Correlation-Flip Visual Injection for Hallucination Mitigation in Large Vision-Language Models

Anonymous ACL submission

Abstract

Large vision-language models (LVLMs) often hallucinate due to over-reliance on language priors. Existing methods implicitly assume that visual tokens maintain consistent semantics across layers and thus indiscriminately inject shallow-layer visual information into deeper layers. We identify a new phenomenon, termed correlation flip, where a visual token positively correlated with the decoding context in an intermediate layer becomes negatively correlated in the final layer, indicating that correct visual evidence is suppressed before prediction. Based on this finding, we propose Correlation-Flip Visual Injection (CFVI), a simple and training-free decoding-time method that selectively identifies correlation-flip tokens and injects corrective visual signals into the final-layer hidden state to restore suppressed evidence and mitigate hallucinations. CFVI can be directly applied to existing LVLMs, and introduces minimal computational overhead. Experiments on multiple benchmarks demonstrate that CFVI achieves strong performance while preserving language fluency.

1 Introduction

Large vision-language models (LVLMs) have achieved strong performance on a wide range of multimodal tasks (Alayrac et al., 2022; Li et al., 2023a; Liu et al., 2023). However, their generations still suffer from hallucinations, where model outputs mention content that is inconsistent with the image or absent from it (Li et al., 2023b; Liu et al., 2024b; Rohrbach et al., 2018). Since many LVLMs are built on large language models (LLMs) (Liu et al., 2024a; Jiang et al., 2024; Bai et al., 2023), they naturally inherit strong language priors, which constitute an important source of hallucination (Khayatan et al., 2026).

Recent work proposes various strategies to mitigate hallucinations. Training-free methods usually intervene in inference to suppress hallucinated content. Contrastive decoding and its variants calibrate predictive probabilities by constructing differential distributions, such as image-missing views or counterfactual images, thereby reducing over-reliance on language priors (Leng et al., 2024; Wang et al., 2024; Zhu et al., 2025; Huo et al., 2024; Fang et al., 2026). These methods often introduce substantial computational overhead.

A complementary direction enhances visual representations. Improving sensitivity to low-level visual features is often viewed as a fundamental way to reduce hallucinations (Wang et al., 2025). Existing approaches introduce multimodal feature compensation by injecting condensed visual features into feed-forward networks (Zhu et al., 2026; Zou et al., 2024) or by using residual connections to fuse accurate visual information from shallow layers (Chen et al., 2026). However, these methods usually inject visual information from intermediate layers into deeper layers without precise selection. They also share an implicit assumption: the semantic role of a visual token remains relatively consistent across layers. In other words, if a visual token reflects certain visual evidence in shallow or intermediate layers, deeper layers are expected to maintain a consistent understanding of what the model has perceived, with only the influence strength changing rather than the semantic interpretation being fundamentally reversed.

Our observation shows that this assumption does not always hold. During the forward pass, some visual tokens are positively correlated with the current decoding context in intermediate layers, but they gradually become negatively correlated as information propagates to deeper layers. We call this phenomenon "correlation flip". Specifically, the model perceives the right evidence in intermediate layers, but during final-layer generation, this visual evidence is suppressed, diluted, or even reversed by language priors, causing the final output to devi-

