CONTENT-BOUNDED PHASE CONSENSUS: MITIGATING POSITION-DRIVEN SOURCE PREFERENCE IN MULTI-IMAGE VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-image vision–language models (VLMs) must associate visual evidence with its source, yet their predictions can depend on source position. We study this behavior through Boundary Keys, the Key representations of model-native tokens delimiting visual sources. Across representative models from Qwen3-VL, InternVL3.5, and Molmo2-O, identical visual inputs exhibit position-dependent source attention, while early-layer Boundary Keys contain a dominant source-shared component. RoPE maps this shared component to source-dependent realizations, and counterfactual phase reassignment changes boundary source preference even when native source-specific residuals are retained. Motivated by this analysis, we propose Content-Bounded Phase Consensus (CBPC), a lightweight training-free intervention that selectively contracts the RoPE-induced phase deviation while preserving the cross-source mean. CBPC derives an input- and layer-dependent correction strength from native visual-key-centroid statistics without additional parameters, answer labels, or strength search. Across Mantis, MuirBench, MIRB, and QBench2, CBPC yields consistent gains, improves correctness under meaning-preserving image reorderings, and incurs modest LM forward overhead with negligible additional peak memory.

1 INTRODUCTION

Large vision–language models (LVLMs) increasingly process multiple or interleaved visual sources within a single context, enabling cross-image comparison, reference resolution, and multi-image reasoning (Alayrac et al., 2022; Zhao et al., 2024b; Jiang et al., 2024). Recent model families extend these capabilities across architectures and scales (Bai et al., 2025; Wang et al., 2025b; Clark et al., 2026). Such settings require models to maintain the association between visual evidence and its originating source, yet benchmarks and controlled studies report failures under diverse inter-image relations, visual distractors, long multimodal contexts, and increasing numbers of tracked visual concepts (Wang et al., 2025a; Zhao et al., 2024a; Liu et al., 2024a; Meng et al., 2025; Xia et al., 2025; Song et al., 2024; Sharma et al., 2024; Wu et al., 2025; Das et al., 2026). Predictions can also depend on where relevant evidence appears, even when image reordering preserves the question semantics and source references.

Recent approaches mitigate multi-image failures at different levels of cross-image processing. SoFA modifies inter-image attention connectivity, FOCUS combines predictions from image-focused contexts, and Delimiter Token Scaling (DTS) rescales native image-delimiter representations (Tian et al., 2025; Park et al., 2025; Lee et al., 2026). We instead ask whether a localized position-dependent contribution can be isolated within the model's native boundary representations. Because boundary tokens recur in the structural role of delimiting visual sources, their Key representations provide a natural site for examining source–position interactions. We call these representations *Boundary Keys* and analyze how their geometry interacts with rotary positional encoding (RoPE) (Su et al., 2021).

Our analysis of Qwen3-VL-8B, InternVL3.5-8B, and Molmo2-O-7B reveals three connected findings. First, duplicating the same image across source positions still produces asymmetric source-conditional attention. Second, early-layer pre-RoPE Boundary Keys exhibit a dominant source-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

shared component together with comparatively small source-specific residuals. Third, although the source-shared component is identical across sources before RoPE, counterfactually reassigning its position-dependent rotations changes total-Key boundary source preference even when native queries and source-specific residuals are retained. Together, these results identify the RoPE-induced phase deviation as a controllable contributor to boundary source preference, without implying that the source-shared component itself should be removed.

Motivated by this analysis, we propose **Content-Bounded Phase Consensus (CBPC)**, a deliberately lightweight, training-free intervention on Boundary Keys. CBPC selectively contracts only the RoPE-induced phase deviation of the source-shared component while preserving its cross-source mean and leaving source-specific residuals directly unchanged. At each selected layer, an input- and layer-dependent correction strength is obtained in closed form from the source-wise query–key (QK) variability of native visual-key centroids. The resulting rule requires no additional parameters, answer labels, or search over correction strengths, and does not directly modify Values.

We evaluate CBPC on nine model configurations from Qwen3-VL, InternVL3.5, and Molmo2-O across Mantis, MuirBench, MIRB, and QBench2 (Jiang et al., 2024; Wang et al., 2025a; Zhao et al., 2024a; Zhang et al., 2024). CBPC yields consistent improvements over vanilla inference, with an equally weighted average gain of 1.36 percentage points, while remaining competitive with or stronger than the evaluated training-free baselines. Under meaning-preserving cyclic reorderings of Mantis, it increases all-orders-correct accuracy by 1.84–3.23 points across three representative models. Controlled ablations support the phase-selective and content-referenced design, while CBPC incurs modest LM forward overhead with negligible additional peak memory.

Our contributions are threefold:

- We identify a dominant source-shared component in Boundary Keys and show that reassigning its RoPE phases modulates total-Key boundary source preference even when native source-specific residuals are retained.
- We introduce CBPC, a lightweight, mean-preserving contraction of the RoPE-induced phase deviation with input- and layer-dependent strength derived from native visual-key-centroid statistics.
- We demonstrate consistent gains across model families, scales, and multi-image benchmarks, together with improved correctness across meaning-preserving image reorderings and controlled evidence supporting the phase-selective and content-referenced design.

2 Related Work

Multi-image LVLMs and evaluation. Flamingo enables interleaved visual–text contexts, while MMICL and Mantis develop multimodal in-context learning and multi-image instruction following (Alayrac et al., 2022; Zhao et al., 2024b; Jiang et al., 2024). Recent LVLM families, including Qwen3-VL, InternVL3.5, and Molmo2, extend these capabilities across architectures and scales (Bai et al., 2025; Wang et al., 2025b; Clark et al., 2026). Evaluation has similarly expanded: MuirBench and MIRB cover diverse multi-image relations and reasoning, QBench2 evaluates pairwise low-level visual comparison, and MIBench, MMIU, MMIE, and VLM2-Bench examine complementary aspects of multi-image instruction following, inter-image relationships, interleaved comprehension, and visual correspondence (Wang et al., 2025a; Zhao et al., 2024a; Zhang et al., 2024; Liu et al., 2024a; Meng et al., 2025; Xia et al., 2025; Zhang et al., 2025). Long-context studies further expose failures under visual distractors and extended multimodal contexts, while controlled analyses show degradation as more visual concepts must be tracked simultaneously (Sharma et al., 2024; Song et al., 2024; Wu et al., 2025; Das et al., 2026). Our work complements these task-level analyses by examining a specific RoPE-induced, position-dependent component of model-native Boundary Keys.

