
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
chenbeitao@gmail.com

Jingkuan Song³
jingkuan.song@gmail.com

Lianli Gao^{1†}
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

*Co-first author.

†Corresponding author.

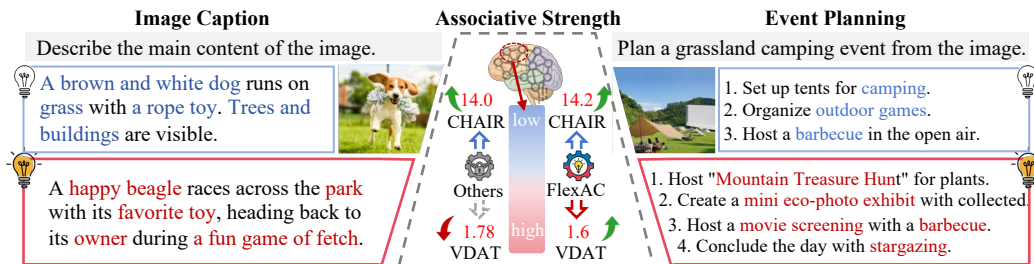


Figure 1: **Different tasks require different levels of associative reasoning:** factual tasks (e.g., image caption) benefit from lower association, while creative tasks (e.g., event planning) thrive on higher association. Existing methods suppress hallucinations at the cost of creativity (e.g., -1.78 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 ↓ in CHAIR) but lack mechanisms for enhancing creativity, resulting in reduced associative reasoning strength (1.78 ↓ 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 **Flexible Association Control (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-4o-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 $\mathcal{D}_{\text{Euc}}$ measures the spatial distribution differences.

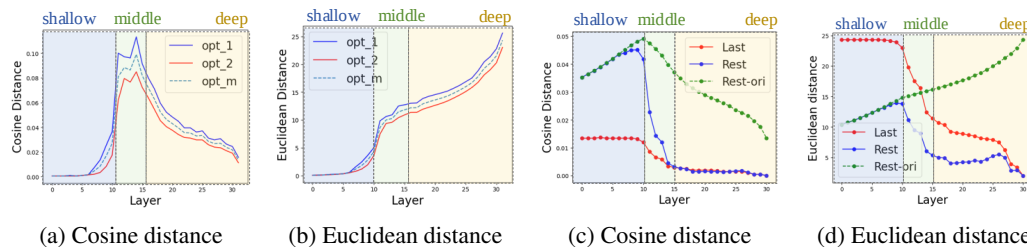


Figure 2: (a) and (b) show the cosine and Euclidean distances between associative and non-associative features across layers. (c) and (d) illustrate the impact of replacing associative features in different layers on subsequent layers. “Last” and “Rest” denotes the final layer difference d_L and the 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

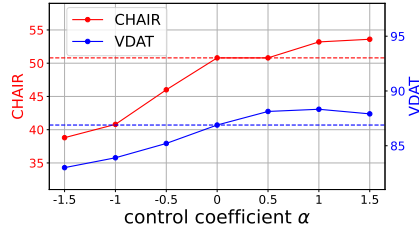


Figure 3: **Impact of Middle Layer Control on Hallucination-Driven Behavior.** Adjusting α increases both hallucination (CHAIR) and creativity (VDAT), suggesting that associative strength can be modulated through middle-layer control using hallucination representations.

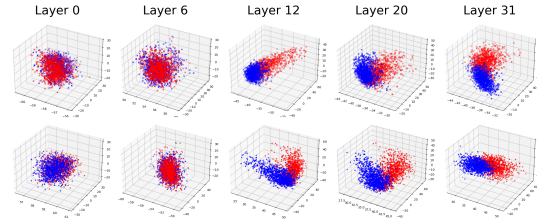


Figure 4: **Visualization of feature representations in LLaVA-1.5-7b**, reduced via PCA, shows red (associative) and blue (non-associative) points. The feature distributions show increasing separation in deeper layers, illustrating how associative distinctions are formed. See Appendix F.1 for all layers.

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 associative 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:

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

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:

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

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

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.

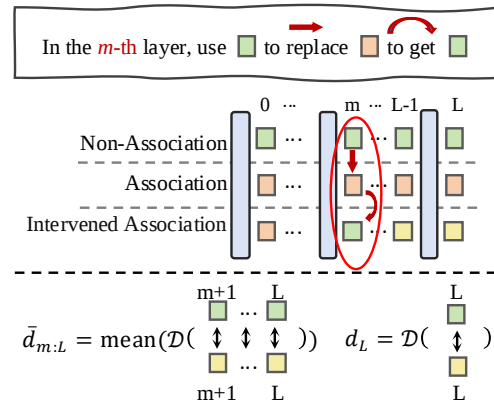


Figure 5: **Layer Intervention for Association Localization.** The goal is that locating the key layers for associative feature generation. Associative features are replaced with non-associative ones at different layers, and the impact on subsequent layers is evaluated using d_L and $\bar{d}_{m:L}$.

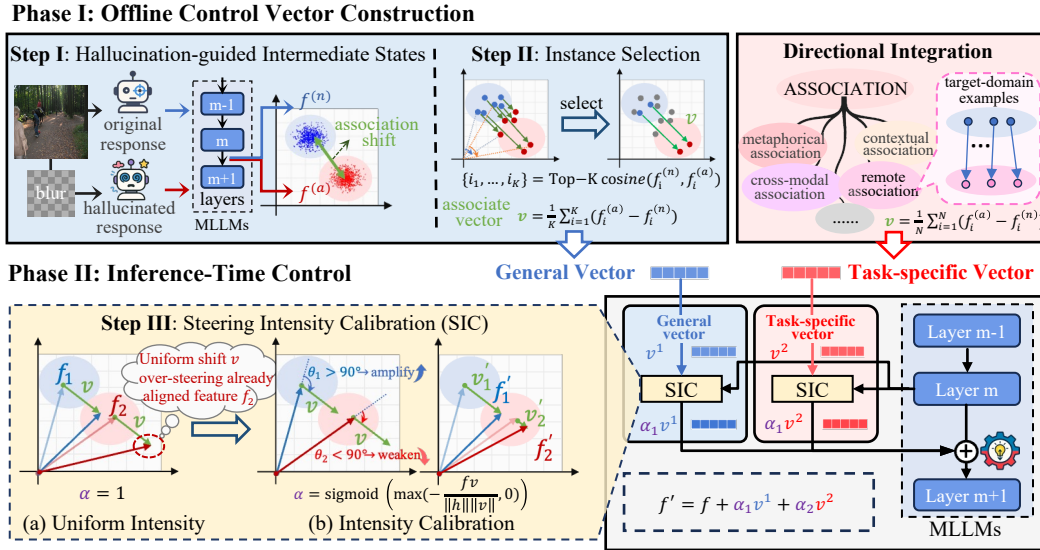


Figure 6: **Overview of the proposed FlexAC framework.** **Phase I: Offline Control Vector Construction** extracts a **general associative vector** from hallucination-guided intermediate features (Step I), by selecting Top-K instance pairs with maximal association shifts (Step II). It also constructs **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:

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

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

$$f_l^{\text{control}} = f_l + \alpha \cdot v_l. \quad (5)$$

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:

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

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-4o 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} :

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

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:

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

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:

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

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.

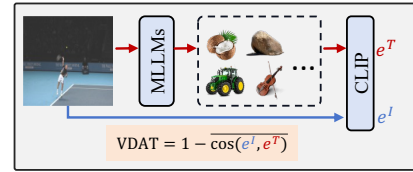


Figure 7: **Visual Divergent Association Test (VDAT)** evaluates a model’s associative reasoning by prompting it to generate unrelated nouns from an image, and quantifies performance through image-text measured using CLIP embeddings.

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			
		CHAIR _S ↓	CHAIR _I ↓	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 (↓21.4) on Qwen-VL, 36.6 (↓14.2 vs. Regular) on LLaVA-1.5, and 28.6 (↓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. 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

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

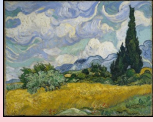
Creation MMBench	Q: Write a prose inspired by Van Gogh's 1889 artwork, capturing its essence and depth.
	Title: Wheatfield with Cypresses The wheatfield, Vincent van Goghs masterpiece, is a testament to the power of nature and the human spirit. The painting depicts a wheatfield in the midst of a storm, with cypress trees standing tall and proud against the backdrop of a dramatic sky ... FlexAC-P ($\alpha = -1$)
	Title: Wheatfield with Cypresses The wheatfield, Vincent van Gogh's masterpiece, is a testament to the power of nature and the human experience. In this painting, the artist captures the essence of life and death , the cycle of growth and decay , and the fleeting nature of existence ... FlexAC-C ($\alpha = 1$)
VDAT	Q: List 10 unrelated, tangible nouns that are not connected to the image or each other.
	snowboarder , black, white, snow, mountain, glove, helmet, goggles, board, jump FlexAC-P ($\alpha = -1$)
	soap , train, piano, cake, guitar, apple, lamp, book, bed, chair FlexAC-C ($\alpha = 1$)

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	
	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

