FlexAC Towards Flexible Control of Associative Reasoning in Multimodal Large Language Models

Shengming Yuan^{1*} shengming.yuan@outlook.com

Xinyu Lyu^{2*} xinyulyu68@gmail.com

Shuailong Wang¹ wslliongliong@gmail.com

Beitao Chen¹

Jingkuan Song³

Lianli Gao^{1†}

chenbeitao@gmail.com jingkuan.song@gmail.com

lianli.gao@uestc.edu.cn

¹University of Electronic Science and Technology of China ²Southwestern University of Finance and Economics, Chengdu, China ³Tongji University

Abstract

Multimodal large language models (MLLMs) face an inherent trade-off between faithfulness and creativity, as different tasks require varying degrees of associative reasoning. However, existing methods lack the flexibility to modulate this reasoning strength, limiting MLLMs' adaptability across factual and creative scenarios. To bridge this gap, we propose equipping MLLMs with mechanisms that enable flexible control over associative reasoning. We begin by investigating the internal mechanisms underlying associative behavior in MLLMs and find that: (1) middle layers play a pivotal role in shaping model's associative tendencies, (2) modifying representations in these layers effectively regulates associative reasoning strength, and (3) hallucinations can be exploited to derive steering vectors that guide this modulation. Building on these findings, we introduce Flexible Association Control (FlexAC), a lightweight and training-free framework for modulating associative behavior in MLLMs. FlexAC first induces hallucination-guided intermediate representations to **encode associative directions**. Then, it selects high-association instances to construct effective associative steering vectors, whose strengths are adaptively calibrated to balance creative guidance with output stability. Finally, recognizing the multi-dimensional nature of associative reasoning, FlexAC incorporates task-specific associative vectors derived from a forward pass on a few target-domain samples, enabling models to follow diverse associative directions and better adapt to **creative tasks**. Notably, our method achieves up to a 5.8× improvement in creativity on Creation-MMBench and a 29% reduction in hallucination rate on CHAIR, surpassing existing baselines and demonstrating its effectiveness in enabling flexible control over associative reasoning in MLLMs. Our code is available at https://github.com/ylhz/FlexAC.

1 Introduction

In cognitive science, divergent and convergent thinking represent two distinct modes of human associative behavior: convergent thinking relies on typical, fact-based associations to support faithful

^{*}Equal contribution.

[†]Corresponding author.

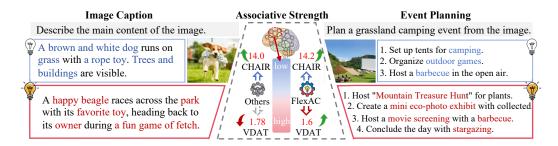


Figure 1: **Different tasks require different levels of associative reasoning**: factual tasks (*e.g.*, image caption) benefit from lower association, while creative tasks (*e.g.*, event planning) thrive on higher association. Existing methods suppress hallucinations at the cost of creativity (*e.g.*, -1.78 on VDAT; "Others" from Ha-DPO). FlexAC enables MLLMs to adjust associative reasoning strength accordingly.

reasoning, whereas divergent thinking engages atypical, context-dependent associations to foster creativity [1]. Recent studies show that multimodal large language models (MLLMs) [2, 3, 4] exhibit brain-like properties, such as structured embedding spaces [5], cross-modal integration [6], and higher-order cognitive functions [7], indicating that they emulate human associative processes. Consequently, like the human brain, MLLMs require the capacity to flexibly regulate associative reasoning strength to support both faithful reasoning and creative generation.

However, existing methods lack the flexibility to modulate associative reasoning strength, limiting MLLMs' adaptability across factual and creative scenarios. On one hand, current hallucination mitigation techniques, such as Contrastive Decoding [8, 9, 10] and Direct Preference Optimization [11], focus on improving faithfulness but often suppress associative reasoning capabilities, thereby hindering performance on tasks involving imaginative understanding and literary expression. On the other hand, how to enhance MLLMs' creativity in a controllable and task-specific manner remains underexplored. For instance, as illustrated in Figure 1, existing hallucination mitigation techniques improve faithfulness (14.0 \downarrow in CHAIR) but lack mechanisms for enhancing creativity, resulting in reduced associative reasoning strength (1.78 \downarrow in VDAT) and poor performance on tasks such as event planning. This gap highlights the need for equipping MLLMs with controllable mechanisms to flexibly modulate associative reasoning strength based on task demands.

To enable controllable modulation of associative reasoning strength, we begin by examining how associative behavior emerges within MLLMs. Drawing inspiration from prior works [12, 13], we hypothesize that hallucination and creativity arise from shared associative mechanisms, whose manifestations vary with task demands. To validate this, we collect input-response pairs containing both grounded (low-association) and hallucinated (high-association) outputs, and analyze their internal representations to uncover how associative behavior is reflected within the model. Our analysis reveals three key findings (see Section 2.1 and Section 2.2): (1) Associative behaviors are primarily encoded in the middle layers, where the representations of grounded and hallucinated responses become distinctly separable; (2) Modifying internal representations at these layers can effectively alter the strength and direction of associative reasoning; (3) Direction of hallucinated representations can stimulate associative reasoning capability, offering a potential control signal for this modulation. These findings indicate that associative tendencies are encoded in middle layers and can be modulated through targeted interventions guided by hallucination.