Position bias and multi-image interventions. Source-order and target-position sensitivity have been documented in multimodal models (Tan et al., 2024; Tian et al., 2025). SoFA mitigates position bias by interpolating between causal and bidirectional inter-image attention (Tian et al., 2025), while CAPL combines cross-image attention calibration with preference optimization (Yang et al., 2026). At the positional-encoding level, Circle-RoPE and related work modify multimodal posi-

tional geometry or frequency allocation (Wang et al., 2026; Huang et al., 2026). Other methods intervene directly at inference time: FOCUS constructs image-focused contexts and combines their output logits (Park et al., 2025), whereas DTS rescales native delimiter hidden states to strengthen image separation (Lee et al., 2026). CBPC instead retains the native attention mask and RoPE operator and selectively contracts the RoPE-induced phase deviation of the source-shared Boundary-Key component using an input- and layer-dependent visual reference.

3 Analysis of Boundary Source Preference

We analyze boundary source preference in Qwen3-VL-8B, InternVL3.5-8B, and Molmo2-O-7B. Our analysis proceeds in three steps: we first test whether source attention is position-sensitive under identical visual content, then characterize the geometry of Boundary Keys before RoPE, and finally isolate how the RoPE phases of their source-shared component affect total-Key boundary source preference.

3.1 Identical Visual Inputs Exhibit Position-Dependent Attention

We first ask whether source attention remains symmetric when visual content is identical. Using 100 single-image MMBench validation examples (Liu et al., 2024b), we duplicate each image across four source positions with a source-neutral prompt. For token class $\tau \in \{B, V\}$, let $\mathcal{T}_s^\tau$ denote the boundary or visual-token positions of source s. We define source-conditional attention as

$$P_s^\tau = \frac{\sum_{t \in \mathcal{T}_s^\tau} \exp(a_t)}{\sum_j \sum_{t \in \mathcal{T}_j^\tau} \exp(a_t)}, \tag{1}$$

where a_t is the native attention logit. Figure 1 shows that the last source is selected as top-1 substantially more often than the 25% uniform reference for both boundary and visual tokens across all three models. Thus, identical visual inputs can still induce position-dependent source attention, motivating analysis of the associated Boundary Keys.

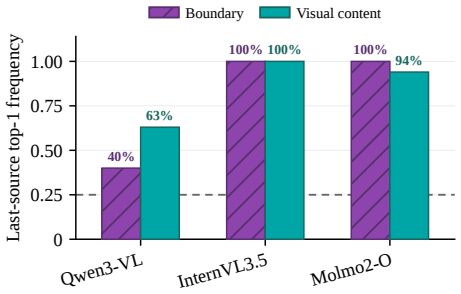


Figure 1: **Position-dependent attention under identical visual inputs.** Last-source top-1 frequency over 100 four-source duplicate-image inputs. The dashed line denotes the 25% uniform reference.

3.2 Boundary Keys Exhibit Dominant Source-Shared Geometry

We next examine pre-RoPE Boundary Keys at the first decoder layer using 200 natural multi-image examples from Mantis and MuirBench (Jiang et al., 2024; Wang et al., 2025a). Here, a boundary role denotes a model-specific delimiter type serving the same structural function across visual sources. For a fixed boundary role and KV head, we decompose the Boundary Key of source s as

$$k_s = c + r_s, \quad c = \frac{1}{S}\sum_j k_j, \quad r_s = k_s - c, \tag{2}$$

where c is the *source-shared component* and r_s is the *source-specific residual*. Figure 2 shows strong cross-source alignment, with mean cosine similarities above 0.97 in all three models, while the residual-to-shared energy ratio remains approximately 1%. These results identify a dominant source-shared geometry in the measured pre-RoPE Boundary Keys. They do not imply that the source-specific residual is functionally unimportant; instead, they motivate testing whether the source-shared component itself can induce source preference after positional encoding.

3.3 RoPE Phases Modulate Boundary Source Preference

Although c is shared before RoPE, its post-RoPE realization depends on the source boundary position p_s. We write

$$u_s = R(p_s)c, \qquad m = \frac{1}{S}\sum_j u_j, \qquad d_s = u_s - m, \tag{3}$$

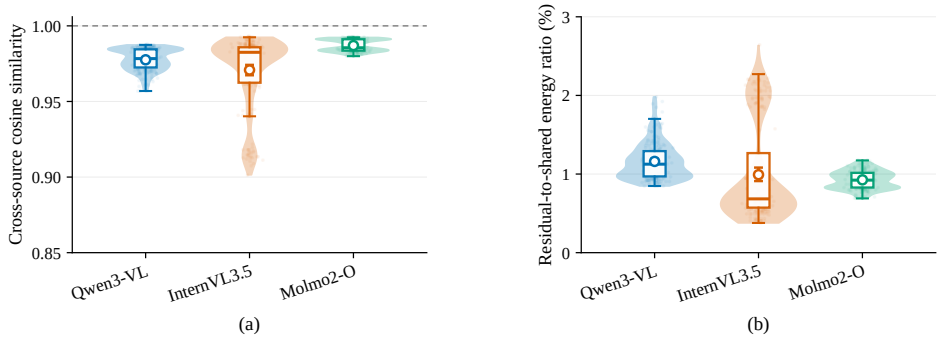


Figure 2: **Pre-RoPE Boundary-Key geometry.** Measurements use 200 natural multi-image inputs from Mantis and MuirBench. (a) Mean cross-source cosine similarity. (b) Residual-to-shared energy ratio. Violins show sample-level distributions, boxes denote interquartile ranges, and error bars denote 95% bootstrap confidence intervals for the mean.

where m is the *cross-source mean* and d_s is the *RoPE-induced phase deviation*, subsequently shortened to *phase deviation*. The native post-RoPE Boundary Key is therefore $K_s = R(p_s)r_s + m + d_s$. Because $q^\top m$ is source-independent for a fixed query q, only the phase deviation can create source-wise variation from the source-shared component.