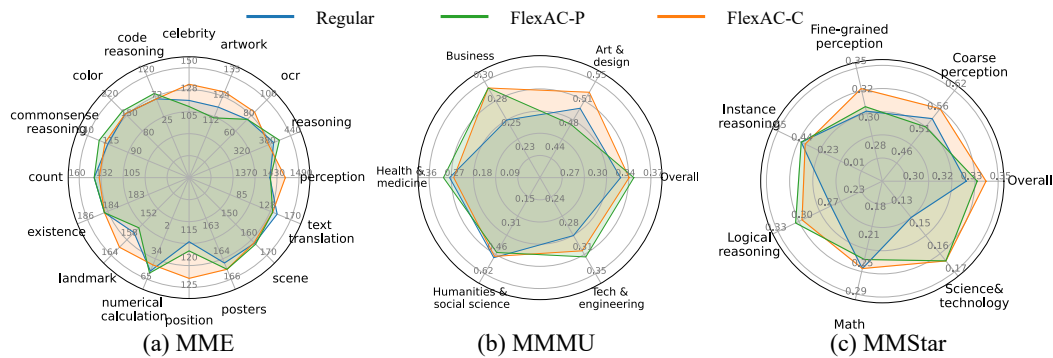


Figure 9: **Performance on general-purpose benchmarks.** Comparison of Regular, FlexAC-P (faithfulness-enhanced, $\alpha = -1$), and FlexAC-C (creativity-enhanced, $\alpha = 1$).

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).

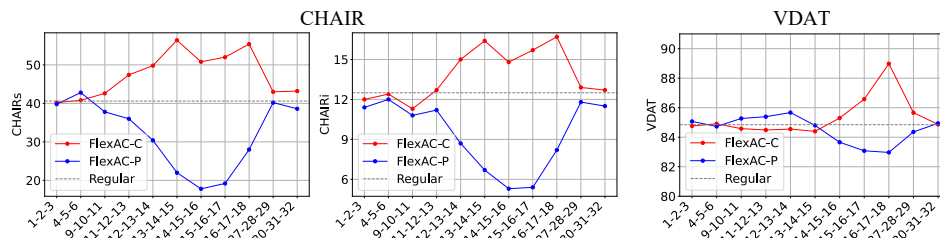


Figure 10: **Layer-wise analysis of control effectiveness in FlexAC.** The x-axis represents the control layers, while the y-axis shows the performance of the model on CHAIR and VDAT metrics.

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.

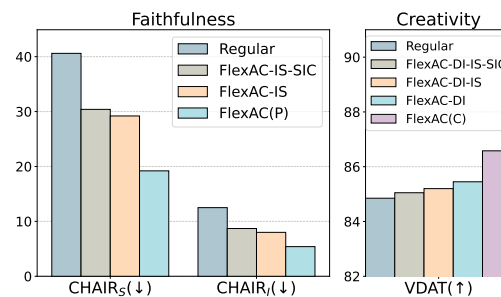


Figure 11: **Ablation study on components,** showing the impact of Instance Selection (IS), Steering Intensity Calibration (SIC), and Directional Integration (DI).

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. DeepSeek-VL2 [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, Beita 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 vlms through hallucination-aware direct preference optimization, 2023. 2, 6, 10
- [12] Nina Rimskey, 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 Wenhua 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. 9
- [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. 9
- [33] Beita 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 Marjeh, 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 llms 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 cross-referenced.
- 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 rigor, 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.

FlexAC: Towards Flexible Control of Associative Reasoning in Multimodal Large Language Models (Supplementary Material)

CONTENTS

A	Broader Impacts	22
B	Data Generation and Feature Extraction	22
C	Metrics details	22
D	Detailed Experimental Results	24
D.1	User study	24
D.2	Extended results on Creation-MMBench	26
D.3	Efficiency comparison	27
D.4	Extended Evaluation on General-Purpose Benchmarks	27
D.5	Detailed results on POPE	28
E	Ablation Study	29
E.1	Effect of dataset Sizes	29
E.2	Effect of control layer.	30
F	Visualizations	31
F.1	Detailed Feature Representation Analysis Using PCA	31
F.2	Detailed Layer Intervention for Association Localization	32
F.3	Visualization of more examples	35
F.4	Feature Distance Analysis on Additional Models	42

A Broader Impacts

FlexAC introduces finer control over the associative behavior of MLLMs, enabling safer and more context-appropriate responses across tasks. This may benefit applications requiring factual precision (e.g., education, medical support) or creative output (e.g., storytelling, art generation). However, enhancing associative capacity also increases the model’s expressive power, which—if misused—could lead to persuasive but unfounded generations. As with all generation-controlling techniques, FlexAC should be deployed alongside robust safeguards to ensure alignment with human intent and ethical use.

B Data Generation and Feature Extraction

Inducing and Representing Model Associations: To investigate the causes of model association, we generate two data distributions: one from the model’s original outputs (non-associative) and another with induced associative content using blurred images and tailored prompts [8, 42]. For example, the model is prompted with: “Describe the image and include some hallucinated objects that are imagined but do not exist in the image, as if they were real.” Following [12], we construct a multiple-choice dataset to capture feature distributions. The model is given an image and prompted to generate detailed responses, with two predefined options (Figure 12): [1] non-associative (factual) and [2] associative (creative). The hidden states corresponding to these inputs are extracted to obtain distinct feature representations, $F_{\text{non-assoc}}^l$ and F_{assoc}^l , capturing the model’s internal response to both associative and non-associative prompts across different layers.



```
<image>
Question: Please describe this image in detail.
[1] responses without association
[2] responses with association
Please select the most appropriate answer: [1 or 2]
```

Figure 12: The prompt for extracting associative and non-associative features

C Metrics details

All comparative experiments are conducted using the VLMEvalKit³. For binary choice questions, we prompt the model with: “Please answer Yes or No.” We evaluate three models in our experiments: LLaVA-1.5 (liuhaotian/llava-v1.5-7b), Qwen-VL (Qwen/Qwen-VL-Chat), and DeepSeek-VL2 (deepseek-ai/deepseek-vl2-tiny).

VDAT : VDAT fills a gap in evaluating the creative potential of multimodal models, which previous metrics did not adequately address. To ensure consistency, both CHAIR and VDAT were evaluated using the same 500 images, randomly selected from the MSCOCO dataset.

Creation-MMBench [16]: Creation-MMBench is a multimodal benchmark designed to evaluate the creative capabilities of MLLMs in real-world, image-grounded scenarios. It contains 765 test cases across 51 fine-grained tasks, with instance-specific criteria that assess both imaginative quality and visual consistency. In contrast to prior work that compares models to GPT-4o, our evaluation focuses on measuring improvements over each model’s own vanilla baseline.

CHAIR [14]: Caption Hallucination Assessment with Image Relevance (CHAIR) is a metric designed to evaluate the hallucination of image caption task. It measures the hallucination rate of the generated text by comparing the generated caption with the ground-truth caption. CHAIR consists of

³<https://github.com/open-compass/VLMEvalKit>

two metrics: CHAIR_S and CHAIR_I. They can be calculated as follows:

$$\text{CHAIR}_S = \frac{|\{\textit{hallucinated objects}\}|}{|\{\textit{all mentioned objects}\}|}, \quad (10)$$

$$\text{CHAIR}_I = \frac{|\{\textit{captions with hallucinated objects}\}|}{|\{\textit{all captions}\}|}. \quad (11)$$

POPE [15]: The Polling-based Object Probing Evaluation (POPE) is a metric developed to evaluate object hallucination in MLLMs. By framing the evaluation as a series of Yes-or-No questions about specific objects in images, POPE avoids issues related to instruction sensitivity. Using three sampling strategies—Random, Popular, and Adversarial—it effectively examines models’ tendencies to hallucinate frequently occurring or co-occurring objects, providing a stable and reliable assessment of object hallucination. Refer to [15], we built POPE on 500 randomly selected MSCOCO [25] validation images, each containing over three ground-truth objects and six constructed questions.

MME [17]: MLLM Evaluation benchmark (MME) is a benchmark designed to assess multimodal large language models (MLLMs) across core skills in perception and cognition, such as object recognition, attribute identification, reasoning, and translation. Using accuracy-based metrics, MME provides objective insights into model capabilities, highlighting areas for improvement in understanding and reasoning.

MMMU [18]: MMMU (Massive Multi-discipline Multimodal Understanding and Reasoning) is a large-scale benchmark targeting expert-level multimodal understanding and reasoning. It comprises 11.5K college-level questions across 6 disciplines and 30 subjects, featuring 30 diverse image types such as charts, medical scans, diagrams, and chemical structures. MMMU emphasizes deep domain knowledge and deliberate reasoning, challenging models to integrate perception, knowledge, and logic in complex tasks. It serves as a necessary testbed for evaluating progress toward Expert AGI.

MMStar [19]: MMStar is a high-quality benchmark designed to evaluate vision-language models on truly vision-dependent tasks. It includes 1,500 human-curated samples across 6 core capabilities and 18 fine-grained skills, ensuring minimal data leakage and strong visual grounding.