Motivated by these findings, we propose **Flex**ible Association **C**ontrol (**FlexAC**), a lightweight and training-free framework for modulating associative behavior in MLLMs. The core idea is to first extract the associative vector from hallucinated responses (Phase I: Offline Control Vector Construction), which exhibit strong associative tendencies, and then apply it at inference time to guide model behavior (Phase II: Inference-Time Control). **In the Offline Control Vector Construction Phase**, FlexAC performs **three key steps:** (1) Hallucination-Guided Intermediate States: We collect grounded—hallucinated response pairs, and measuring the differences between their hidden states within model's middle layers, which encode the associative direction. (2) Instance Selection: To reduce noise from individual samples, we select the top-K response pairs with the largest association shifts and average their differences to obtain a reliable steering vector. (3) Directional Integration: To further support tasks requiring multi-dimensional associations (e.g., storytelling or metaphor generation), we augment the general associative vector with task-specific associative vectors derived

from GPT-40-generated, high-association samples. These vectors are incorporated at inference time for fine-grained and controllable modulation. **In the Inference-Time Control Phase**, we apply the combined steering vector during inference. However, uniformly applying this vector can lead to over-steering, especially for inputs already exhibit strong associative behavior, causing irrelevant outputs or stylistic drift. To mitigate this, we introduce **Steering Intensity Calibration**, which adaptively scales the steering vector: amplifying it when associative behavior is weak, and attenuating it when the desired level has been reached.

To evaluate the effectiveness of FlexAC in controlling associative behavior, we conduct experiments across three fronts: hallucination mitigation (CHAIR [14] and POPE [15] for low-association tasks), creativity enhancement (VDAT and Creation-MMBench [16] for high-association tasks), and general-purpose evaluation (MME [17], MMMU [18], and MMStar [19]). Results show that FlexAC enables flexible modulation of associative reasoning capability, achieving state-of-the-art performance on both low- and high-association tasks while enhancing general capabilities.

In summary, our contributions are fourfold: (1)We present a unified perspective that links hallucination and creativity to associative reasoning, identifying middle-layer representations as key control points. (2)We propose **FlexAC**, a lightweight and training-free framework for flexible modulation of associative strength, enabling task-aware switching between hallucination suppression and creativity enhancement. (3)We introduce VDAT, a benchmark specifically designed to evaluate associative reasoning strength. (4)We conduct comprehensive experiments demonstrating that FlexAC effectively controls associative behavior across hallucination, creativity, and general-purpose benchmarks.

2 Analyzing and modulating associative behavior in MLLMs

2.1 Analyzing layer-wise localization of associative processes

Feature Distance Analysis: Quantifying layer-wise differences between associative and non-associative representations. To identify where associative behavior emerges, we analyze layer-wise representations in LLaVA-1.5-7b using 1000 images from COCO2024. For each image, we collect two type of responses: a grounded (non-associative) response from the model's default output, and a hallucinated (associative) response induced via blurred inputs and specific prompts [8]. Here, we use hallucinated responses to represent associative behavior, as they often include many imaginative contents, objects that do not exist in the image but are semantically related to the scene, reflecting the model's associative tendencies. We then extract the associative features $f^{(a)}$ and non-associative features $f^{(n)}$ from all intermediate layers for both data types (visualized in Figure 4). The full data construction and feature extraction process is detailed in Appendix B. Next, we compute the cosine distance and Euclidean distance between $f^{(a)}$ and $f^{(n)}$ across all layers. The cosine distance \mathcal{D}_{cos} is used to evaluate the directional alignment between associative and non-associative features, while Euclidean distance \mathcal{D}_{Euc} measures the spatial distribution differences.

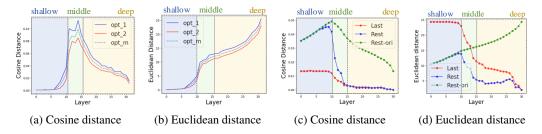
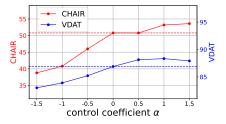
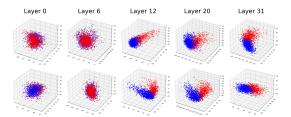


Figure 2: (a) and (b) show the cosine and Euclidean distances between associative and non-associative features across layers. (c) and (d) illustrate the impact of replacing associative features in different layers on subsequent layers. "Last" and "Rest" denotes the final layer difference d_L and the average layer difference $\bar{d}_{m:L}$, respectively. "Rest-ori" represents the original mean feature distance $\bar{d}_{m:L}$ without replacement.

As shown in Figures 2a and 2b, both cosine and Euclidean distances remain consistently low in the **shallow layers** (layers 0–9), indicating shared low-level perception. However, for middle and deep layers, we observe distinct patterns between cosine and Euclidean distance when comparing grounded





control using hallucination representations.

Figure 3: Impact of Middle Layer Control Figure 4: Visualization of feature representations on Hallucination-Driven Behavior. Adjusting in LLaVA-1.5-7b, reduced via PCA, shows red (as- α increases both hallucination (CHAIR) and sociative) and blue (non-associative) points. The creativity (VDAT), suggesting that associative feature distributions show increasing separation in strength can be modulated through middle-layer deeper layers, illustrating how associative distinctions are formed. See Appendix F.1 for all layers.

and hallucinated responses across layers. Cosine distance peaks in the **middle layers** (layers 10–15), indicating that this stage is where feature directions diverge most significantly—suggesting that associative behavior is primarily introduced and shaped in this range. In contrast, Euclidean distance increases steadily across both middle and deep layers (layers 10-31), implying that the overall feature magnitudes continue to drift even in later stages. This discrepancy raises a key question: Is associative behavior actively introduced in the deep layers, or are these differences merely the propagated result of associative shifts originating in the middle layers?