Reassigning phase changes the preferred source. We perform local counterfactuals that keep the visual inputs, source order, native queries, and source-specific residuals fixed while permuting only the phases applied to the source-shared component:

$$K_s^{\text{shared}}(\pi) = R(p_{\pi(s)})c, \qquad K_s^{\text{total}}(\pi) = R(p_{\pi(s)})c + R(p_s)r_s. \tag{4}$$

On the same 200 natural multi-image examples, shared-only preferred-source agreement with the applied phase permutation is 100% for all three models, serving as a sanity check. With native source-specific residuals retained, total-Key agreement remains 86.5% for Qwen3-VL and 100% for InternVL3.5 and Molmo2-O (Figure 3(a)). Thus, phase assignment remains a strong contributor to total-Key boundary source preference.

Shared-only phase selection modulates total-Key routing margins. We further quantify the effect using 100 controlled four-source MMBench inputs with a designated target source. For each input, we enumerate all 24 source-phase permutations. Let $Z_s^\chi(\pi)$ denote the source score aggregated over decoder layers 0 and 1 and KV heads, where $\chi \in \{\text{shared}, \text{total}\}$, and define the *routing margin* as $Z_{\text{route}}^\chi(\pi) = Z_t^\chi(\pi) - \max_{j \neq t} Z_j^\chi(\pi)$. We select target-favoring and target-disfavoring assignments using only shared-component scores:

$$\pi_+ = \arg\max_\pi Z_{\text{route}}^{\text{shared}}(\pi), \qquad \pi_- = \arg\min_\pi Z_{\text{route}}^{\text{shared}}(\pi). \tag{5}$$

We then evaluate their effect using total-Key scores:

$$\Delta Z_{\text{route}}^{\pm} = Z_{\text{route}}^{\text{total}}(\pi_\pm) - Z_{\text{route}}^{\text{total}}(\text{id}). \tag{6}$$

Figure 3(b) shows that favoring assignments increase the mean routing margin by 0.45–1.28, whereas disfavoring assignments decrease it by 0.20–0.54 across the three models; all six paired bootstrap 95% confidence intervals exclude zero. Because selection uses shared-only QK scores whereas evaluation retains the native source-specific residual, these results show that phase assignments selected from the source-shared component produce two-sided mean changes in total-Key boundary source preference even when native source-specific residuals are retained.

Together, these findings identify the phase deviation d_s as a localized, controllable contributor to boundary source preference. This motivates selectively contracting d_s while preserving the cross-source mean m and without directly modifying the source-specific residual r_s.

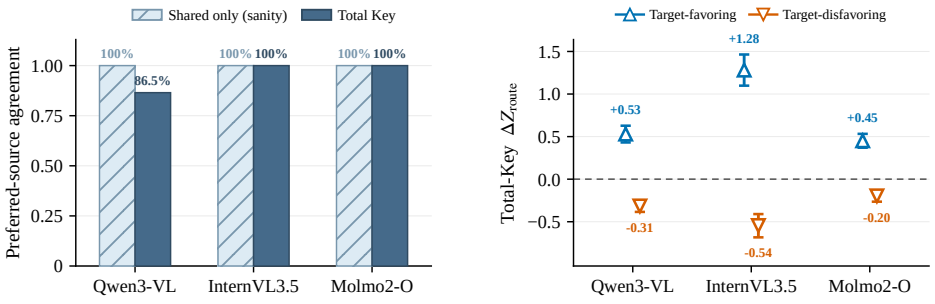


Figure 3: **Shared-component phase reassignment modulates boundary source preference.** (a) Preferred-source agreement with the applied phase permutation on 200 natural multi-image examples; shared-only serves as a sanity check, while total-Key retains native source-specific residuals. (b) Mean change in total-Key routing margin on 100 controlled four-source inputs. Target-favoring and target-disfavoring assignments are selected from all 24 phase permutations using shared-only QK scores and evaluated with native source-specific residuals retained. Error bars in (b) denote paired bootstrap 95% confidence intervals.

4 Content-Bounded Phase Consensus

Section 3 identifies the RoPE-induced phase deviation d_s as a controllable contributor to boundary source preference. We introduce Content-Bounded Phase Consensus (CBPC), which selectively contracts this phase deviation while preserving the cross-source mean m and leaving the source-specific residual r_s directly unchanged. A visual-key-centroid reference determines an input- and layer-dependent correction strength from the native QK statistics at each selected layer. Figure 4 summarizes the overall CBPC pipeline before we present the formal derivation.

4.1 Phase-Selective Consensus

Using the decomposition from Section 3, the native post-RoPE Boundary Key is $K_s = R(p_s)r_s + m + d_s$. Suppressing input, layer, KV-head, and boundary-role indices, CBPC defines the consensus path

$$K_s(\lambda) = K_s - \lambda d_s = R(p_s)r_s + m + (1-\lambda)d_s, \qquad 0 \le \lambda \le 1. \tag{7}$$

Here λ is the correction strength. Setting $\lambda = 0$ recovers the native Boundary Key, while $\lambda = 1$ yields full consensus of the source-shared component. Because $\sum_s d_s = 0$, the correction preserves the cross-source mean of the post-RoPE Boundary Keys. It therefore leaves the consensus term m and the source-specific residual directly unchanged and modifies only the phase deviation.