D Detailed Experimental Results

D.1 User study

To validate the effectiveness of the VDAT metric as a measure of associative creativity, we conducted a human evaluation study comparing FlexAC against several baselines. Specifically, we randomly selected 30 image-response examples from the Qwen-VL evaluation set and presented them to 15 human raters. For each example, two responses were shown—one from FlexAC and one from a baseline method (Regular, VAF, or VCD). Participants were asked to judge which response contained objects more unrelated to the image, as a proxy for stronger remote association. The response options were presented as “Answer A” and “Answer B,” with the method-to-label mapping randomized in each trial to eliminate bias. Raters evaluated each pair on a five-point scale ranging from “A is much better than B” to “B is much better than A.” These choices were then converted to numeric scores for aggregation—for example, “ $A \gg B$ ” assigns 3 points to A, “ $A = B$ ” assigns 1 point to both A and B.

Figure 13 shows that FlexAC consistently receives higher average scores than all baselines, with low variance across users. Figure 14 further reveals that over 70% of responses favored FlexAC ($A > B$ or $A \gg B$), while fewer than 6% favored the baseline. These results demonstrate strong alignment between the VDAT metric and human judgment. Figure 15 provides a screenshot of the evaluation interface. Together, these findings support VDAT as a valid and human-aligned metric for measuring associative creativity in vision-language generation tasks.

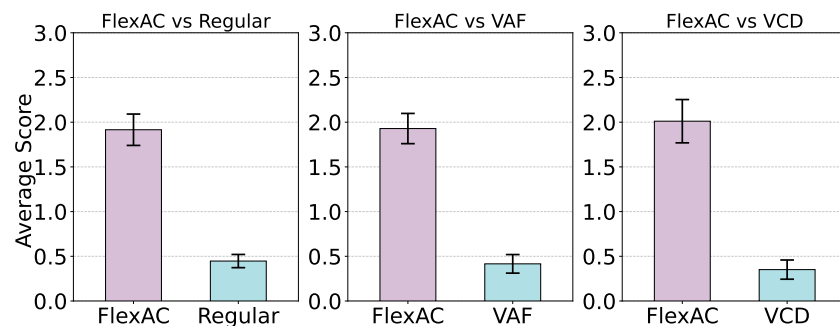


Figure 13: Average user ratings comparing FlexAC with baseline methods on the VDAT task. Each bar represents the average score across 15 users for 30 randomly selected image-response pairs. Error bars indicate the maximum and minimum individual user scores, reflecting rating consistency. Higher scores indicate stronger perceived remote association ability.

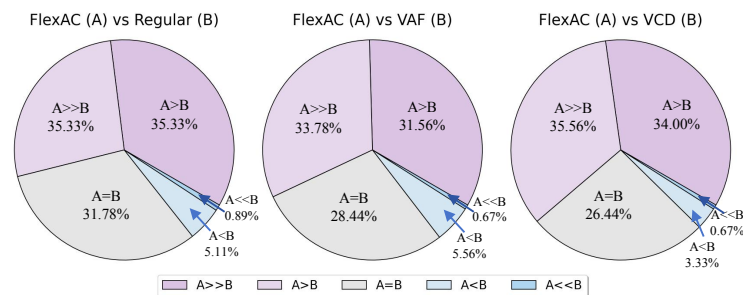


Figure 14: Distribution of user rating preferences when comparing FlexAC with each baseline on the VDAT task. $A = B$ indicates equal preference; $A \gg B$ and $A > B$ mean FlexAC is preferred; $A < < B$ and $A < B$ mean the baseline is preferred. Results show strong preference for FlexAC in most cases.

Task: Test the model's remote association ability

Question: Please generate 10 objects that are **most unrelated** to the image

Your Name (Required)

Please enter your name

Image 21

Comparison: Method A vs Method B (Item 1/90)



© Ryan Taylor

For each image, please compare the two sets of answers. The more unrelated the answers are to the image content, the better we consider the model's performance. Objects generated earlier are more important!!!

Answer A :

book, apple, guitar, cake, tv, car, fridge, plane, computer, lamp

Answer B :

train, platform, trees, fence, signal, signal, signal, signal, signal, signal

Your Evaluation (Which answer is better? Single choice)

☐ A>>B (A is much better than B)

☐ A>B (A is somewhat better than B)

☐ A=B (A and B are equally good)

☐ A<B (B is somewhat better than A)

☐ A<<B (B is much better than A)

Previous

Progress: 1/90 (0%)

Next

Submit Survey

Figure 15: Interface of the user study for evaluating remote association. Participants are presented with an image and two model-generated answers, and asked to judge which set of objects is more unrelated to the image. The label A or B was randomly assigned to FlexAC or baseline in each trial to prevent method identification.

2822

<https://doi.org/10.52202/085713-0088>

D.2 Extended results on Creation-MMBench

To further evaluate FlexAC’s effectiveness in enhancing associative behavior for creative generation, we report additional results on the Creation-MMBench benchmark using two base models: LLaVA-1.5 and DeepSeek-VL2, as shown in Table 4. This benchmark covers four creative subcategories—Literary Writing (LW), Common Functional Writing (CFW), Professional Functional Writing (PFW), and Creative Multimodal Understanding (CMU). For each subtask, we report two metrics: VFS (Visual Fidelity Score), which measures the alignment between the image and the generated response, and Reward, which quantifies creativity improvements relative to the base model (i.e., vanilla LLaVA-1.5 or vanilla DeepSeek-VL2, respectively).

In this experiment, FlexAC is configured to enhance associative behavior, with the goal of generating more creative content. Across both models, FlexAC achieves the highest overall reward scores, demonstrating its effectiveness in promoting creative generation without sacrificing visual grounding. Notably, on DeepSeek-VL2, FlexAC obtains a reward of +10.35 on PFW and +6.73 overall, clearly outperforming all baselines. To test whether performance gains stem from meaningful control or arbitrary perturbation, we also evaluate a variant that injects random vectors into the representation. As shown in the “Random” rows, this leads to large performance drops across all metrics—highlighting that FlexAC’s improvements do not come from noise or randomness, but from targeted modulation of associative features. These results further support FlexAC’s ability to improve creative reasoning across diverse multimodal architectures.

Table 4: **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).

Models	Methods	Overall		LW		CFW		PFW		CMU	
		VFS	Reward	VFS	Reward	VFS	Reward	VFS	Reward	VFS	Reward
LLaVA1.5	Regular	5.32	0.00	6.28	0.00	nan	0.00	4.26	0.00	6.08	0.00
	Random	3.53	-60.49	3.11	-69.58	2.19	-72.22	2.93	-60.35	4.80	-52.69
	Ha-DPO	4.84	-26.41	5.09	-30.00	3.68	-19.72	4.37	-26.23	5.67	-27.22
	VCD	5.56	2.00	6.69	7.08	4.87	5.00	4.86	3.00	6.23	-2.31
	VAF	5.30	-5.86	6.15	-3.54	4.27	-5.00	4.74	-6.34	6.01	-6.67
	FlexAC (Ours)	5.45	4.39	6.52	11.88	4.76	-3.89	4.72	3.62	6.18	4.63
DeepSeek-VL2	Regular	6.12	0.00	6.98	0.00	6.35	0.00	5.71	0.00	6.21	0.00
	Random	2.34	-77.47	1.32	-78.96	3.28	-75.83	1.96	-82.46	2.75	-72.08
	VCD	6.42	4.80	7.37	5.63	6.58	-3.33	5.98	6.40	6.55	5.46
	VAF	6.26	-0.39	6.70	-1.25	6.46	-3.06	5.93	2.46	6.42	-2.13
	FlexAC (Ours)	6.29	6.73	6.76	0.63	6.37	4.17	5.99	10.35	6.44	6.48

D.3 Efficiency comparison

To assess the computational efficiency of FlexAC, we compare the inference runtime of different methods on the Qwen-VL model when evaluating the CHAIR benchmark. Specifically, we measure the total time required to process the full test set under each method’s configuration. As shown in Figure 16, FlexAC incurs only minimal additional overhead compared to the original model, demonstrating that its control mechanism introduces negligible runtime cost. In contrast, VCD exhibits significantly higher latency due to its reliance on dual forward passes—one for the original image and another for a perturbed version—highlighting its inefficiency. These results confirm that FlexAC achieves controllable reasoning with minimal impact on inference speed.

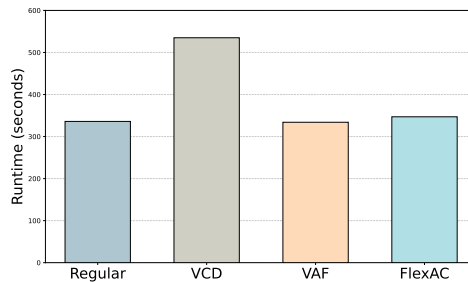


Figure 16: Inference runtime (in seconds) of different methods on Qwen-VL when evaluating CHAIR. FlexAC adds minimal overhead, while VCD incurs high cost due to dual-pass processing.