Layer Intervention: Verifying the source of as**sociative signals.** To answer this, we conduct a layer intervention experiment (Figure 5), in which we replace the associative feature $f_m^{(a)}$ with the corresponding non-associative feature $f_m^{(n)}$ at different layers m, and observe the influence on downstream representations. The modified feature propagation is defined as:

$$f_l^{\text{modified}} = \begin{cases} f_l^{(a)} & l < m \\ \mathcal{M}^l \circ \dots \circ \mathcal{M}^{m+1}(f_m^{(n)}) & l \ge m, \end{cases}$$
(1)

where \mathcal{M}^l denotes the *l*-th layer of the model. We evaluate the impact by calculating the final layer difference d_L and the average layer difference $\bar{d}_{m+1:L}$ as follows:

$$d_L = \mathcal{D}(f_L^{\text{modified}}, f_L^{(n)}) \tag{2}$$

$$\bar{d}_{m:L} = \frac{1}{L-m} \sum_{i=m+1}^{L} \mathcal{D}(f_i^{\text{modified}}, f_i^{(n)}), \quad (3)$$

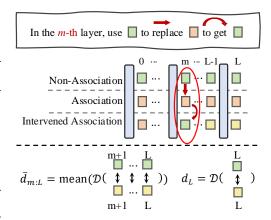


Figure 5: Layer Intervention for Association **Localization.** The goal is that locating the key layers for associative feature generation. Associative features are replaced with non-associative ones at different layers, and the impact on subsequent layers is evaluated using d_L and $d_{m:L}$.

where $\mathcal{D}(\cdot)$ denotes either cosine or Euclidean distance.

Results in Figures 2c and 2d show that replacing features in shallow layers (layers 0-9) leads to minimal changes in downstream representations, indicating limited influence on associative processing. In contrast, replacing features in middle layers (layers 10-15) significantly reduces divergence in later layers, suggesting that these layers are the primary source of associative behavior. Replacements in deep layers (layers 16-31) again have limited impact, implying that these layers mainly propagate rather than generate associative features. More visualization in Appendix F.2.

Finding 1: Middle layers are critical for shaping MLLM's associative behavior.

Phase I: Offline Control Vector Construction

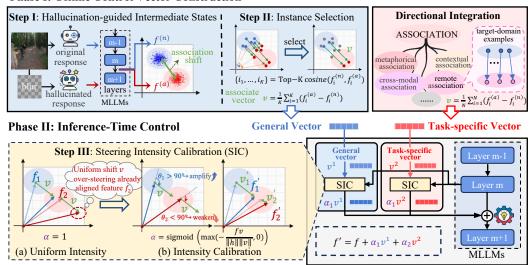


Figure 6: Overview of the proposed FlexAC framework. Phase I: Offline Control Vector Construction extracts a general associative vector from hallucination-guided intermediate features (Step I), by selecting Top-K instance pairs with maximal association shifts (Step II). It also constructs task-specific associative vectors from a few target-domain examples (Step III), reflecting diverse associative needs. Phase II: Inference-Time Control injects these vectors into middle-layer features. A Steering Intensity Calibration (SIC) module adaptively adjusts the influence of each vector per sample to achieve controllable associative reasoning strength.

2.2 Analyzing control strategies for associative behavior modulation

This analysis investigates whether associative behavior can be modulated by manipulating middle-layer representations, and whether hallucinated responses reveal effective directions for such control. Using the same grounded and hallucinated feature pairs from Section 2.1, we compute feature differences layer by layer to derive the control direction:

$$v_l = f_l^{(a)} - F_l^{(n)}. (4)$$

We then apply this steering vector during inference to modulate the model's output by adjusting the middle-layer features with control coefficient α :

$$f_l^{\text{control}} = f_l + \alpha \cdot v_l. \tag{5}$$

To assess the impact of steering on associative behavior, we introduce VDAT (Visual-Divergent Association Test), a benchmark that evaluates a model's associative reasoning by prompting it to generate unrelated nouns to the input image, thereby measuring its capacity for visual-driven divergent thinking (details in Section 3.1). As shown in Figure 3, increasing α from -1.5 to 1.5 raises CHAIR from approximately 38.8 to 53.6 and VDAT from around 83 to 87.9, indicating that higher α values lead to both more hallucination and stronger associative ability. Conversely, decreasing α reduces both scores. These results highlight that α provides a controllable mechanism for modulating associative behavior in MLLMs. These results yield two key findings:

Finding 2: Modifying middle layers enables control over associative reasoning strength.

Finding 3: Hallucinations help derive steering vectors to guide associative reasoning.

2.3 Flexible association control

Based on our findings in Sections 2.1 and 2.2, we propose Flexible Association Control (FlexAC), a lightweight, training-free framework for modulating associative behavior in MLLMs. As illustrated in Figure 6, FlexAC operates in two phases: (I) Offline Control Vector Construction, which derives general and task-specific associative directions, and (II) Inference-Time Control, which injects these directions into middle-layer features for dynamic modulation.