4.2 Visual-Key-Centroid Reference

For each input i and selected layer ℓ, CBPC independently determines how strongly the phase deviation should be contracted. Let $\mathcal{V}_{i,s}$ denote the visual-token positions belonging to source s. For each KV head, we form a *visual-key centroid* from the native post-RoPE visual Keys,

$$K^{\text{vis}}_{i,\ell,s} = \frac{1}{|\mathcal{V}_{i,s}|} \sum_{t \in \mathcal{V}_{i,s}} K_{i,\ell,t}. \tag{8}$$

Using post-RoPE queries from all prefill positions after the final image boundary, we measure source-wise variation in phase-deviation and visual-key-centroid scores:

$$B_{i,\ell} = \mathbb{E}_q \operatorname{Var}_s\left(q^\top d_{i,\ell,s}\right), \qquad C_{\text{vis},i,\ell} = \mathbb{E}_q \operatorname{Var}_s\left(q^\top K^{\text{vis}}_{i,\ell,s}\right). \tag{9}$$

At each selected layer, we average source variances over queries, KV heads, and boundary roles to obtain $(B_{i,\ell}, C_{\text{vis},i,\ell})$ and one correction strength $\lambda_{i,\ell}$. We refer to $C_{\text{vis},i,\ell}$ as the visual-key-centroid

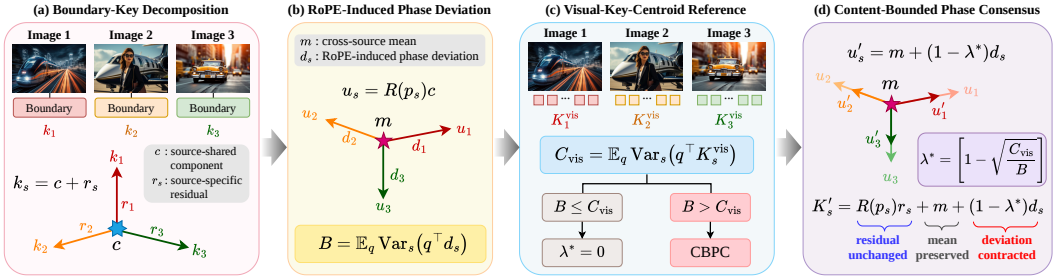


Figure 4: **Overview of Content-Bounded Phase Consensus (CBPC).** (a) Boundary Keys are decomposed into source-shared and source-specific components. (b) RoPE maps the source-shared component to source-dependent realizations, yielding phase deviations around their cross-source mean. (c) Source-wise visual-key centroids define a native reference for source-wise query–key variability. (d) CBPC uses the phase variation and visual reference to determine an input- and layer-dependent correction strength, contracting only the phase deviation while preserving the cross-source mean and leaving the source-specific residual directly unchanged.

reference, or visual reference. It provides an empirical scale for source-wise QK variability rather than a guarantee of semantic preservation.

Along Eq. 7, a common correction strength scales the phase-induced variation as

$$B_{i,\ell}(\lambda) = (1-\lambda)^2 B_{i,\ell}. \tag{10}$$

Let $A_{i,\ell} = \sum_{h,b,s} \|d_{i,\ell,h,b,s}\|_2^2$ denote the phase-deviation energy over KV heads h, boundary roles b, and sources s. We choose the smallest correction along the consensus path that brings the pooled phase variation to the visual-reference level:

$$\min_{\lambda \in [0,1]} \lambda^2 A_{i,\ell} \qquad \text{s.t.} \qquad (1-\lambda)^2 B_{i,\ell} \le C_{\text{vis},i,\ell}. \tag{11}$$

Since the objective is nondecreasing for $\lambda \ge 0$, the smallest feasible correction strength solves the unregularized problem:

$$\lambda^\star_{i,\ell} = \begin{cases} 0, & B_{i,\ell} \le C_{\text{vis},i,\ell}, \\ 1 - \sqrt{C_{\text{vis},i,\ell}/B_{i,\ell}}, & B_{i,\ell} > C_{\text{vis},i,\ell}. \end{cases} \tag{12}$$

Thus, the unregularized solution performs no correction when the measured phase variation is already below the visual reference and otherwise applies the minimum Euclidean displacement required along the prescribed consensus path. This minimum-displacement statement concerns the pooled current-layer statistics and this one-dimensional path, rather than arbitrary Key perturbations or separate constraints for individual heads and roles.

For numerical stability, the evaluated implementation uses

$$\lambda_{i,\ell} = \left[1 - \sqrt{\frac{C_{\text{vis},i,\ell}}{B_{i,\ell} + \epsilon}}\right]_+, \qquad [x]_+ = \max(x, 0), \tag{13}$$

where $\epsilon > 0$ stabilizes the division. We distinguish this implementation from the exact unregularized solution in Eq. 12; finite ϵ can introduce additional contraction. Neither form uses answer labels or a search over correction strengths.

Application and implementation. Let $\mathcal{L}$ denote the selected decoder layers. At each $\ell \in \mathcal{L}$, CBPC independently recomputes the Boundary-Key decomposition, phase deviation, visual reference, pooled statistics $(B_{i,\ell}, C_{\text{vis},i,\ell})$, and correction strength $\lambda_{i,\ell}$, and applies

$$K'_{i,\ell,s} = K_{i,\ell,s} - \lambda_{i,\ell} d_{i,\ell,s}, \qquad \ell \in \mathcal{L}. \tag{14}$$

During prefill, the equivalent Key correction is applied after native Q/K normalization and before RoPE and attention:

$$k'_{i,\ell,s} = k_{i,\ell,s} - \lambda_{i,\ell} R(p_s)^{-1} d_{i,\ell,s}, \qquad R(p_s) k'_{i,\ell,s} = K'_{i,\ell,s}. \tag{15}$$

Decoding reuses the corrected native KV cache without further intervention. Only Boundary Keys are directly modified; Values and native RoPE are retained, without Key-norm restoration.