D.4 Extended Evaluation on General-Purpose Benchmarks

To rigorously evaluate FlexAC’s impact on general capabilities, we extended our analysis to a diverse suite of 11 benchmarks, as detailed in Table 5. Our evaluation spans three representative categories: **general multimodal reasoning**, comprising MM-Vet [43], MMBench [44], SEED-Bench [45], and MMBB [46]; **vision-centric understanding**, which includes RealWorldQA [47], CVBench [48], and AI2D [49]; and **OCR/document-based question answering**, covering TextVQA [50], ChartQA [51], DocVQA [52], and OCRVQA [53]. This comprehensive approach verifies that our control mechanism does not introduce performance degradation.

The results are presented in Table 5. Across all three categories, both FlexAC-C (creativity-enhanced) and FlexAC-P (faithfulness-enhanced) maintain performance closely comparable to the baseline model. This provides strong evidence that our targeted control mechanism effectively modulates associative reasoning without degrading the model’s fundamental, general-purpose capabilities.

Table 5: Performance of FlexAC on an extended suite of 11 general-purpose benchmarks, grouped by capability. The results demonstrate that FlexAC maintains performance comparable to the baseline across general multimodal, vision-centric, and OCR/document tasks, indicating our method does not harm general capabilities.

Category	Benchmark	Regular	FlexAC-C	FlexAC-P
General Multimodal	MM-Vet	39.81	38.17	37.33
	MMBench	0.581	0.598	0.576
	SEED-Bench	0.638	0.625	0.640
	MMBB	0.703	0.678	0.699
Vision-centric	RealWorldQA	0.486	0.490	0.495
	CVBench	0.549	0.524	0.560
	AI2D	0.612	0.614	0.616
OCR & Document	TextVQA	60.66	60.78	59.81
	ChartQA	48.36	49.40	45.92
	DocVQA	57.79	56.85	57.59
	OCRVQA	47.46	49.74	45.83

D.5 Detailed results on POPE

To complement the summary results in Figure 1, we report detailed POPE evaluation metrics across all settings (random, popular, adversarial) and models in Table 6. These include accuracy, precision, recall, and F1 scores for all baselines and our FlexAC variants.

Table 6: Performance on POPE. FlexAC here denotes the version configured to suppress associative behavior, aiming to improve factual accuracy (faithfulness).

Modal	Setting	Method	Accuracy↑	Precision ↑	Recall↑	F1 Score↑
Qwen-VL	Overall	Regular	86.64	92.92	79.33	85.59
		VCD	87.62	91.91	82.53	86.97
		VAF	87.17	91.45	82.0	86.47
		FlexAC (Ours)	87.44	89.31	85.07	87.14
	random	Regular	88.6	97.38	79.33	87.44
		VCD	89.97	97.02	82.53	89.19
		VAF	89.5	96.47	82.0	88.65
		FlexAC (Ours)	90.0	94.38	85.07	89.48
	popular	Regular	87.0	93.7	79.33	85.92
		VCD	87.97	92.6	82.53	87.28
		VAF	87.7	92.55	82.0	86.96
		FlexAC (Ours)	88.47	91.27	85.07	88.06
	adversarial	Regular	84.33	88.15	79.33	83.51
		VCD	84.93	86.69	82.53	84.56
		VAF	84.3	85.95	82.0	83.93
		FlexAC (Ours)	83.87	83.07	85.07	84.06
LLaVA-1.5	Overall	Regular	87.18	91.47	82.0	86.48
		HA-DPO	85.29	92.57	76.73	83.91
		VCD	84.91	88.09	80.73	84.25
		VAF	87.07	87.93	85.93	86.92
		FlexAC (Ours)	87.84	87.13	88.8	87.96
	random	Regular	89.3	96.02	82.0	88.46
		HA-DPO	86.97	96.48	76.73	85.48
		VCD	87.5	93.37	80.73	86.59
		VAF	90.07	93.68	85.93	89.64
		FlexAC (Ours)	91.43	93.74	88.8	91.2
	popular	Regular	87.53	92.2	82.0	86.8
		HA-DPO	86.0	94.19	76.73	84.57
		VCD	85.27	88.78	80.73	84.57
		VAF	87.93	89.51	85.93	87.69
		FlexAC (Ours)	88.7	88.62	88.8	88.71
	adversarial	Regular	84.7	86.68	82.0	84.28
		HA-DPO	82.9	87.53	76.73	81.78
		VCD	81.97	82.78	80.73	81.74
		VAF	83.2	81.48	85.93	83.65
		FlexAC (Ours)	83.4	80.14	88.8	84.25
DeepSeek-VL	Overall	Regular	88.42	88.13	88.8	88.47
		VCD	87.82	87.64	88.07	87.85
		VAF	88.37	87.59	89.4	88.49
		FlexAC (Ours)	88.52	88.36	88.73	88.55
	random	Regular	92.0	94.87	88.8	91.74
		VCD	91.03	93.62	88.07	90.76
		VAF	91.87	94.04	89.4	91.66
		FlexAC (Ours)	91.8	94.53	88.73	91.54
	popular	Regular	88.13	87.63	88.8	88.21
		VCD	87.27	86.68	88.07	87.37
		VAF	88.13	87.19	89.4	88.28
		FlexAC (Ours)	88.37	88.09	88.73	88.41
	adversarial	Regular	85.13	82.73	88.8	85.66
		VCD	85.17	83.24	88.07	85.58
		VAF	85.1	82.32	89.4	85.71
		FlexAC (Ours)	85.4	83.19	88.73	85.87

E Ablation Study

E.1 Effect of dataset Sizes

To analyze the sensitivity of FlexAC to the number of instances used in control vector construction, we vary Top-K over a wide range: $\{1, 5, 10, 20, 50, 100, 200, 500, 1000, 1500, 2000\}$, and evaluate performance on CHAIRs ($\downarrow$), CHAIRi ($\downarrow$), and VDAT ($\uparrow$) using Qwen-VL.

As shown in Figure 17, both FlexAC-P (measured on CHAIR for faithfulness) and FlexAC-C (measured on VDAT for creativity) exhibit similar trends: performance is relatively high but unstable when K is very small, and stabilizes near its peak around $K = 50$. Further increasing K leads to slight performance degradation, likely due to noise introduced by instances. These results highlight the effectiveness of our Instance Selection strategy, which focuses on selecting a small, high-quality set of associative and non-associative samples.

Notably, across all K values, FlexAC-C and FlexAC-P consistently appear on opposite sides of the Regular baseline, reflecting two associative reasoning strength. This clear separation demonstrates FlexAC’s capacity to bidirectionally modulate reasoning behavior, enabling controllable transitions between creative and faithful outputs.

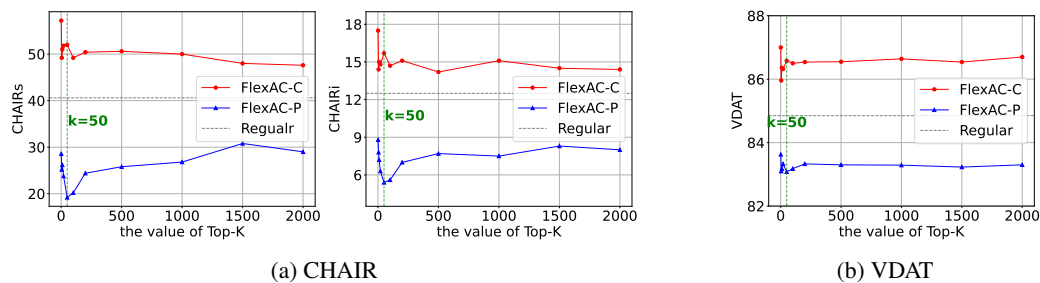


Figure 17: Sensitivity analysis of the Top-K hyperparameter used in general control vector construction on Qwen-VL. We vary the number of selected instances (K) and evaluate performance on CHAIRs, CHAIRi, and VDAT benchmarks.

E.2 Effect of control layer.

To validate the generality of our control layer findings beyond Qwen-VL, we conduct additional layer-wise control effectiveness analysis on LLaVA-1.5 and DeepSeek-VL2, as shown in Appendix Figure 18 and Figure 19. Similar to the trends observed in Qwen-VL, we find that both FlexAC-C and FlexAC-P exhibit consistent improvements in their respective metrics (VDAT and CHAIR) when applied to middle layers. Specifically, the performance peaks around middle layers (layers 10-15) for LLaVA-1.5 and Layers 4-6 for DeepSeek-VL2, which aligns with our feature distance analysis (see Appendix F.4). These results further support our choice of control layers and demonstrate that the effectiveness of FlexAC’s modulation strategy generalizes across diverse MLLM architectures.