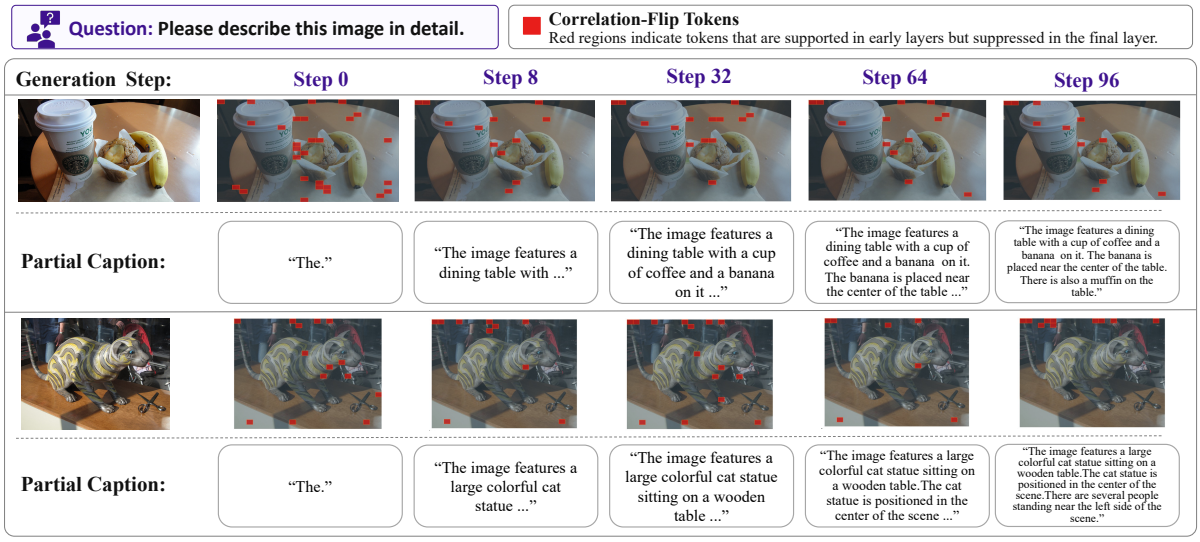

Figure 1: Spatial distribution of correlation-flip tokens during generation with CFVI. These tokens are abundant throughout decoding and concentrate around salient image regions that correspond to key objects.

ate from image facts. As shown in Figure 1, many correlation-flip tokens appear during the model generation process and concentrate around salient regions corresponding to key objects. We argue that correlation flip is an informative signal associated with hallucination in LVLMs.

Based on this finding, we propose Correlation-Flip Visual Injection (CFVI), a targeted visual correction framework for hallucination mitigation. CFVI identifies correlation-flip tokens by measuring cross-layer cosine correlations between visual tokens and decoding hidden states, selecting tokens that are positively correlated in an intermediate layer but negatively correlated in the final layer. It then computes adaptive weights based on the flip magnitude and aggregates these tokens into a corrective visual signal, which is injected into the final-layer hidden state to influence language prediction. Unlike methods that globally enhance visual representations, CFVI selectively corrects only visually grounded tokens whose semantic alignment reverses during decoding, enabling more precise hallucination mitigation.

This design improves both visual feature quality and the use of visual information by the model. We validate CFVI on multiple hallucination benchmarks and show that it achieves strong efficiency. Experiments demonstrate that precise injection improves faithfulness to image facts while preserving fluent generation. Our main contributions are summarized as follows:

- We discover and formalize the "correlation flip" phenomenon of visual tokens: the same token is positively correlated with the context in intermediate layers but becomes negatively correlated in the final layer. This reveals, from a new perspective, that hallucinations in LVLMs may be associated with "seeing but ultimately failing to use" rather than merely "not seeing."

- We propose Correlation-Flip Visual Injection (CFVI), a targeted decoding-time correction method that identifies visual tokens with reversed cross-layer semantic alignment, and adaptively injects aggregated corrective signals into the final-layer hidden state to recover suppressed visual evidence for fine-grained hallucination mitigation.

- Our method not only improves visual feature quality but also enhances the use of visual information by the model. It achieves strong performance and maintains high efficiency across multiple benchmarks, demonstrating that correlation flip is a useful signal for mitigating hallucinations in LVLMs.

2 Related Work

2.1 Hallucination in LVLMs

Hallucination in large vision-language models (LVLMs) is commonly understood as text generation that is inconsistent with image content. Instead, it results from multiple internal mechanisms. Prior

studies show that LVLMs rely on both visual evidence and language priors during generation, and language priors can dominate in many cases (Khayatan et al., 2026; Li et al., 2026; Zhang et al., 2026). Consequently, even when a model receives correct visual information, its final output may still deviate from image facts. From the perspective of representation learning, hallucination is also related to cross-layer representation evolution (Jiang et al., 2025; Lin et al., 2026). Visual information often retains strong perceptual properties in intermediate layers, while deeper layers gradually couple it with language modeling objectives (Favero et al., 2024).

2.2 Hallucination Mitigation in LVLMs

Existing training-free methods for hallucination mitigation in LVLMs typically reduce hallucinations by modifying generation strategies, reranking candidate outputs, or introducing confidence-based constraints (Leng et al., 2024; Wang et al., 2024; Zhu et al., 2025; Huo et al., 2024; Fang et al., 2026; Min et al., 2025). Among these, a line of work improves grounding by enhancing or injecting visual evidence (Wang et al., 2025; Zhu et al., 2026; Zou et al., 2024; Chen et al., 2026), which is also the focus of this paper. These methods are appealing because they require no model retraining and directly address a key cause of hallucination: language priors often dominate the generation process in LVLMs, making visual signals difficult to fully preserve. As a result, visual injection has emerged as a simple yet effective solution. However, these methods generally assume that visual features maintain relatively stable semantic roles across layers and therefore enhance visual information uniformly, without distinguishing which evidence is actually suppressed in deeper layers.