Table 1: **Accuracy (%) across model families and scales.** Best results are **bolded**; tied best results are all bolded.

Benchmark	Method	Qwen3-VL				InternVL3.5				Molmo2-O
		2B	4B	8B	32B	2B	4B	8B	14B	7B
Mantis	Vanilla	49.31	63.59	67.74	72.35	55.30	64.52	64.98	68.20	61.75
	DTS	50.23	64.52	68.66	73.73	53.46	58.99	59.91	61.29	62.67
	FOCUS	43.32	63.59	64.98	72.81	56.68	61.75	64.06	68.66	62.67
	CBPC	**51.15**	**65.44**	**69.12**	**74.19**	**57.14**	**65.44**	**65.90**	**69.59**	**64.06**
MuirBench	Vanilla	37.65	52.42	51.58	56.19	42.65	52.96	55.19	53.92	47.85
	DTS	39.96	52.46	46.35	56.96	**43.54**	50.65	52.12	54.04	48.04
	FOCUS	30.73	50.77	49.19	56.77	38.42	39.42	49.15	42.50	43.85
	CBPC	**40.38**	**52.69**	**52.19**	**57.12**	43.50	**54.42**	**56.35**	**55.96**	**48.62**
MIRB	Vanilla	44.48	48.50	57.89	56.97	48.81	53.15	55.42	58.10	52.32
	DTS	45.30	48.71	58.20	57.17	49.54	53.97	52.63	58.72	52.63
	FOCUS	39.73	47.37	47.88	56.24	41.59	49.74	52.22	46.65	48.50
	CBPC	**47.47**	**49.02**	**58.72**	**57.59**	**50.05**	**56.55**	**56.14**	**59.55**	**53.35**
QBench2	Vanilla	61.20	75.80	76.80	80.10	65.00	73.80	75.80	78.10	74.60
	DTS	62.90	77.00	77.60	**81.20**	65.90	75.00	75.90	78.30	75.20
	FOCUS	59.00	73.60	73.10	79.40	65.50	74.80	74.80	77.80	72.70
	CBPC	**63.10**	**77.50**	**78.00**	**81.20**	**66.70**	**75.30**	**76.20**	**78.50**	**75.90**

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

We evaluate CBPC on nine model configurations from Qwen3-VL (Bai et al., 2025), InternVL3.5 (Wang et al., 2025b), and Molmo2-O (Clark et al., 2026) across Mantis (Jiang et al., 2024), MuirBench (Wang et al., 2025a), MIRB (Zhao et al., 2024a), and QBench2 (Zhang et al., 2024). We compare against vanilla inference, Delimiter Token Scaling (DTS) (Lee et al., 2026), and FOCUS (Park et al., 2025), following their released implementations. SoFA (Tian et al., 2025) is omitted because no public implementation was available at the time of our experiments. All methods use the same inputs, preprocessing, prompts, and decoding configuration within each model–benchmark setting. Unless otherwise specified, CBPC uses decoder layers 0 and 1 and independently recomputes its statistics and correction strength at each layer. Experiments use NVIDIA RTX A5000 GPUs with 24 GB memory. Additional benchmark and evaluation details are provided in the appendix.

5.2 MAIN RESULTS

Table 1 reports accuracy across the evaluated model families, scales, and benchmarks. CBPC consistently improves over vanilla inference, with an equally weighted average gain of 1.36 percentage points. The improvement is also stable across model families, with average gains of 1.39 points for Qwen3-VL, 1.34 for InternVL3.5, and 1.35 for Molmo2-O. This trend holds across benchmarks with different multi-image reasoning and visual-comparison demands, suggesting that the benefit is not confined to a particular architecture, model scale, or task type. CBPC also remains competitive with or stronger than the evaluated training-free baselines across the table.

5.3 ROBUSTNESS TO SOURCE REORDERING

Using all Mantis questions, we evaluate every cyclic shift of each image sequence while remapping image references in questions, choices, and answers to preserve the original semantics. This yields 540 meaning-preserving presentations. We report original-order accuracy, MeanAcc over all evaluated presentations, and OrderRobustAcc, the fraction of original questions answered correctly under every evaluated cyclic order.

Table 2: **Robustness to meaning-preserving cyclic image reordering on Mantis.** Orig., Mean, and Robust denote OriginalAcc, MeanAcc, and OrderRobustAcc (%). Best results are **bolded**; tied best results are all bolded.

	Qwen3-VL-8B			InternVL3.5-8B			Molmo2-O-7B		
Method	**Orig.**	**Mean**	**Robust**	**Orig.**	**Mean**	**Robust**	**Orig.**	**Mean**	**Robust**
Vanilla	67.74	64.81	59.91	64.98	63.89	56.22	61.75	61.85	52.53
DTS	68.66	65.00	59.91	59.91	58.52	47.00	62.67	62.22	53.00
FOCUS	64.98	61.85	55.30	64.06	62.78	57.14	62.67	61.85	51.15
Equal-phase	68.20	65.93	61.29	**65.90**	64.44	58.99	63.59	62.41	53.46
Full consensus	67.74	65.56	59.91	65.44	64.26	57.14	63.13	62.22	53.46
CBPC	**69.12**	**67.04**	**62.67**	**65.90**	**64.81**	**59.45**	**64.06**	**63.15**	**54.38**

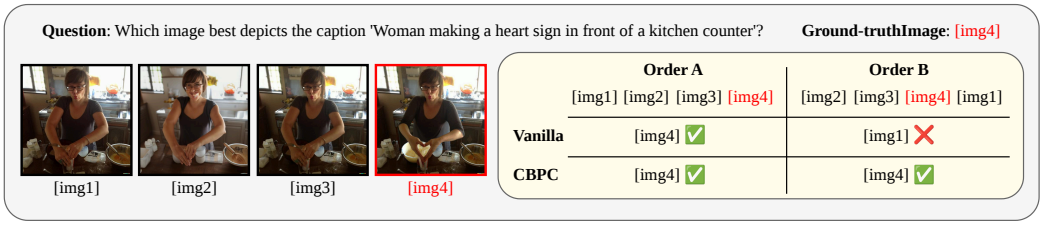


Figure 5: **Qualitative source-order example.** The same Mantis problem is shown under two meaning-preserving cyclic orders. Vanilla changes its selected image from `[img4]` to `[img1]`, whereas CBPC selects the target `[img4]` under both.