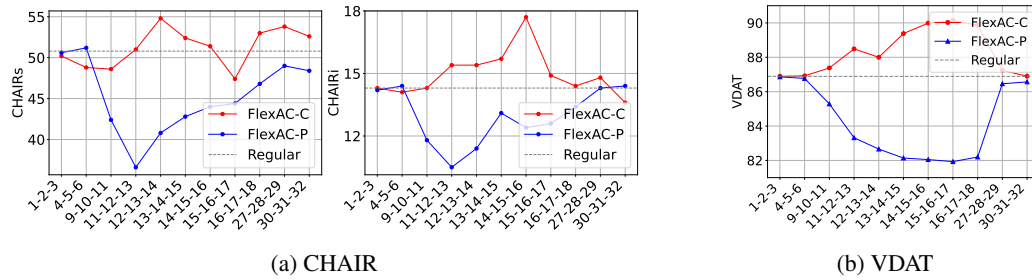


Figure 18: Layer-wise analysis of control effectiveness in FlexAC on LLaVA-1.5. The x-axis represents the control layers, while the y-axis shows the performance of the model on CHAIR and VDAT metrics.

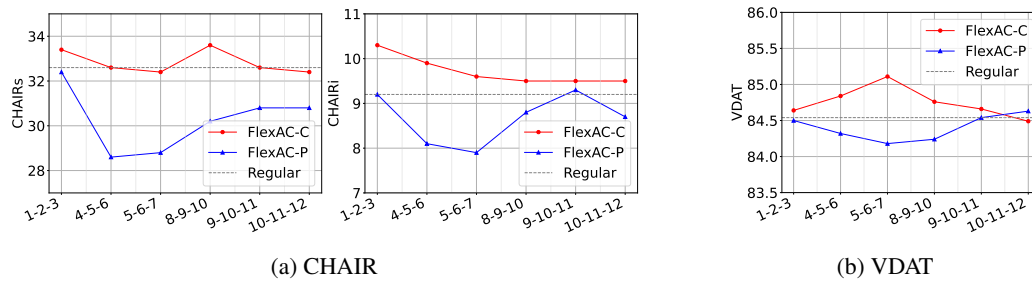


Figure 19: Layer-wise analysis of control effectiveness in FlexAC on DeepSeek-VL2. The x-axis represents the control layers, while the y-axis shows the performance of the model on CHAIR and VDAT metrics.

F Visualizations

F.1 Detailed Feature Representation Analysis Using PCA

To provide a more detailed view of how associative and non-associative representations evolve across the model, we present an expanded version of Figure 4 in Figure 20. This visualization shows the PCA-reduced features layer by layer in LLaVA, with red points representing associative features and blue points representing non-associative ones. Compared to the summary visualization, this version reveals how feature separation progressively emerges across layers. In shallow layers (e.g., Layer 0), the two feature types show significant overlap, indicating similar low-level representations. However, starting from the middle layers (around Layer 12), the separation becomes increasingly distinct, highlighting that the model's associative behavior is primarily shaped in the deeper stages.

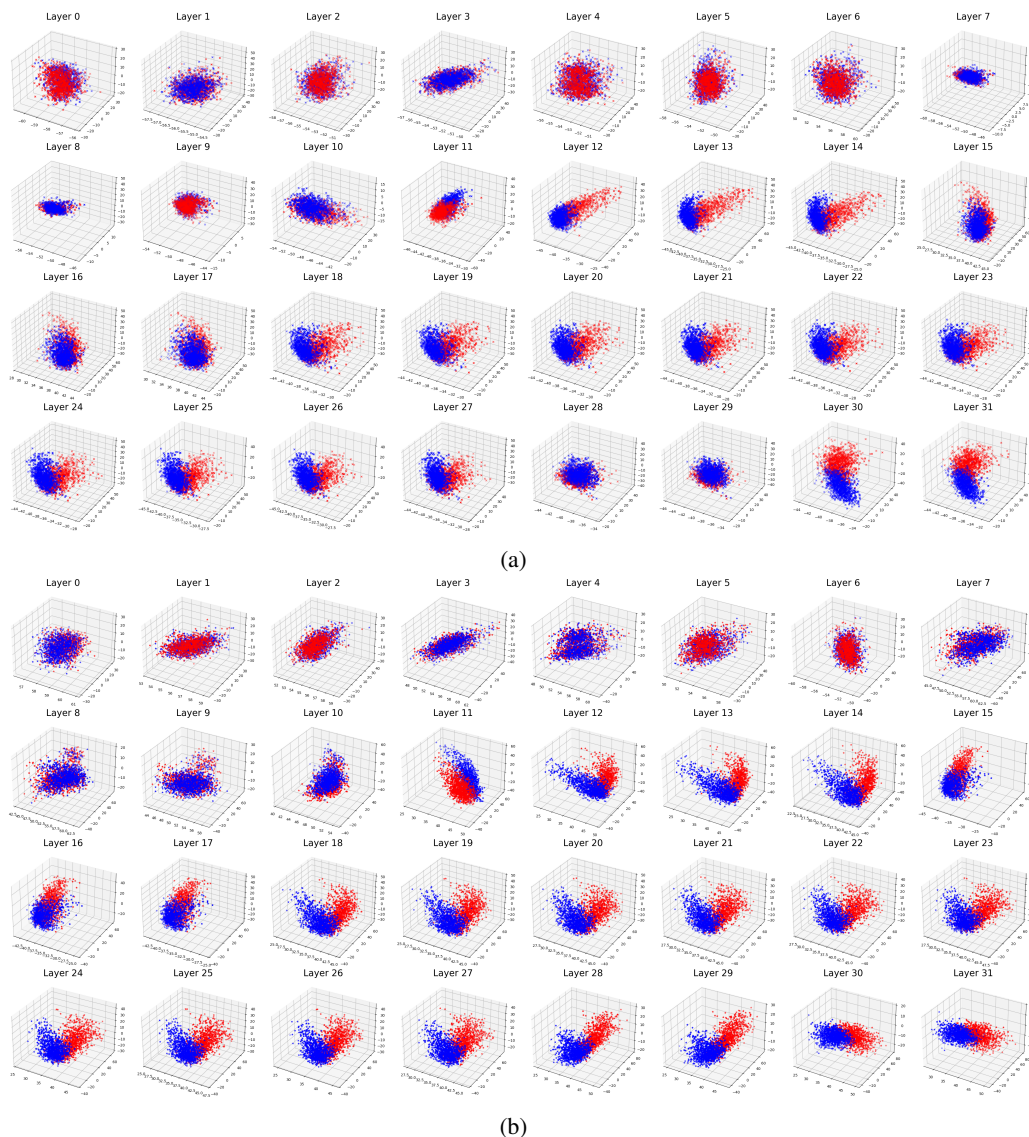


Figure 20: Visualization of feature representations in LLaVA, reduced via PCA, shows the distribution of associative and non-associative data points, represented by red and blue colors, respectively. Subplots (a) and (b) represent the results for different option orders. In deeper layers, the red and blue points exhibit clearer separation, indicating enhanced differentiation between associative and non-associative representations.

F.2 Detailed Layer Intervention for Association Localization

To gain a more comprehensive understanding of how different layers contribute to associative content generation, we expanded on the analysis presented in Figures 2c and 2d by examining each layer individually. As shown in Figure 21 and Figure 22, the detailed version presented here provides a layer-by-layer breakdown of how the interventions affect the model's internal representations.

In each subplot of this detailed version, we intervened at a specific layer (denoted by the subplot title, e.g., “Layer 0,” “Layer 1,” etc.) by replacing its associative features with non-associative features. We then analyzed the impact of this intervention on feature distances across all layers. Figure 21 and Figure 22 shows that when shallow layers (e.g., layers before Layer 11) are replaced, the feature distances in subsequent layers do not change significantly. However, when middle layers such as Layer 11 are replaced, the subsequent feature distances drop sharply, indicating that these layers have a crucial impact on the model's associative tendencies. In contrast, when deeper layers (e.g., layers after Layer 14) are replaced, the changes in subsequent layers become more stable, suggesting that deeper layers have a weaker influence on associative tendencies.

This detailed analysis highlights that replacing features at specific layers has a distinct influence on subsequent layers, with the greatest impact often observed in middle layers. This is consistent with the averaged results in Figures 2c and 2d, which pointed towards the critical role of middle layers in maintaining associative characteristics.