Phase I: Offline Control Vector Construction. To capture a general associative direction, we first induce hallucinated responses that exhibit high associative behavior (**Finding 3**). For each input, we extract hidden features from the middle layer l, where associative distinctions are most prominent (**Finding 1**), resulting in paired features $f_l^{(a)}$ and $f_l^{(n)}$. We select the top-K pairs with the highest cosine distances to construct a representative direction vector:

$$\mathcal{I} = \text{Top} - K\left(\mathcal{D}_{\cos}(f_{l,i}^{(a)}, f_{l,i}^{(n)})\right); v_l = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \left(f_{l,i}^{(a)} - f_{l,i}^{(n)}\right)$$
(6)

To handle tasks requiring diverse associative patterns (e.g., metaphorical, contextual), we further construct task-specific associative vectors from a few high-association, instruction-aligned examples. As vanilla MLLMs struggle to produce such outputs, we leverage GPT-40 to generate high-quality associative outputs.

Phase II: Inference-Time Control. During inference-time phase, we adjust the hidden state f_l at middle layer l (**Finding 2**) by injecting a combination of general associative vector v_l^{gen} and task-specific associative vector v_l^{task} :

$$f_l^{\text{control}} = f_l + \alpha_{\text{gen}} \cdot v_l^{\text{gen}} + \alpha_{\text{task}} \cdot v_l^{\text{task}}$$
 (7)

where α is the tunable coefficient that controls the steering intensity. This formulation is grounded in recent theoretical findings [20], which reveal that task-specific differences in model weights exhibit linearly decomposable structures. This property supports our assumption that associative directions can be independently extracted and combined within the hidden space.

However, directly applying a uniform steering vector across all inputs can lead to over-steering, especially when the input already exhibits strong associative behavior, causing deviation from the intended semantic space (see Step III of Figure 6). To mitigate it, we introduce steering intensity calibration strategy, which adjusts the steering strength α based on:

$$\alpha = \operatorname{sigmoid}\left(\max\left(-\frac{f_l \cdot v_l}{\|f_l\|\|v_l\|}, 0\right)\right) \tag{8}$$

This formulation increase steering strength when the current representations is misaligned with the associate direction, and suppresses it when already aligned. We further normalize the modulated feature to preserve its scale:

$$f_l^{\text{control}} \leftarrow f_l^{\text{control}} \cdot \frac{\|f_l\|}{\|f_l^{\text{control}}\|}$$
 (9)

This mechanism enables precise, interpretable modulation of associative behavior, allowing MLLMs to shift smoothly between factual accuracy and creative generation (Figure 8).

3 Experiments

3.1 Experimental setup

Evaluation Metric: To evaluate the effectiveness of FlexAC, we conduct experiments on three benchmark types: (1) **hallucination**, using CHAIR [14] and POPE [15] to assess object-level factual consistency; (2) **creativity**, using our proposed VDAT for associative reasoning and Creation-MMBench [21] for open-ended image-grounded generation; and (3) **general-purpose capability**, using MME [17], MMMU [18] and MMStar [19] to ensure core perception and reasoning are preserved. Metric details are in Appendix C.

VDAT: Visual Divergent Association Test. To measure a model's associative reasoning and creative potential more directly, we introduce VDAT, a diagnostic benchmark that complements Creation-MMBench by focusing specifically on associative reasoning strength. Inspired by [22], VDAT prompts the model to generate multiple nouns that are unrelated both to the input image, capturing its capacity for visual-driven divergent thinking (Figure 7). The metric is computed using CLIP ViT-L/14 embeddings.

Implementation Details. We evaluate the effectiveness of our FlexAC on LLaVA-1.5 [23], Qwen-VL [3], and Deepseek-VL [4], comparing it with Ha-DPO [11], VCD [8], and VAF [24]. From the COCO2014 [25] dataset, we randomly selected 2000 images and then applied Instance Selection

to choose 50 images for generating the general association vector. For the layer intervention, we manipulated the following layers based on each model's associative strength: Qwen-VL (layers 15, 16, 17), LLaVA-1.5 (layers 11, 12, 13), and Deepseek-VL (layers 4, 5, 6). For FlexAC-P (faithfulness-enhanced) and FlexAC-C (creativity-enhanced), the control coefficient α is set to -1 and 1, respectively. All experiments were conducted on 8×RTX 4090 GPUs. The parameter analysis of the number of images is provided in Appendix E.1.

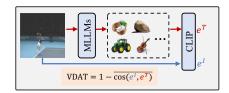


Figure 7: **Visual Divergent Association Test** (**VDAT**) evaluates a model's associative reasoning by prompting it to generate unrelated nouns from an image, and quantifies performance through image-text measured using CLIP embeddings.

Table 1: **Performance on hallucination benchmarks.** FlexAC here denotes the version configured to suppress associative behavior, aiming to improve factual accuracy (faithfulness).

