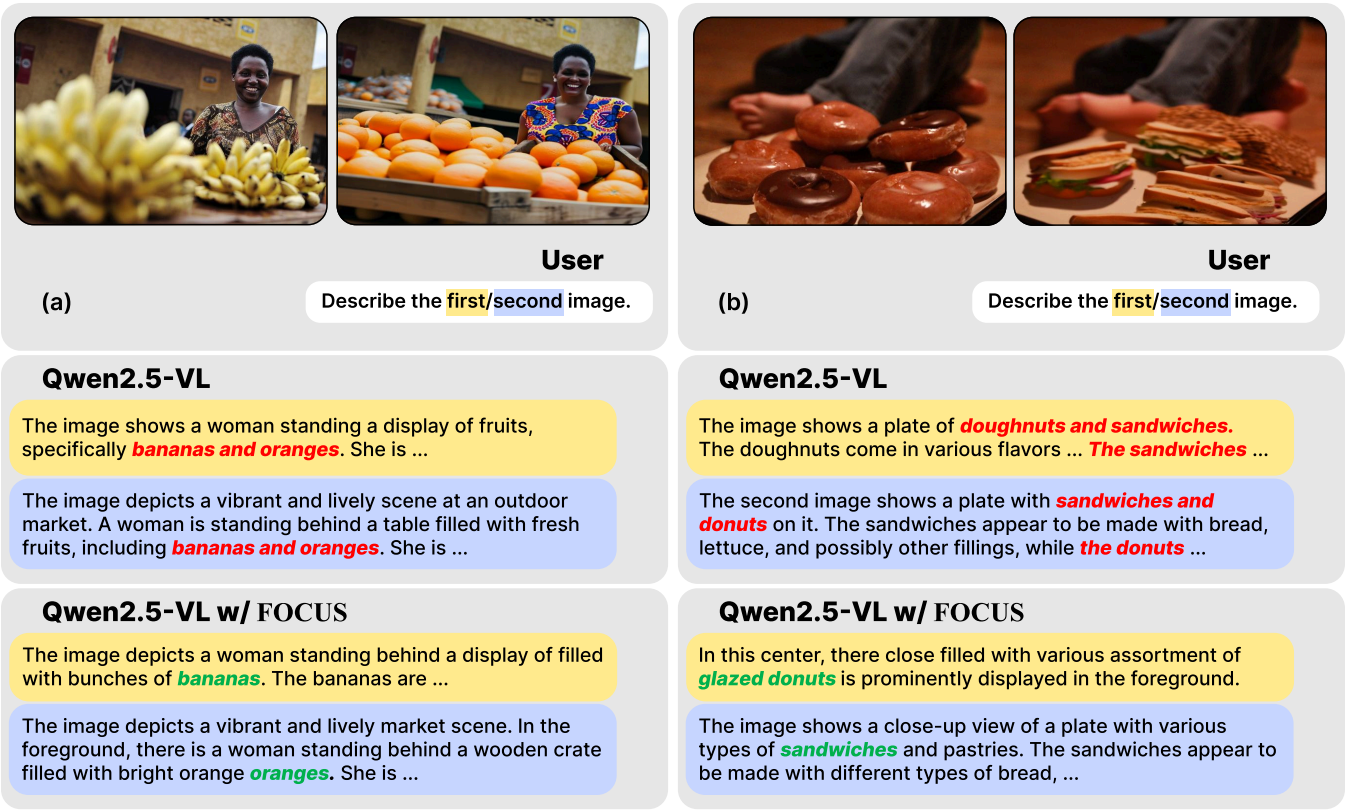

Figure 3: Qualitative samples using the Qwen2.5-VL 3B model. In both multi-image settings, the baseline decoding strategy often produces a mixed information of the other image not indicated by the question. On the other hand, FOCUS disentangles cross-image information well, resulting in a better multi-image understanding.

Comparison with a Noise-based Decoding Method

In this section, we compare VCD (Leng et al. 2024) with our method on the multi-image understanding task. VCD was originally proposed to mitigate hallucinations induced by language priors, but among decoding-time methods, it is the most comparable to ours. Specifically, VCD shares some similarity with our method in that it also leverages noise-masked images. However, the fundamental motivations of the two approaches are entirely different. While VCD aims to reduce hallucinated outputs by introducing noise to counteract language-driven biases, FOCUS is designed to enhance multi-image understanding by forcing the model to attend to one image at a time.