We compare CBPC with vanilla inference, DTS, FOCUS, and two phase controls that retain the source-specific residual. Equal-phase uses a common shared-component phase, $K_s^{\mathrm{eq}} = R(p_s)r_s + R(p_1)c$, eliminating source-dependent shared phase variation without preserving the original cross-source mean. Full consensus instead uses $K_s^{\mathrm{full}} = R(p_s)r_s + m$, corresponding to $\lambda = 1$ and preserving the cross-source mean.

Table 2 shows that CBPC obtains the highest MeanAcc and OrderRobustAcc for all three representative models. Relative to vanilla inference, OrderRobustAcc increases by 2.76, 3.23, and 1.84 points for Qwen3-VL-8B, InternVL3.5-8B, and Molmo2-O-7B, respectively, and exceeds equal-phase by 1.38, 0.46, and 0.92 points. Figure 5 illustrates this behavior in a concrete Mantis example, where vanilla changes its prediction after reordering while CBPC remains consistent across both presentations.

5.4 Design Ablations

We use Qwen3-VL-8B, InternVL3.5-8B, and Molmo2-O-7B on Mantis and QBench2 to examine the intervention direction and correction-strength rule. We further evaluate alternative intervention-layer schedules in Appendix A.2.2, where the default Early-2 configuration performs best across the evaluated settings.

Intervention direction. We compare phase-deviation-only correction at fixed strength $\lambda = 0.5$ with whole-Key centered correction and source-shared attenuation. For each input, layer, KV head, and boundary role, alternative directions are scaled to match the perturbation norm of the phase-deviation-only intervention, with a maximum observed relative mismatch of 2.502×10^{-7}. As shown in Table 3(a), phase-deviation-only correction is best or tied-best among the matched directions in five of six settings; source-shared attenuation is higher by 0.46 points on Molmo2-O/Mantis.

Correction strength. We next fix the intervention direction to $-d_s$ and compare CBPC with the best fixed coefficient over $\lambda \in \{0.25, 0.50, 0.75\}$, full consensus, and a shuffled-C_{vis} control. At each selected layer, the shuffled control retains the current input's phase-variation statistic $B_{i,\ell}$ but replaces its visual reference $C_{\mathrm{vis},i,\ell}$ with a shuffled reference from the same model, benchmark, and

Table 3: **Design ablations.** Accuracy (%) on Mantis and QBench2. In (a), bold indicates the best or tied-best result among the three energy-matched intervention directions; adaptive CBPC is shown as a non-matched reference. In (b), bold indicates the best result including ties, and shuffled-C_{vis} reports mean ± sample standard deviation over three shuffles.

Condition	Qwen3-VL-8B		InternVL3.5-8B		Molmo2-O-7B	
	Mantis	QBench2	Mantis	QBench2	Mantis	QBench2
(a) Intervention direction at matched perturbation energy						
Vanilla	67.74	76.80	64.98	75.80	61.75	74.60
Whole-Key centered	67.74	77.60	**64.98**	75.90	63.13	**75.70**
Source-shared attenuation	68.20	77.00	64.52	75.90	**63.59**	75.30
Phase deviation d_s-only	**68.66**	**77.90**	**64.98**	**76.00**	63.13	**75.70**
CBPC (adaptive)	69.12	78.00	65.90	76.20	64.06	75.90
(b) Correction strength with direction fixed to $-d_s$						
Vanilla	67.74	76.80	64.98	75.80	61.75	74.60
Fixed $\lambda = .25$	67.74	77.50	64.98	76.10	62.67	75.00
Fixed $\lambda = .50$	68.66	77.90	64.98	76.00	63.13	75.70
Fixed $\lambda = .75$	68.20	77.70	65.44	75.90	63.59	75.80
Full consensus	67.74	77.80	65.44	76.00	63.13	**75.90**
Shuffled-C_{vis}	67.90±0.70	77.70±0.10	64.52±0.46	75.97±0.06	63.29±0.27	75.73±0.12
CBPC	**69.12**	**78.00**	**65.90**	**76.20**	**64.06**	**75.90**

Table 4: **Inference efficiency.** LM forward time and peak allocated GPU memory on the fixed Mantis subset. Lower is better for both metrics.

Model	LM forward time (s/sample)		Peak memory (GiB)	
	Vanilla	CBPC	Vanilla	CBPC
Qwen3-VL-8B	1.046	1.133	16.9688	16.9688
InternVL3.5-8B	1.392	1.420	17.0014	17.0015
Molmo2-O-7B	0.624	0.676	16.4034	16.4057

source-count group. CBPC exceeds the best evaluated fixed coefficient in all six settings and also exceeds the shuffled-reference mean throughout. These results support the use of an input-matched visual reference over the evaluated fixed-strength and shuffled-reference alternatives.

Inference efficiency. We compare vanilla inference and CBPC on the same fixed Mantis subset; full measurement details are provided in the appendix. As shown in Table 4, CBPC increases LM forward time by 2.01–8.33% across the three representative models, while peak allocated GPU memory changes by at most 0.0023 GiB. Thus, CBPC incurs only a modest increase in LM forward time, with negligible additional peak allocated memory.