Models	Methods	CHAIR				POPE				
Widels		$\overline{\operatorname{CHAIR}_S}\downarrow$	$CHAIR_I\downarrow$	Recall	Len	F1-score ↑	Accuracy ↑	Precision ↑	Recall	
Qwen-VL	Regular	40.6	12.5	71.7	94.6	85.6	86.6	92.9	79.3	
	VCD	42.0	11.2	71.7	91.2	86.3	87.2	92.4	81.0	
	VAF	38.0	11.7	72.2	91.4	86.5	87.2	91.4	82.0	
	FlexAC (Ours)	19.2	5.4	62.5	74.8	87.1	87.4	89.3	85.1	
LLaVA-1.5	Regular	50.8	14.3	79.7	97.3	86.5	87.2	91.5	82.0	
	Ha-DPO	36.8	10.4	74.0	88.3	83.9	85.3	92.6	76.7	
	VCD	51.0	15.5	79.1	98.9	84.3	84.9	88.1	80.7	
	VAF	47.8	13.7	79.2	96.1	86.9	87.1	87.9	85.9	
	FlexAC (Ours)	36.6	10.4	75.0	95.1	87.9	87.8	87.1	88.8	
Deepseek-VL2	Regular	32.6	9.2	67.0	121.0	88.5	88.4	88.1	88.8	
	VCD	36.6	11.3	67.2	128.2	87.9	87.8	87.6	88.1	
	VAF	32.0	9.2	66.2	119.0	88.5	88.4	87.6	89.4	
	FlexAC (Ours)	28.6	8.1	64.7	117.0	88.6	88.5	88.4	88.7	

3.2 Main results

Results on Hallucination Benchmark. To evaluate FlexAC's ability to improve factual accuracy in faithfulness-focused tasks, we conduct experiments on CHAIR and POPE. To this end, we set α in FlexAC to 1, selecting the precision-optimized variant. As shown in Table 1, FlexAC consistently achieves the lowest hallucination scores on most models and metrics. For examples, on CHAIR_S, FlexAC reduces hallucination to 19.2 (\uparrow 21.4) on Qwen-VL, 36.6 (\uparrow 14.2 vs. Regular) on LLaVA-1.5, and 28.6 (\uparrow 4.0) on Deepseek-VL2. On CHAIR_I, it similarly achieves the best scores (5.4, 10.4, and 8.1 respectively). In terms of POPE accuracy, FlexAC achieves the highest F1-score on LLaVA-1.5 (87.9) and comparable or superior precision and recall across the board. These results highlight FlexAC's ability to flexibly suppress excessive associative behavior in factual tasks, leading to improved accuracy across models.

Results on Creativity Benchmark. To evaluate FlexAC's ability to enhance associative reasoning in creative tasks, we conduct experiments on VDAT(Table 2) and Creation-MMBench(Table 3).

As shown in Table 2, hallucination mitigation methods like Ha-DPO reduce hallucinations but impair associative capacity, leading to lower creativity (*e.g.*, VDAT score of 85.11 vs.

Table 2: **Performance on VDAT.** FlexAC here denotes the version optimized to enhance associative behavior for creative tasks (creativity).

Methods	Qwen-VL	LLaVA-1.5	DeepSeek-VL2
Regular	84.85	86.89	84.54
Ha-DPO	-	85.11	-
VCD	83.69	86.83	84.62
VAF	84.95	86.79	84.61
FlexAC (Ours)	86.58	88.49	84.76

86.89 for the regular model). In contrast, FlexAC improves remote associative reasoning, achieving a higher VDAT score of 88.49. To further verify the validity of the VDAT metric, we conduct a user study presented in Appendix D.1. Further, on Creation-MMBench Table 3, we report VFS (Visual Fidelity Score), which evaluates image-text alignment, and Reward, which quantifies creativity

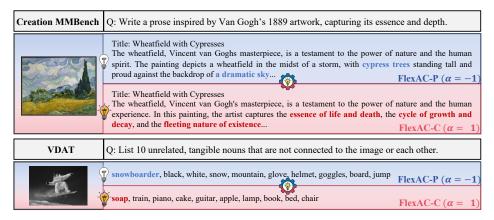


Figure 8: **Visualization of FlexAC's Control on Associative Reasoning.** This figure illustrates example outputs from Creation MMBench and VDAT, showing how FlexAC-P (faithfulness) and FlexAC-C (creativity) adjust the level of associative strength in the generated responses.

improvements relative to the base model (Qwen-VL). FlexAC achieves the highest Reward (10.92), outperforming methods like VCD (-3.86) and VAF (-1.63), while maintaining competitive VFS.

Qualitative examples in Figure 8 further support this: in Creation-MMBench, FlexAC-P focuses on concrete visual elements (e.g., "cypress trees"), while FlexAC-C introduces abstract themes (e.g., "life and death"). In VDAT, FlexAC-P outputs image-relevant nouns (e.g., "snowboarder"), whereas FlexAC-C generates semantically distant words (e.g., "guitar", "apple"), demonstrating enhanced divergent thinking. These examples confirm that FlexAC effectively modulates associative strength to meet diverse creative demands. For additional examples, see Appendix F.3.

Table 3: **Performance on Creation-MMBench.** We report results on four subcategories: Literary Writing (LW), Common Functional Writing (CFW), Professional Functional Writing (PFW), and Creative Multimodal Understanding (CMU). FlexAC here denotes the version optimized to enhance associative behavior for creative tasks (creativity).

Methods	Overall		LW		CMU		PFW		CFW	
Witting	VFS	Reward	VFS	Reward	VFS	Reward	VFS	Reward	VFS	Reward
Regular	6.10	0.00	6.83	0.00	5.53	0.00	5.58	0.00	6.66	0.00
VCD	6.05	-3.86	6.68	-2.71	5.67	2.50	5.61	-3.77	6.46	-6.57
VAF	6.06	-1.63	6.39	-3.96	5.57	-4.17	5.61	-0.53	6.64	-0.93
FlexAC (Ours)	6.25	10.92	7.20	15.63	5.83	6.11	5.43	5.96	7.00	15.65

Results on General-Purpose Benchmark. To evaluate the generalization capabilities of FlexAC across a range of tasks, we conduct experiments on three standard multimodal benchmarks using Qwen-VL: MME, MMMU, and MMStar. These benchmarks cover a wide range of capabilities including fine-grained grounding, reasoning, and instruction following.