Since VCD was developed for single-image understanding, we applied a slight modification to adapt it to multi-image scenarios. Formally, the computation can be expressed as: $f_{\text{orig}} + \alpha \cdot (f_{\text{orig}} - f_{\text{noise}})$, where f_{orig} denotes the logits from the original multi-image input, and f_{noise} denotes the logits from the same input with noise applied. This calculation is performed using our f_{noise} interface while retaining VCD's original noise type (e.g., diffusion noise).

As shown in Table 7, our method consistently outperforms the modified VCD across all multi-image benchmarks in terms of Image Score. This highlights the limitation of directly applying a decoding method designed for single-image inputs to multi-image reasoning tasks.

Model Size	Method	Winoground	VisMin
Qwen2.5-VL-3B	VCD variant	32.5	29.3
	FOCUS (ours)	**45.0**	**44.2**
Qwen2.5-VL-7B	VCD variant	35.0	59.5
	FOCUS (ours)	**60.0**	**71.0**

Table 7: Comparison of image scores between our method and a VCD-style extension on a multi-image reasoning task.

Conclusion

This paper identifies underexplored limitation of LVLMs in multi-image settings: *cross-image information leakage* problem, where a model fails to disentangle visual information from multiple images. We propose a novel training-free decoding strategy that mitigates the problem without any architecture modification or additional training. Our method introduces a noise-guided image focusing decoding, whereby non-target images are corrupted to encourage the model to focus on a single image at a time. The resulting logits are then refined and aggregated through a contrastive logit aggregation scheme, effectively isolating information from

each image while preserving positional context. We demonstrate the effectiveness of our method on four multi-image benchmarks using three LVLM families, showing consistent improvements in multi-image reasoning. Ablation studies confirm the impact of noise design, masking strength, and contrastive weighting. Without altering model weights, our method offers a practical and generalizable solution.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*. 1

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433. 1

Awadalla, A.; Gao, I.; Gardner, J.; Hessel, J.; Hanafy, Y.; Zhu, W.; Marathe, K.; Bitton, Y.; Gadre, S.; Sagawa, S.; et al. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*. 1, 2

Awal, R.; Ahmadi, S.; Zhang, L.; and Agrawal, A. 2024. Vismin: Visual minimal-change understanding. *Advances in Neural Information Processing Systems*, 37: 107795–107829. 2, 4, 5

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*. 1, 5

Chen, J.; Zhang, T.; Huang, S.; Niu, Y.; Zhang, L.; Wen, L.; and Hu, X. 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)*, 4209–4221. 2

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; Li, B.; Luo, P.; Lu, T.; Qiao, Y.; and Dai, J. 2024a. 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)*, 24185–24198. 1, 2, 5

Chen, Z.; Zhao, Z.; Luo, H.; Yao, H.; Li, B.; and Zhou, J. 2024b. HALC: Object Hallucination Reduction via Adaptive Focal-Contrast Decoding. In *Forty-first International Conference on Machine Learning*. 2

Dai, W.; Li, J.; Li, D.; Tiong, A.; Zhao, J.; Wang, W.; Li, B.; Fung, P. N.; and Hoi, S. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36: 49250–49267. 1, 2

Dong, X.; Dong, S.; Wang, J.; Huang, J.; Zhou, L.; Sun, Z.; Jing, L.; Lan, J.; Zhu, X.; and Zheng, B. 2025. INTER: Mitigating Hallucination in Large Vision-Language Models by Interaction Guidance Sampling. *arXiv preprint arXiv:2507.05056*. 2

Herdade, S.; Kappeler, A.; Boakye, K.; and Soares, J. 2019. Image captioning: Transforming objects into words. *Advances in neural information processing systems*, 32. 1

Huang, Q.; Dong, X.; Zhang, P.; Wang, B.; He, C.; Wang, J.; Lin, D.; Zhang, W.; and Yu, N. 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*, 13418–13427. 2

Jiang, D.; He, X.; Zeng, H.; Wei, C.; Ku, M.; Liu, Q.; and Chen, W. 2024. Mantis: Interleaved multi-image instruction tuning. *arXiv preprint arXiv:2405.01483*. 1, 2