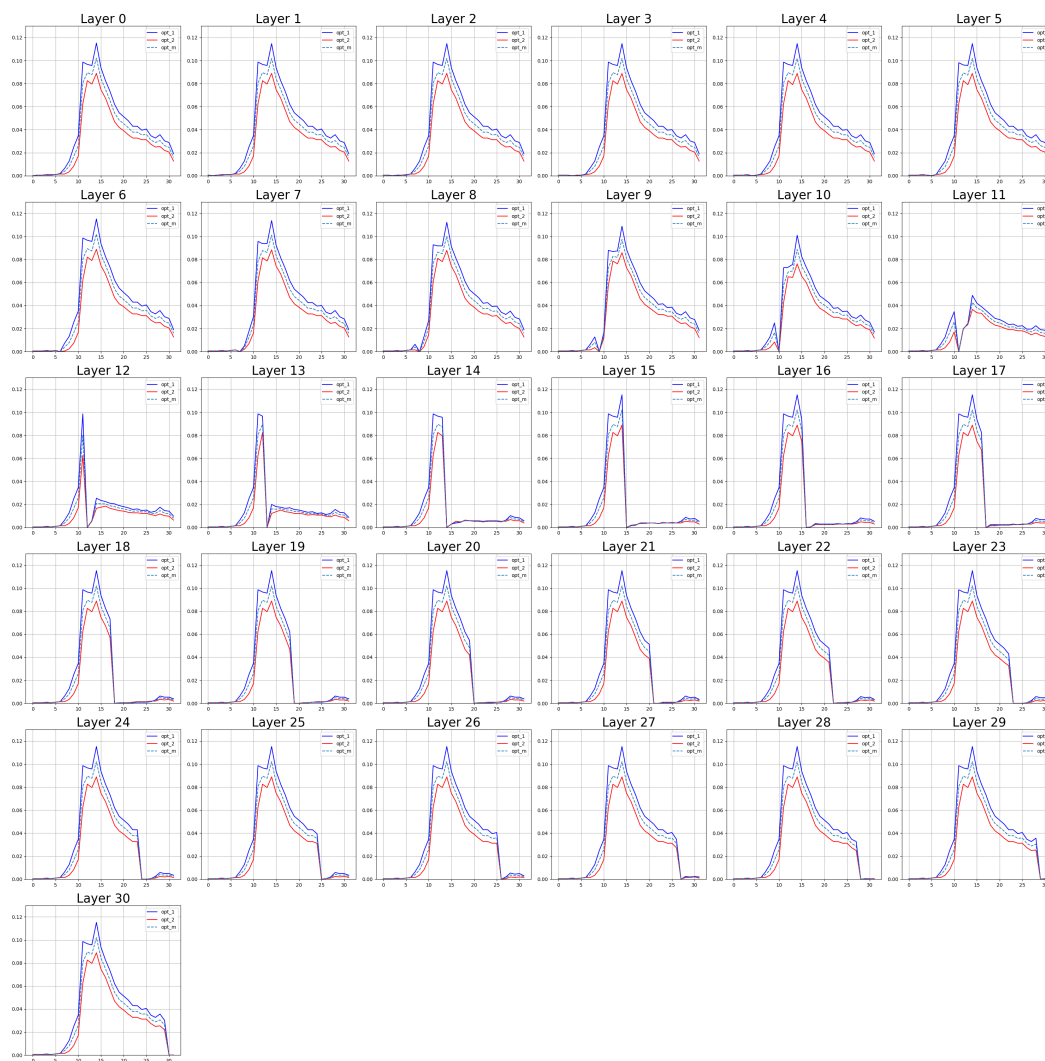


Figure 21: Feature distance analysis across layers after layer intervention. For example, the subplot titled “Layer 12” shows the feature distances across all layers after replacing associative features at Layer 12 with non-associative features. The X-axis represents the different layers, and the Y-axis represents the **Cosine distance** between associative and non-associative data.

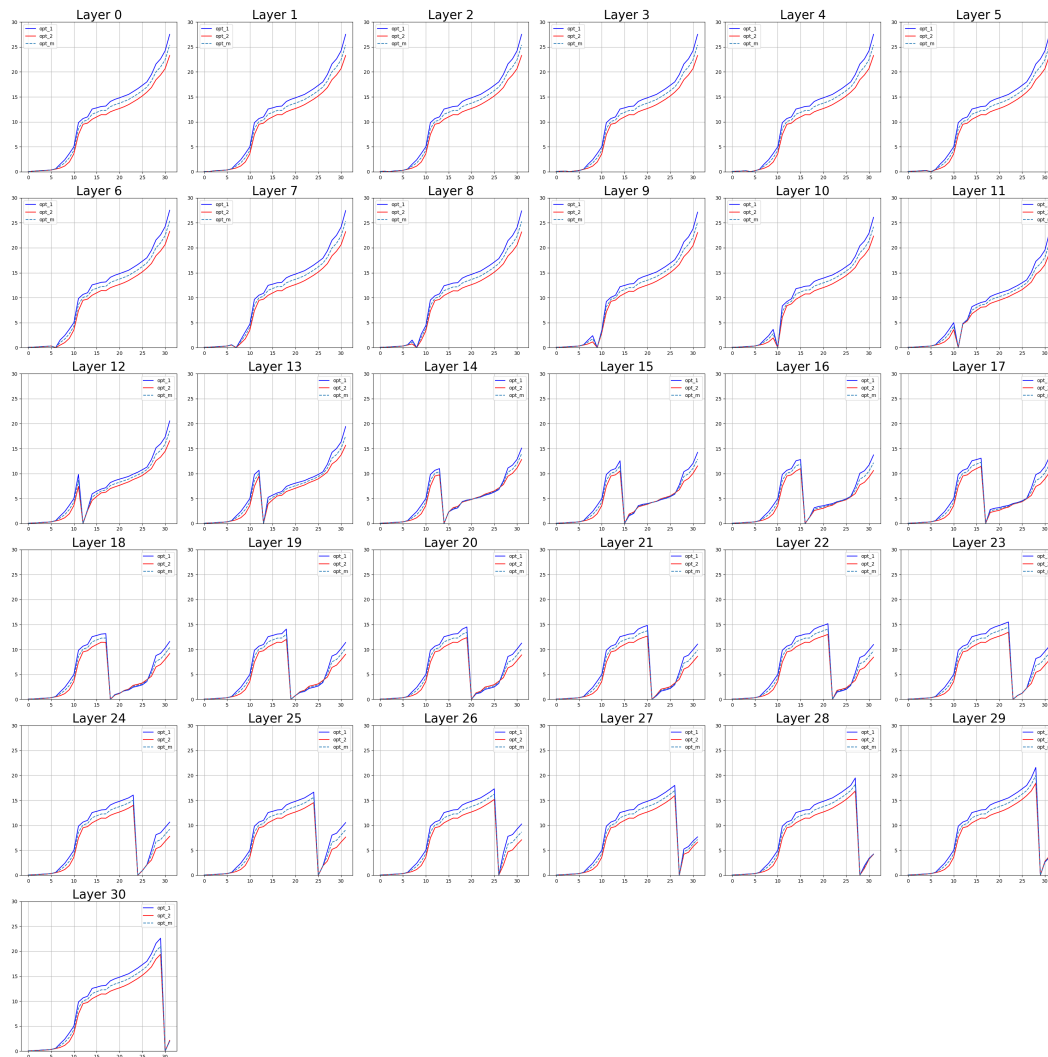


Figure 22: Feature distance analysis across layers after layer intervention. For example, the subplot titled “Layer 12” shows the feature distances across all layers after replacing associative features at Layer 12 with non-associative features. The X-axis represents the different layers, and the Y-axis represents the **Euclidean distance** between associative and non-associative data.

E.3 Visualization of more examples

We visualize sample outputs from both the creativity- and precision-enhancing variants of FlexAC on Qwen-VL. Figure 23 shows examples from the VDAT benchmark, highlighting differences in associative strength. Additional qualitative results on Creation-MMBench are provided in Figures 24–29, illustrating FlexAC’s ability to adjust associative reasoning across creative tasks.

VDAT	Q: List 10 unrelated, tangible nouns that are not connected to the image or each other.
	 grill, food, plate, vegetable, meat, oven, table, kitchen, cooking, dinner  FlexAC-P ($\alpha = -1$)  book, chair, tv, bag, car, computer, phone, clock, guitar, cake  FlexAC-C ($\alpha = 1$)
	 plate, broccoli, food, plastic, container, table, fork, dinner, meal, plate  FlexAC-P ($\alpha = -1$)  banana, car, book, cake, chair, cup, dog, fork, guitar, house  FlexAC-C ($\alpha = 1$)
	 lamp post, traffic light, building, street, sidewalk, traffic, people, traffic signal, street light, street sign.  FlexAC-P ($\alpha = -1$)  umbrella, elephant, pizza, guitar, book, cake, car, chair, cat, lamp  FlexAC-C ($\alpha = 1$)
	 kitchen, dishwasher, cabinets, windows, curtains, sink, counter, floor, wall, light  FlexAC-P ($\alpha = -1$)  table, chair, book, window, curtain, wall, floor, counter, cabinet, drawer  FlexAC-C ($\alpha = 1$)
	 zebra, grass, field, animal, grassy, meadow, herd, grazing, grasslands, wild animal  FlexAC-P ($\alpha = -1$)  book, cat, car, tree, pizza, guitar, moon, elephant, cake, plane  FlexAC-C ($\alpha = 1$)
	 motorcycle, sidewalk, brick, pavement, bike, street, sidewalk, bike, street, pavement  FlexAC-P ($\alpha = -1$)  book, chair, car, clock, cake, cat, couch, cup, door, dog  FlexAC-C ($\alpha = 1$)
	 watermelon, scooter, helmet, market, people, car, street, helmet, scooter, market  FlexAC-P ($\alpha = -1$)  apple, car, truck, helmet, bag, chair, table, umbrella, person, market  FlexAC-C ($\alpha = 1$)

Figure 23: **Visualization of FlexAC’s Control on VDAT**, showing how FlexAC-P (faithfulness) and FlexAC-C (creativity) adjust the level of associative strength in the generated responses.