As shown in Figure 9, both FlexAC-P (faithfulness-enhanced) and FlexAC-C (creativity-enhanced) maintain performance similar to the vanilla model across most categories, indicating no significant compromise in general capabilities. Notably, FlexAC-C outperforms the baseline on the OCR task in MME, likely due to its enhanced ability to associate text with related visual entities, improving inference and disambiguation under challenging conditions.

3.3 Ablation study

Layer-wise Control Analysis. We investigate the impact of middle layers on associative reasoning and identify the optimal control layers by testing interventions on shallow, middle, and deep layers, evaluating their effects on both CHAIR and VDAT metrics.

The results in Figure 10 demonstrate that middle layers have the most significant impact on performance: FlexAC-P achieves the best CHAIR results when suppressing associative behavior, while FlexAC-C shows the highest VDAT scores when enhancing creativity. In contrast, controlling shallow

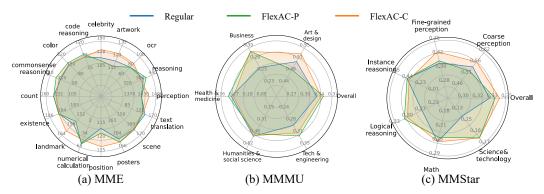


Figure 9: **Performance on general-purpose benchmarks.** Comparison of Regular, FlexAC-P (faithfulness-enhanced, $\alpha = -1$), and FlexAC-C (creativity-enhanced, $\alpha = 1$).

or deep layers has minimal effect. Based on these findings, we select layers 15, 16, and 17 as the control layers for Qwen-VL; results for other models are provided in Appendix E.2.

Effectiveness of different Components. We conducted an ablation study to assess the impact of components within FlexAC, including Instance Selection (IS), Steering Intensity Calibration (SIC), and Directional Integration (DI), on faithfulness (CHAIR) and creativity (VDAT).

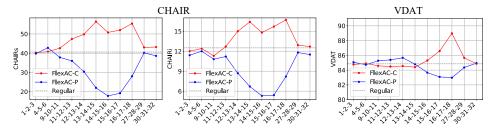


Figure 10: **Layer-wise analysis of control effectiveness in FlexAC.** The x-axis represents the control layers, while the y-axis shows the performance of the model on CHAIR and VDAT metrics.

As shown in Figure 11, for CHAIR, FlexAC(P) achieves the lowest CHAIR_S score (19.2), indicating effective hallucination reduction compared to the regular model (40.6). When IS and SIC are removed from FlexAC (FlexAC-IS-SIC), performance slightly worsens (30.4), confirming their role in enhancing faithfulness. Similarly, for creativity, FlexAC-C scores the highest VDAT (86.58). Removing IS and SIC in FlexAC-IS-SIC leads to a small decrease (85.05), while FlexAC-DI results in a slight improvement, highlighting the importance of DI for creativity. In summary, FlexAC enables flexible adjustment of associative strength to meet the needs of different tasks, balancing hallucination reduction and creativity enhancement effectively.

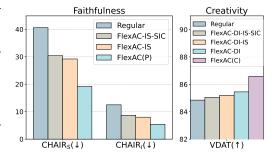


Figure 11: **Ablation study on components**, showing the impact of Instance Selection (IS), Steering Intensity Calibration (SIC), and Directional Integration (DI).

4 Related work

Multimodal Large Language Models. Recent advances in large language models (LLMs) [26, 27, 28] have led to the emergence of multimodal LLMs (MLLMs) that incorporate visual inputs for enhanced capabilities [29, 30]. LLaVA [2, 23] improves instruction-following via visual instruction tuning, while Qwen-VL [3] enhances spatial reasoning through visual grounding. DeepSeekdVL2 [4] adopts a Mixture-of-Experts architecture to improve multimodal comprehension.

Hallucination in MLLMs. MLLMs still face various safety risks [31, 32, 33]. Hallucination is one of the core challenges, where MLLMs generate content misaligned with visual input [34]. To address

this, VCD [8] employs contrastive decoding, and VAF [24] enhances visual signal processing during fusion. HA-DPO [11] reduces hallucinations via preference optimization.

Creativity in Large Models. Creativity, involving divergent thinking and novel associations [35, 36], has been explored in LLMs via cognitive theories like dual-pathway [37]. Olson *et al.* [38] and Chen & Ding [39] promote remote associations; MacGyver [40] and CLOT [41] tackle functional fixedness and divergent thinking, respectively. Creation-MMBench [21] provides a benchmark for evaluating image-grounded creative generation.

5 Conclusion

In this work, we investigate the root of associative behavior in MLLMs, finding that middle-layer representations govern associative reasoning strength and that hallucinated responses encode reliable steering directions. Based on these insights, we propose FlexAC, a lightweight, training-free framework that combines hallucination-guided steering with adaptive calibration and in-context augmentation. FlexAC enables controllable creativity and achieves state-of-the-art performance across hallucination, creativity, and general-purpose benchmarks. **Limitations:** FlexAC requires white-box access to hidden states and is not applicable to black-box models like ChatGPT.