6 Conclusion

We studied position-dependent source preference in multi-image VLMs through model-native Boundary Keys and identified a dominant source-shared component whose RoPE-induced phase deviations modulate boundary source preference even with native source-specific residuals retained. We introduced Content-Bounded Phase Consensus (CBPC), which selectively contracts these deviations while preserving the cross-source mean using an input- and layer-dependent visual-key-centroid reference. CBPC consistently improves vanilla inference across diverse models and benchmarks and increases correctness under meaning-preserving source reorderings. It requires no training or additional parameters, modifies only Boundary Keys, and incurs modest LM forward overhead with negligible additional peak allocated memory. These results show that CBPC provides a lightweight way to improve multi-image source handling while retaining the model's native positional encoding and architecture.

AI USE STATEMENT

In this work, we used generative AI tools in a limited capacity to assist with English translation and LaTeX typesetting and formatting of the manuscript. All AI-assisted work was carefully reviewed and revised by the authors. We take responsibility for the final content of this work, and all text and other manuscript content were reviewed and approved by the authors.

REPRODUCIBILITY STATEMENT

The main paper describes CBPC and the experimental setup, while the appendix provides additional evaluation details. We also provide supplementary code implementing CBPC and the evaluation pipeline to support reproduction of the reported results.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

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

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Rohun Tripathi, Sangho Lee, Mohammadreza Salehi, Jason Ren, Chris Dongjoo Kim, Yinuo Yang, et al. Molmo2: Open weights and data for vision-language models with video understanding and grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28652–28668, 2026.

Anurag Das, Adrian Bulat, Alberto Baldrati, Ioannis Maniadis Metaxas, Bernt Schiele, Georgios Tzimiropoulos, and Brais Martinez. More images, more problems? a controlled analysis of vlm failure modes. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 7423–7442, 2026.

Jie Huang, Xuejing Liu, Sibo Song, RuiBing Hou, Hong Chang, Junyang Lin, and Shuai Bai. Revisiting multimodal positional encoding in vision–language models. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=sCCF4ygDAw.

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

Minyoung Lee, Yeji Park, Dongjun Hwang, Yejin Kim, Seong Joon Oh, and Junsuk Choe. Enhancing multi-image understanding through delimiter token scaling. *arXiv preprint arXiv:2602.01984*, 2026.

Haowei Liu, Xi Zhang, Haiyang Xu, Yaya Shi, Chaoya Jiang, Ming Yan, Ji Zhang, Fei Huang, Chunfeng Yuan, Bing Li, et al. Mibench: Evaluating multimodal large language models over multiple images. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22417–22428, 2024a.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*. Springer, 2024b.

Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Tianshuo Yang, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, et al. Mmiu: Multimodal multi-image understanding for evaluating large vision-language models. In *International Conference on Learning Representations*, 2025.

Yeji Park, Minyoung Lee, Sanghyuk Chun, and Junsuk Choe. Mitigating cross-image information leakage in lvlms for multi-image tasks. *arXiv preprint arXiv:2508.13744*, 2025.

Aditya Sharma, Michael Saxon, and William Yang Wang. Losing visual needles in image haystacks: Vision language models are easily distracted in short and long contexts. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5429–5451, 2024.

Dingjie Song, Shunian Chen, Guiming Hardy Chen, Fei Yu, Xiang Wan, and Benyou Wang. Milebench: Benchmarking mllms in long context. *arXiv preprint arXiv:2404.18532*, 2024.

Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021.

Zhijie Tan, Xu Chu, Weiping Li, and Tong Mo. Order matters: Exploring order sensitivity in multimodal large language models. *arXiv preprint arXiv:2410.16983*, 2024.

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

Chengcheng Wang, Jianyuan Guo, Hongguang Li, Yuchuan Tian, Ying Nie, Chang Xu, and Kai Han. Circle-rope: Cone-like decoupled rotary positional embedding for vision-language models. In *Forty-third International Conference on Machine Learning*, 2026.

Fei Wang, Xingyu Fu, James Y Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, et al. Muirbench: A comprehensive benchmark for robust multi-image understanding. In *International Conference on Learning Representations*, 2025a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b.

Tsung-Han Wu, Giscard Biamby, Jerome Quenum, Ritwik Gupta, Joseph E Gonzalez, David Chan, et al. Visual haystacks: A vision-centric needle-in-a-haystack benchmark. In *International Conference on Learning Representations*, 2025.

Peng Xia, Siwei Han, Shi Qiu, Yiyang Zhou, Zhaoyang Wang, Wenhao Zheng, Zhaorun Chen, Chenhang Cui, Mingyu Ding, Linjie Li, et al. Mmie: Massive multimodal interleaved comprehension benchmark for large vision-language models. In *International Conference on Learning Representations*, 2025.

Xiaochen Yang, Hao Fang, Jiawei Kong, Yaoxin Mao, Bin Chen, and Shu-Tao Xia. Looking back and forth: Cross-image attention calibration and attentive preference learning for multi-image hallucination mitigation. In *European Conference on Computer Vision*, pp. 259–277. Springer, 2026.

Jianshu Zhang, Dongyu Yao, Renjie Pi, Paul Pu Liang, and Yi R Fung. Vlm2-bench: A closer look at how well vlms implicitly link explicit matching visual cues. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7510–7545, 2025.

Zicheng Zhang, Haoning Wu, Erli Zhang, Guangtao Zhai, and Weisi Lin. Q-bench+: A benchmark for multi-modal foundation models on low-level vision from single images to pairs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10404–10418, 2024.

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

Haozhe Zhao, Zefan Cai, Shuzheng Si, Xiaojian Ma, Kaikai An, Liang Chen, Zixuan Liu, Sheng Wang, Wenjuan Han, and Baobao Chang. Mmicl: Empowering vision-language model with multi-modal in-context learning. In *International Conference on Learning Representations*, 2024b.

Table 5: **Shuffled-C_{vis} results by permutation seed.** Accuracy (%).