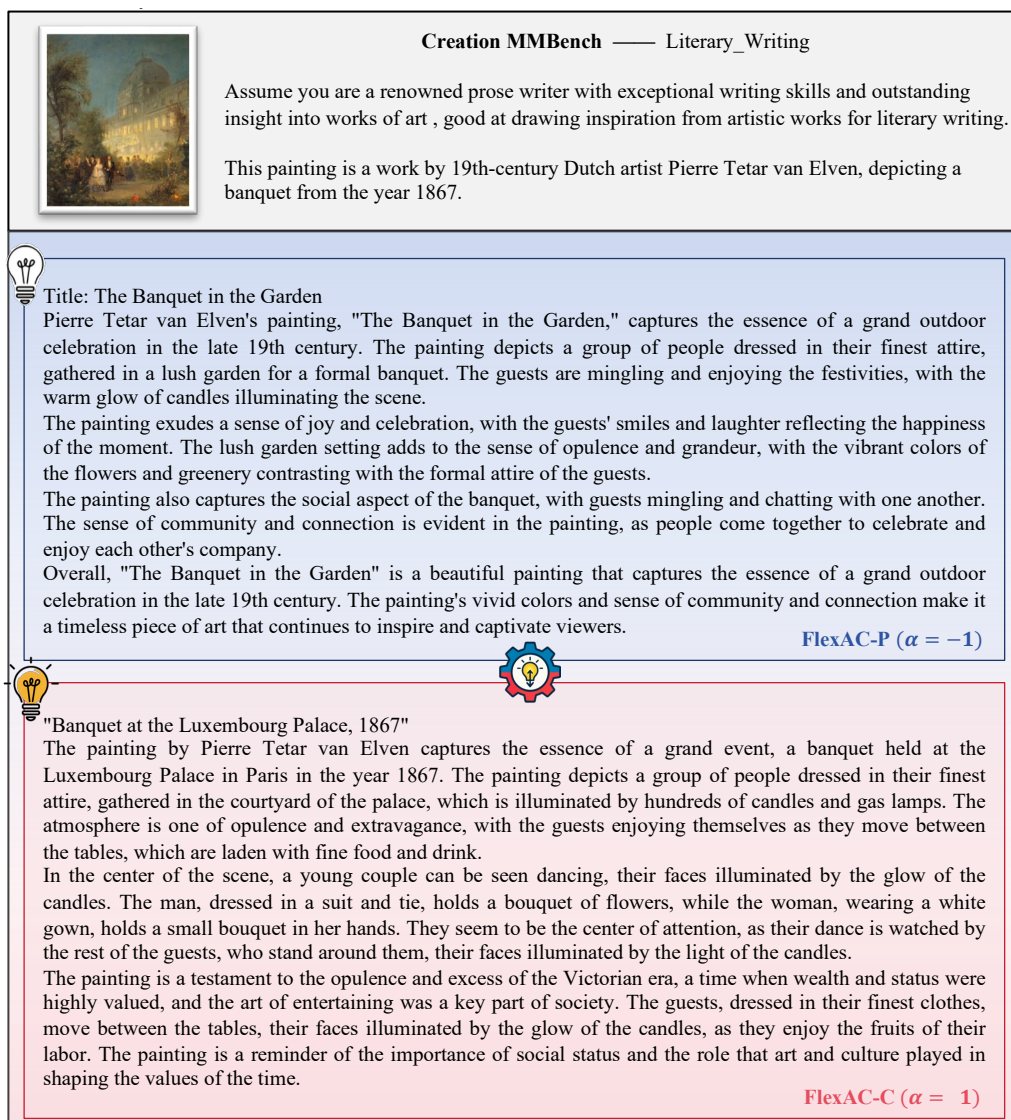


Figure 24: **Visualization of FlexAC's Control on Creation MMBench**, showing how FlexAC-P (faithfulness) and FlexAC-C (creativity) adjust the level of associative strength in the generated responses.



Figure 25: **Visualization of FlexAC's Control on Creation MMBench**, showing how FlexAC-P (faithfulness) and FlexAC-C (creativity) adjust the level of associative strength in the generated responses.

Creation MMBench — Literary_Writing

Assume you are talented children's animation creators who have a profound insight into the story plot and a unique understanding of the picture content, and have the ability to continue to narrate and complete a wonderful story according to the content of the previous foreshadows.

These are the first few scenes of the children's animation: A group of dinosaur friends is standing on a grassy field.



As the sun began to set, the group of dinosaur friends, including a cute little T-Rex, a friendly Triceratops, and a wise old Stegosaurus, gathered together on the grassy field. They had been on a long adventure, exploring the lush green landscapes and encountering new friends along the way. As they sat together, they shared stories of their adventures and the fun they had had together. The T-Rex, who was known for his love of adventure, spoke of the excitement of discovering new lands and meeting new creatures. The Triceratops, who was known for his calm and steady nature, spoke of the joy of simply being surrounded by the beauty of nature. The Stegosaurus, who was known for his wise words, spoke of the importance of friendship and the value of the bonds we form with others. As the sun set behind them, the group of dinosaur friends sat in silence, content in the knowledge that they had each other and the memories of their adventures.

FlexAC-P ($\alpha = -1$)

As the group of dinosaur friends stands on the grassy field, they notice a small, colorful butterfly flitting about. The baby dinosaur, who is still a bit too small to walk, is fascinated by the fluttering insect.

"Wow, look at that pretty butterfly!" the baby dinosaur says, pointing excitedly at the tiny creature.

"I've never seen one before," says the baby triceratops, who is standing next to the baby dinosaur. "Do you think it's a dragonfly?"

"No, it's a butterfly," the baby pteranodon corrects, with a grin on his face. "Butterflies are way cooler than dragonflies."

The baby tyrannosaurus, who is standing at the back of the group, chimes in with a giggle. "I want to catch one! Can we, Mommy?"

The mother triceratops, who is standing nearby, gives the baby a warning look. "We can't catch the butterfly, honey. It's too fast and too fragile. We have to be careful not to hurt it."

The baby dinosaur pouts. "But I want to play with it! Can't we just touch it?"

The mother dinosaur sighs. "I'm sorry, but we can't touch the butterfly. It's too delicate. We have to be careful not to harm it."

The baby dinosaur huffs, but the mother dinosaur is right. Butterflies are very fragile creatures, and they need to be treated with care. The baby dinosaur decides to watch the butterfly as it flutters around them, fascinated by its colors and movements.

As they watch, the baby dinosaur notices a small, wiggly worm crawling nearby. "Hey, look at that!" he says, pointing excitedly at the worm. "It's so cool! Can we touch it?"

The mother dinosaur shakes her head. "No, honey. That's a worm. They're not very pretty, but they're important to the ecosystem. We have to be careful not to harm them, too."

The baby dinosaur pouts again, but the mother dinosaur is right. Worms are important to the soil and the plants, and they help the plants grow. The baby dinosaur decides to watch the worm as it wriggles around them, fascinated by its movements.

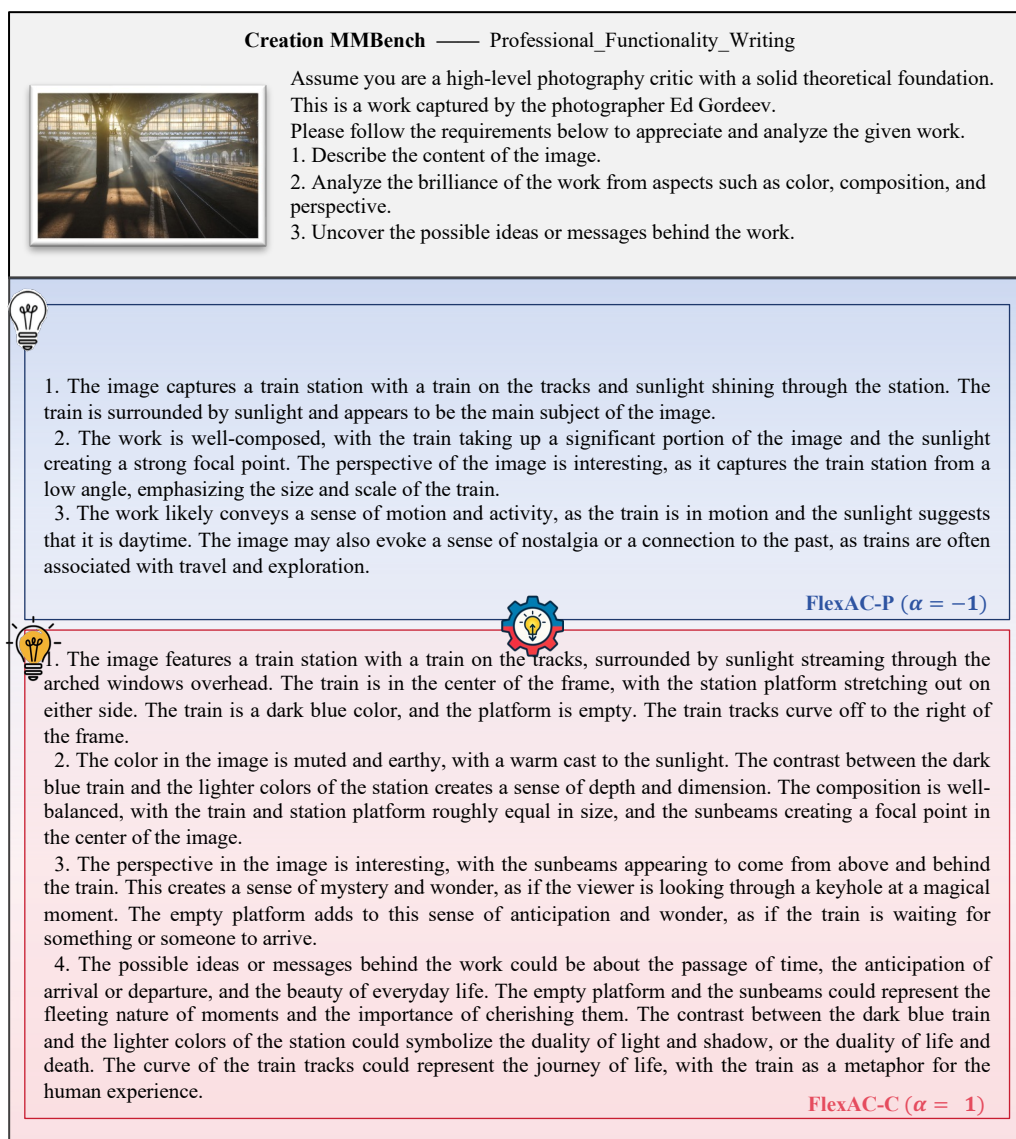
As they watch, the baby dinosaur notices a small, colorful flower blooming nearby. "Wow, look at that!" he says, pointing excitedly at the flower. "It's so pretty! Can