6 Acknowledgements

This study is supported by grants from the National Natural Science Foundation of China (Grant No. U23A20315, No. 62425208, No. U22A2097, No. 62122018, No. 62020106008), Shenzhen Science and Technology Program (No.JCYJ20240813114208012), Fundamental Research Funds for the Central Universities, and Natural Science Foundation of Sichuan Province (Grant No. 2025ZNSFSC1463).

References

- [1] Liane Gabora. The neural basis and evolution of divergent and convergent thought. *The Cambridge handbook of the neuroscience of creativity*, pages 58–70, 2018. 2
- [2] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 2, 9
- [3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *CoRR*, 2023. 2, 6, 9
- [4] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024. 2, 6, 9
- [5] Ariel Goldstein, Haocheng Wang, Leonard Niekerken, Mariano Schain, Zaid Zada, Bobbi Aubrey, Tom Sheffer, Samuel A Nastase, Harshvardhan Gazula, Aditi Singh, et al. A unified acoustic-to-speech-to-language embedding space captures the neural basis of natural language processing in everyday conversations. *Nature Human Behaviour*, pages 1–15, 2025. 2
- [6] Jerry Tang, Meng Du, Vy Vo, Vasudev Lal, and Alexander Huth. Brain encoding models based on multimodal transformers can transfer across language and vision. *Advances in neural information processing systems*, 36:29654–29666, 2023. 2
- [7] Xuhui Jiang, Yuxing Tian, Fengrui Hua, Chengjin Xu, Yuanzhuo Wang, and Jian Guo. A survey on large language model hallucination via a creativity perspective. *arXiv preprint* arXiv:2402.06647, 2024. 2
- [8] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *CVPR*, pages 13872–13882, 2024. 2, 3, 6, 10, 22

- [9] Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024.* Association for Computational Linguistics, 2024. 2
- [10] Xinyu Lyu, Beitao Chen, Lianli Gao, Hengtao Shen, and Jingkuan Song. Alleviating hallucinations in large vision-language models through hallucination-induced optimization. In *NeurIPS*, 2024. 2
- [11] Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization, 2023. 2, 6, 10
- [12] Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *ACL*, pages 15504–15522, 2024. 2, 22
- [13] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In *ICLR*, 2024. 2
- [14] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *EMNLP*, pages 4035–4045, 2018. 3, 6, 22
- [15] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *EMNLP*, pages 292–305, 2023. 3, 6, 23
- [16] Xinyu Fang, Zhijian Chen, Kai Lan, Lixin Ma, Shengyuan Ding, Yingji Liang, Xiangyu Zhao, Farong Wen, Zicheng Zhang, Guofeng Zhang, Haodong Duan, Kai Chen, and Dahua Lin. Creation-mmbench: Assessing context-aware creative intelligence in MLLM. *CoRR*, abs/2503.14478, 2025. 3, 22
- [17] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. MME: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, 2023. 3, 6, 23
- [18] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *CVPR*, 2024. 3, 6, 23
- [19] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *NeurIPS*, 2024. 3, 6, 23
- [20] Hongkang Li, Yihua Zhang, Shuai Zhang, Meng Wang, Sijia Liu, and Pin-Yu Chen. When is task vector provably effective for model editing? a generalization analysis of nonlinear transformers. In *ICLR*, 2025. 6
- [21] Xinyu Fang, Zhijian Chen, Kai Lan, Lixin Ma, Shengyuan Ding, Yingji Liang, Xiangyu Zhao, Farong Wen, Zicheng Zhang, Guofeng Zhang, Haodong Duan, Kai Chen, and Dahua Lin. Creation-mmbench: Assessing context-aware creative intelligence in MLLM. *CoRR*, abs/2503.14478, 2025. 6, 10
- [22] Honghua Chen and Nai Ding. Probing the "creativity" of large language models: Can models produce divergent semantic association? In *EMNLP*, pages 12881–12888, 2023. 6
- [23] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pages 26286–26296, 2024. 6, 9
- [24] Hao Yin, Guangzong Si, and Zilei Wang. Clearsight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. In *CVPR*, 2025. 6, 10

- [25] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, pages 740–755, 2014. 6, 23
- [26] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*, 2023. 9
- [27] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 9
- [28] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 9
- [29] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. arXiv preprint arXiv:2306.04387, 2023. 9
- [30] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 9
- [31] Shengming Yuan, Qilong Zhang, Lianli Gao, Yaya Cheng, and Jingkuan Song. Natural color fool: Towards boosting black-box unrestricted attacks. In *NeurIPS*, 2022.
- [32] Youheng Sun, Shengming Yuan, Xuanhan Wang, Lianli Gao, and Jingkuan Song. Any target can be offense: Adversarial example generation via generalized latent infection. In ECCV, 2024.
- [33] Beitao Chen, Xinyu Lyu, Shengming Yuan, Jingkuan Song, Heng Tao Shen, and Lianli Gao. SafePTR: Token-level jailbreak defense in multimodal LLMs via prune-then-restore mechanism. In *NeurIPS*, 2025. 9
- [34] Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *CoRR*, abs/2402.00253, 2024. 9
- [35] J. P. GUILFORD. Creativity: Yesterday, today and tomorrow. The Journal of Creative Behavior, pages 3–14, 1967. 10
- [36] Mark Runco and Garrett Jaeger. The standard definition of creativity. *Creativity Research Journal CREATIVITY RES J*, pages 92–96, 2012. 10
- [37] Roger E. Beaty, Paul J. Silvia, Emily C. Nusbaum, Emanuel Jauk, and Mathias Benedek. The roles of associative and executive processes in creative cognition. *Memory & Cognition*, pages 1186–1197, 2014. 10
- [38] Jay A. Olson, Johnny Nahas, Denis Chmoulevitch, Simon J. Cropper, and Margaret E. Webb. Naming unrelated words predicts creativity. *Proceedings of the National Academy of Sciences*, page e2022340118, 2021. 10
- [39] Honghua Chen and Nai Ding. Probing the "creativity" of large language models: Can models produce divergent semantic association? In *EMNLP*, pages 12881–12888, 2023. 10
- [40] Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi, Thomas L Griffiths, and Faeze Brahman. Macgyver: Are large language models creative problem solvers? In *NAACL*, pages 5303–5324, 2024. 10
- [41] Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zitnik, and Pan Zhou. Let's think outside the box: Exploring leap-of-thought in large language models with creative humor generation. In *CVPR*, pages 13246–13257, 2024. 10
- [42] Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15840–15853, 2024. 22