Benchmark	Model	Seed 0	Seed 1	Seed 2
Mantis	Qwen3-VL-8B	67.28	68.66	67.74
	InternVL3.5-8B	64.52	64.98	64.06
	Molmo2-O-7B	63.13	63.13	63.59
QBench2	Qwen3-VL-8B	77.80	77.60	77.70
	InternVL3.5-8B	76.00	75.90	76.00
	Molmo2-O-7B	75.80	75.60	75.80

A Appendix

A.1 Benchmark Details

Mantis. Mantis-Eval (Jiang et al., 2024) is a human-annotated benchmark for challenging multi-image reasoning, requiring models to integrate evidence across multiple visual sources for tasks such as cross-image comparison and co-reference. We evaluate all 217 examples.

MuirBench. MuirBench (Wang et al., 2025a) evaluates robust multi-image understanding across 12 tasks and 10 categories of inter-image relations, including temporal, multiview, narrative, and ordering relations. Its pairwise design couples answerable instances with minimally modified unanswerable counterparts to test whether models reliably use the full multi-image context. We evaluate all 2,600 questions.

MIRB. MIRB (Zhao et al., 2024a) evaluates reasoning that requires comparing and integrating multiple images across four dimensions: perception, visual world knowledge, reasoning, and multi-hop reasoning. Its tasks are designed so that multiple visual inputs must be jointly considered rather than treating each image independently. We evaluate 969 examples.

QBench2. QBench2 (Zhang et al., 2024), the image-pair component of Q-Bench+, evaluates low-level visual perception through pairwise comparison and joint analysis of attributes such as clarity, lighting, color, composition, and visual distortions. We evaluate 1,000 examples.

A.2 Additional Details and Ablations

Numerical stabilization. We use $\epsilon = 64\,\epsilon_{\text{FP32}} = 7.6293945 \times 10^{-6}$ in Eq. 13.

A.2.1 Shuffled Visual-Reference Runs

The shuffled-C_{vis} control in Table 3 is evaluated with three permutation seeds. For each seed, donor assignments are generated by a deterministic cycle permutation within each model, benchmark, and source-count group. Singleton groups use the input itself as the donor. The donor's layer-wise C_{vis} profile is applied to the corresponding recipient layers. Table 5 reports the individual runs underlying the mean and sample standard deviation shown in Table 3.

A.2.2 Intervention-Layer Ablation

We vary the decoder layers at which CBPC is applied while using the same layer-specific recalibration rule at every selected layer. For Qwen3-VL and InternVL3.5, Middle-2 and Late-2 correspond to layers 17–18 and 34–35, respectively; for Molmo2-O, they correspond to layers 15–16 and 30–31. All layers denotes layers 0–35 for Qwen3-VL and InternVL3.5 and layers 0–31 for Molmo2-O. Table 6 reports the complete results. Early-2 obtains the highest accuracy across all evaluated model–benchmark settings.

Table 6: **Intervention-layer ablation.** Accuracy (%) on Mantis and QBench2. Early-2 is the default CBPC configuration.

Schedule	Mantis			QBench2		
	Qwen	InternVL	Molmo	Qwen	InternVL	Molmo
Vanilla	67.74	64.98	61.75	76.80	75.80	74.60
L0 only	68.20	64.98	62.67	77.70	76.00	75.60
L1 only	68.20	64.52	63.59	77.50	75.90	75.30
Middle-2	67.28	64.98	63.13	77.40	76.00	75.10
Late-2	67.74	64.98	62.21	77.70	76.00	75.40
All layers	67.74	64.52	62.67	77.80	75.70	75.50
Early-2 (CBPC)	**69.12**	**65.90**	**64.06**	**78.00**	**76.20**	**75.90**

Table 7: **Composition of the source-reordering evaluation.**

Images	Original questions	Presentations
2	145	290
3	39	117
4	32	128
5	1	5
Total	**217**	**540**

A.3 Source-Reordering Protocol

Our robustness evaluation uses all 217 Mantis-Eval questions containing between two and five images. For an example with S_i images, we evaluate all S_i cyclic shifts rather than all $S_i!$ permutations, so that every image occupies each source position exactly once. This produces 540 evaluated presentations in total.

For example, a three-image question is evaluated under $[I_1, I_2, I_3]$, $[I_2, I_3, I_1]$, and $[I_3, I_1, I_2]$. The transformation changes the image presentation order and remaps the corresponding source references while preserving the original semantics. Original image bytes are reused without re-encoding, and explicit references such as "Image 1" or "Image 2" in questions and answer choices are remapped to the corresponding source in the reordered presentation. The answer option is also transformed when necessary to preserve the semantics of the original problem. The native ordering is stored as `variant_index=0`, and all transformed versions of a question share the same `base_sample_id`. We therefore refer to this procedure as a meaning-preserving, lossless order transformation.

A.4 Efficiency Measurement Details

For the efficiency evaluation, we randomly order the 217 Mantis examples with seed 42 and fix the first 100 valid inputs. Vanilla and CBPC use exactly the same inputs and ordering. Each method is executed in a fresh process with batch size 1 and `max_new_tokens=16`; model-loading time is excluded. The first five requests in each process are treated as warm-up and omitted from measurement.

LM forward time. The LM forward time reported in Table 4 is measured using CUDA Events and includes the first prefill forward pass and all subsequent decoding forwards:

$$T_{\mathrm{LM}} = T_{\mathrm{prefill}} + \sum_{t} T_{\mathrm{decode}}^{(t)}. \tag{16}$$

To avoid confounding computational overhead with differences in generation length, we retain only paired Vanilla–CBPC runs with identical numbers of generated tokens. This leaves 94 paired examples for Qwen3-VL-8B and 95 each for InternVL3.5-8B and Molmo2-O-7B.

Peak allocated memory. Immediately before generation for each example, we reset CUDA peak-memory statistics using `torch.cuda.reset_peak_memory_stats()`, and after generation we record `torch.cuda.max_memory_allocated()`. The reported quantity is therefore peak allocated CUDA memory during generation; it is not reserved memory, device-wide utilization, or a model-loading peak.