Laurençon, H.; Saulnier, L.; Tronchon, L.; Bekman, S.; Singh, A.; Lozhkov, A.; Wang, T.; Karamcheti, S.; Rush, A.; Kiela, D.; et al. 2023. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36: 71683–71702. 1, 2

Leng, S.; Zhang, H.; Chen, G.; Li, X.; Lu, S.; Miao, C.; and Bing, L. 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*, 13872–13882. 2, 7

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; et al. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*. 1, 2, 5

Li, F.; Zhang, R.; Zhang, H.; Zhang, Y.; Li, B.; Li, W.; Ma, Z.; and Li, C. 2024b. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*. 2

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, 19730–19742. PMLR. 1

Li, X. L.; Holtzman, A.; Fried, D.; Liang, P.; Eisner, J.; Hashimoto, T.; Zettlemoyer, L.; and Lewis, M. 2023b. Contrastive Decoding: Open-ended Text Generation as Optimization. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12286–12312. Toronto, Canada: Association for Computational Linguistics. 2

Lin, J.; Yin, H.; Ping, W.; Molchanov, P.; Shoeybi, M.; and Han, S. 2024. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 26689–26699. 1, 2

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge. 2

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916. 1, 2

Malkin, N.; Wang, Z.; and Jojic, N. 2022. Coherence boosting: When your pretrained language model is not paying enough attention. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting*

of the Association for Computational Linguistics (Volume 1: Long Papers), 8214–8236. Dublin, Ireland: Association for Computational Linguistics. 2

Park, Y.; Lee, D.; Choe, J.; and Chang, B. 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, 6434–6442. 2

Shi, W.; Han, X.; Lewis, M.; Tsvetkov, Y.; Zettlemoyer, L.; and Yih, W.-t. 2024. Trusting Your Evidence: Hallucinate Less with Context-aware Decoding. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, 783–791. Mexico City, Mexico: Association for Computational Linguistics. 2

Suhr, A.; Lewis, M.; Yeh, J.; and Artzi, Y. 2017. A Corpus of Natural Language for Visual Reasoning. In Barzilay, R.; and Kan, M.-Y., eds., *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 217–223. Vancouver, Canada: Association for Computational Linguistics. 4

Suhr, A.; Zhou, S.; Zhang, A.; Zhang, I.; Bai, H.; and Artzi, Y. 2018. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*. 4

Sun, Q.; Cui, Y.; Zhang, X.; Zhang, F.; Yu, Q.; Wang, Y.; Rao, Y.; Liu, J.; Huang, T.; and Wang, X. 2024. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14398–14409. 1, 2

Suo, W.; Zhang, L.; Sun, M.; Wu, L. Y.; Wang, P.; and Zhang, Y. 2025. Octopus: Alleviating Hallucination via Dynamic Contrastive Decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 29904–29914. 2

Thrush, T.; Jiang, R.; Bartolo, M.; Singh, A.; Williams, A.; Kiela, D.; and Ross, C. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5238–5248. 2, 4

Tian, X.; Zou, S.; Yang, Z.; and Zhang, J. 2025. Identifying and mitigating position bias of multi-image vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 10599–10609. 2, 3

Wang, C.; Chen, X.; Zhang, N.; Tian, B.; Xu, H.; Deng, S.; and Chen, H. 2025a. MLLM can see? Dynamic Correction Decoding for Hallucination Mitigation. In *The Thirteenth International Conference on Learning Representations*. 2

Wang, F.; Fu, X.; Huang, J. Y.; Li, Z.; Liu, Q.; Liu, X.; Ma, M. D.; Xu, N.; Zhou, W.; Zhang, K.; Yan, T. L.; Mo, W. J.; Liu, H.-H.; Lu, P.; Li, C.; Xiao, C.; Chang, K.-W.; Roth, D.; Zhang, S.; Poon, H.; and Chen, M. 2025b. MuirBench: A Comprehensive Benchmark for Robust Multi-image Understanding. In *The Thirteenth International Conference on Learning Representations*. 1, 2

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2024. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models. In *The Twelfth International Conference on Learning Representations*. 2