- [43] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: evaluating large multimodal models for integrated capabilities. In *ICML*, pages 57730–57754, 2024. 27
- [44] 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*, pages 216–233. Springer, 2024. 27
- [45] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal Ilms with generative comprehension. arXiv preprint arXiv:2307.16125, 2023. 27
- [46] Hai-Long Sun, Da-Wei Zhou, Yang Li, Shiyin Lu, Chao Yi, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, De-Chuan Zhan, and Han-Jia Ye. Parrot: Multilingual visual instruction tuning. In *ICML*, 2025. 27
- [47] xAI. Grok-1.5 vision preview. https://x.ai/news/grok-1.5v, April 2024. Accessed: 2025-10-14. 27
- [48] Nannan Zhu, Yonghao Dong, Teng Wang, Xueqian Li, Shengjun Deng, Yijia Wang, Zheng Hong, Tiantian Geng, Guo Niu, Hanyan Huang, et al. Cvbench: Evaluating cross-video synergies for complex multimodal understanding and reasoning. arXiv preprint arXiv:2508.19542, 2025. 27
- [49] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images, 2016. 27
- [50] Amanpreet Singh, Vivek Natarjan, Meet Shah, Yu Jiang, Xinlei Chen, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR*, pages 8317–8326, 2019. 27
- [51] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings* of the Association for Computational Linguistics: ACL 2022, pages 2263–2279, 2022. 27
- [52] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In WACV, pages 2200–2209, 2021. 27
- [53] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, 2019. 27

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The main claims in the abstract and introduction accurately reflect the contributions of the paper, including introducing the FlexAC framework to control associative reasoning in MLLMs and demonstrating its effectiveness in modulating creativity and factual accuracy across various tasks.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals
 are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: We discuss the limitations in conclusion.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include formal theoretical proofs. It is empirical in nature, with a focus on the practical application of the FlexAC framework.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and crossreferenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: In Section 3, we provide comprehensive details of the experimental setup and all results required to reproduce our main findings, accompanied by thorough analysis and corresponding conclusions.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
- (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
- (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
- (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
- (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The core code is included in the supplementary material for reproducibility, and we plan to release the full codebase publicly upon acceptance.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be
 possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not
 including code, unless this is central to the contribution (e.g., for a new open-source
 benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how
 to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The paper provides sufficient experimental details, including the datasets used, model configurations, and the benchmarks in Section 3.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Section 3 and the appendix include detailed analysis of hyperparameters and consistent results across multiple datasets and models, demonstrating robustness even without explicit error bars.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error
 of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Section 3.1 provides details on the compute resources and settings used in our experiments.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: The research conform to the NeurIPS Code of Ethics. There is no indication of any ethical concerns, and the work focuses on improving deep learning techniques with no harmful implications.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a
 deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The broader societal impacts of this work, including both the potential benefits and risks, are discussed in Appendix 1.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper does not release any models or datasets that pose a high risk for misuse. All experiments are conducted using existing publicly available MLLMs and datasets, and no new potentially sensitive or high-risk assets are introduced.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with
 necessary safeguards to allow for controlled use of the model, for example by requiring
 that users adhere to usage guidelines or restrictions to access the model or implementing
 safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All external assets used in the paper, including datasets (e.g., COCO2014) and pretrained models (e.g., LLaVA, Qwen-VL, DeepSeek-VL), are properly cited in the references. Their usage complies with the respective licenses and terms of use as outlined by the original creators.

Guidelines:

• The answer NA means that the paper does not use existing assets.

- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The paper introduces a new method (FlexAC) and provides its core implementation code in the supplementary material.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: The paper includes a user study in which internal volunteers rated model outputs. Participants were informed of the study's purpose and procedure prior to participation. Although they were internal to the organization, compensation was provided in accordance with fair labor practices.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [Yes]

Justification: The paper includes a user study in which participants rated model outputs. Although the task posed minimal risk, participants' involvement was disclosed, and appropriate ethical considerations were followed.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent)
 may be required for any human subjects research. If you obtained IRB approval, you
 should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: The paper discusses the use of multimodal large language models (MLLMs) as part of the core methodology to demonstrate the effectiveness of the FlexAC framework, including its application in hallucination reduction and creativity enhancement.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.