FlexAC-C ($\alpha = 1$)

Figure 26: **Visualization of FlexAC's Control on Creation MMBench**, showing how FlexAC-P (faithfulness) and FlexAC-C (creativity) adjust the level of associative strength in the generated



Figure 27: **Visualization of FlexAC's Control on Creation MMBench**, showing how FlexAC-P (faithfulness) and FlexAC-C (creativity) adjust the level of associative strength in the generated responses.





Assume you are an experienced adventurer. You have the habit of writing a travel journal every day. You tell about the beautiful things you have seen, the interesting things you have encountered, and the feelings you have had during your trip.

This is a scenic photo taken during the tour today.

Please follow the requirements below to write today's travel journal.

1. Describe the Scenery: Use sensory details to vividly convey what you saw, heard, or felt in the landscape.
2. Chronological Storytelling: Narrate the day's events from a first-person perspective, starting with your arrival and ending with your reflections.
3. Highlight Adventure: Include one or two adventurous moments or challenges (e.g., navigating tough terrain or encountering wildlife).
4. Emotional Reflection: Share how the scenery and the day's experiences made you feel as an adventurer.
5. Tie to the Photo: Explain why you captured this photo and how it connects to the most exciting or meaningful part of your day.



Today was an incredible day of adventure and exploration. I arrived at my destination and was immediately struck by the stunning scenery. The photo I have captured perfectly captures the beauty of the landscape, with the bright green Northern Lights dancing across the sky.

As I explored the area, I heard the sound of rushing water, which led me to a beautiful waterfall. The sound of the water was so peaceful and calming, it was a welcome break from the excitement of the day.

One of the most exciting moments of the day was when I encountered a group of hikers who were also out to see the Northern Lights. We all huddled together and watched in awe as the Northern Lights lit up the sky in a stunning display.

As the day came to an end, I reflected on the beauty of nature and the sense of adventure that comes with exploring new places. I am so grateful to have captured this moment in time, and I know that it will be a memory that I will treasure for years to come.

FlexAC-P ($\alpha = -1$)



Day 3: Northern Lights Tour

I woke up early this morning, excited to embark on a Northern Lights tour. The sky was clear and the stars were shining bright as I made my way to the tour bus. The tour guide, a friendly and knowledgeable fellow, greeted us all and gave us a quick briefing on what to expect during the tour.

We drove for a few hours, driving through the rugged terrain of Iceland, passing by stunning landscapes and waterfalls. The tour group was made up of travelers from all over the world, and we all shared our excitement and anticipation for what was to come.

As the night fell, the sky turned a beautiful shade of pink and purple, and we could see the Northern Lights starting to dance in the distance. The tour guide explained that the Northern Lights are caused by charged particles from the sun colliding with the Earth's atmosphere, creating a beautiful display of light and color.

We drove further into the wilderness, and soon we were surrounded by nothing but the stars and the Northern Lights. The display was like nothing I had ever seen before - a beautiful, otherworldly sight that left me in awe. The Northern Lights were like a dance, moving and changing colors in a way that was almost too beautiful to be real.

We spent the next few hours watching the Northern Lights, taking pictures and simply enjoying the moment. It was a truly magical experience, and one that I will never forget.

As the night came to an end, we headed back to our hotel, tired but happy. I felt grateful to have been able to witness such a beautiful natural phenomenon, and I knew that this would be a trip that I would always remember.

I took a photo of the Northern Lights, knowing that this would be a memory that I would always treasure. The Northern Lights were like a reminder that there is so much more to this world than we can ever imagine, and that every day, we have the opportunity to be amazed by something new.

I went to bed that night feeling grateful and humbled by the beauty of the world around us. I knew that I would never forget this trip, and that I would always look back on it with a sense of wonder and awe.

FlexAC-C ($\alpha = 1$)

Figure 29: **Visualization of FlexAC's Control on Creation MMBench**, showing how FlexAC-P (faithfulness) and FlexAC-C (creativity) adjust the level of associative strength in the generated responses.

F.4 Feature Distance Analysis on Additional Models

To complement the analysis in Section 3.1, we extend the feature distance evaluation to two additional MLLMs: Qwen-VL and Deepseek-VL2. As in the main study, we compute the cosine and Euclidean distances between associative and non-associative representations extracted from each transformer layer. The results are shown in Figure 30.

Consistent with our findings on LLaVA, we observe that cosine distance peaks in the middle layers, while Euclidean distance gradually increases throughout the network. These patterns reinforce the conclusion that associative behavior primarily emerges and diverges in the middle layers, while deep layers largely propagate those effects.

Importantly, this analysis also informs the design of our control strategy. In Qwen-VL, the middle layers are approximately 13 – 20, and in DeepSeek-VL2, 3 – 7. Accordingly, we select Layers 15 – 17 for Qwen-VL and Layers 4 – 6 for DeepSeek-VL2 as control points in FlexAC. These ranges correspond to the regions of maximal divergence between associative and non-associative features, enabling targeted yet lightweight intervention.

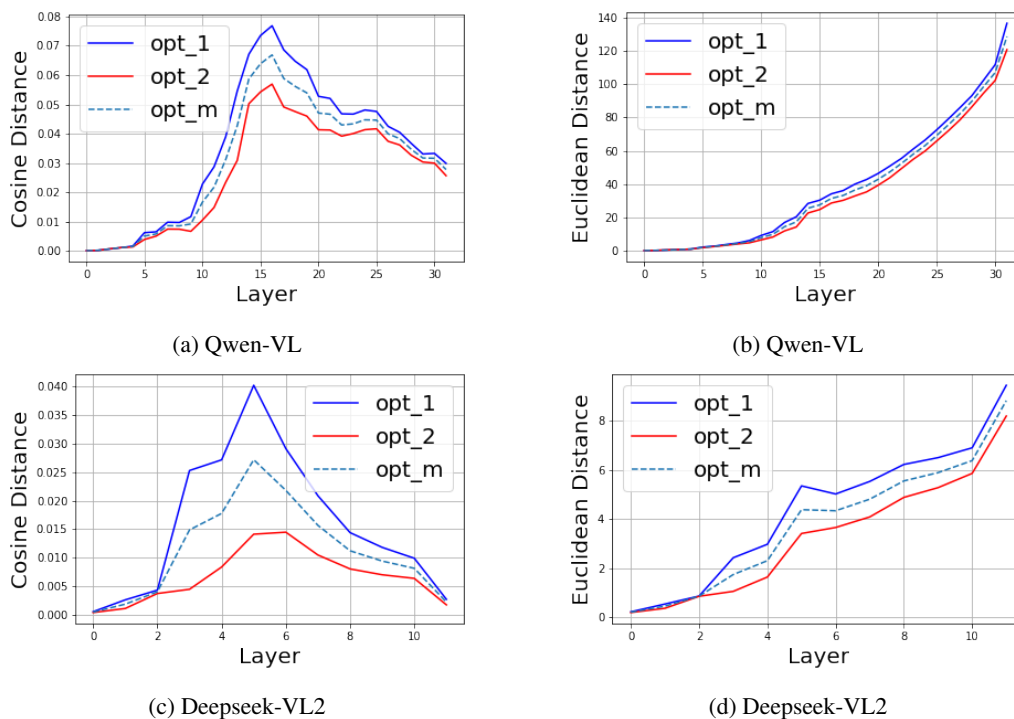


Figure 30: Layer-wise feature distance trends between associative and non-associative representations on Qwen-VL and Deepseek-VL2, extending the LLaVA results from Section 3.1.