In contrast, CFVI explicitly identifies visual tokens whose correlations flip between intermediate and final layers and injects only the key visual evidence suppressed by deep layers into the final hidden state. This avoids noise from indiscriminate enhancement and improves the use of visual information by the model.

3 Our Observations

3.1 Correlation-Flip Tokens

To analyze how visual grounding evolves during cross-layer propagation, we first note that the decoding hidden state $\mathbf{h}_t^\ell$ encodes the model's current contextual belief over the next-token prediction, while the visual token representation $\mathbf{v}_i^\ell$ reflects localized visual evidence. A natural way to measure whether a visual token is supportive of the current generation decision is to evaluate their directional alignment in the shared representation space. Therefore, we define the correlation between a visual token and the decoding context as

$$C_{i,t}^\ell = \cos(\mathbf{v}_i^\ell, \mathbf{h}_t^\ell), \tag{1}$$

where $C_{i,t}^\ell$ measures the semantic alignment between the i-th visual token and the current generation state in layer ℓ.

We choose an anchor layer ℓ_a that retains relatively strong visual grounding information (Favero et al., 2024). In practice, the anchor layer can be selected from middle-to-late layers, which have integrated prompt context but have not fully collapsed into final decisions dominated by language priors. For each visual token at decoding step t, we compute its alignment with the current decoding context in both the anchor layer and the final layer:

$$C_{i,t}^a = \cos(\mathbf{v}_i^{\ell_a}, \mathbf{h}_t^{\ell_a}), \quad C_{i,t}^L = \cos(\mathbf{v}_i^L, \mathbf{h}_t^L). \tag{2}$$

A positive $C_{i,t}^a$ indicates that the token may provide useful visual support for the next-token prediction at the anchor layer. In contrast, a negative $C_{i,t}^L$ indicates that the token becomes semantically opposed to the final decision state. Such a token is defined as a correlation-flip token.

3.2 Flip Intensity Is Correlated with Hallucination Severity

We observe many correlation-flip tokens during LVLM generation. Figure 1 visualizes their spatial distribution across decoding steps. Two patterns emerge. First, correlation-flip tokens are abundant and they concentrate around key regions in the image. Second, the number of flip tokens gradually decreases as generation continues, suggesting that successful use of these tokens in earlier decoding steps alleviates visual grounding drift. This phenomenon indicates that suppressed visual evidence is closely associated with LVLM hallucination.

To further verify this observation, we study the relation between flip intensity and hallucination severity. For token i, we define its flip strength as

$$s_{i,t} = C_{i,t}^a - C_{i,t}^L. \tag{3}$$

For decoding step t, the overall flip intensity is

$$S_t = \frac{1}{|\mathcal{V}|} \sum_{i \in \mathcal{V}} s_{i,t}, \tag{4}$$

Flip intensity	CHAIR (CHAIR$_s$)	POPE (F$_1$)
low	5.5	91.52
mid	6.6	90.45
high	7.0	87.05

Table 1: Relationship between Correlation-Flip Intensity and Hallucination Severity. Mean hallucination severity increases monotonically with flip intensity.

where $\mathcal{V}$ denotes the set of visual tokens. We use CHAIR$_s$ from the CHAIR data as the hallucination severity metric. We use CHAIR$_s$ as the evaluation metric for hallucination severity, derived from the CHAIR dataset. For each image and its corresponding model-generated caption, we compute the flip intensity and partition all captions into three groups—low, medium, and high flip intensity—based on the interquartile range (IQR), using Q_1 and Q_3 as thresholds. As shown in Table 1, the mean hallucination severity increases monotonically in each group. Samples with stronger flip intensity exhibit more severe hallucination, which provides additional evidence for the positive relation between the two factors. We applied the same bucketing approach to the POPE dataset (random split) and recorded the evaluation metrics for hallucination severity (F_1 score) within each group in the table. It was observed that samples with higher flip intensity exhibit more severe hallucination and lower F_1 scores.

This result shows that cross-layer correlation flip is not random, but directly related to hallucination generation. It further motivates the central idea of CFVI: if the model restores suppressed visual evidence, it may recover more stable visual grounding and reduce hallucination.

4 The Proposed Method

We propose Correlation-Flip Visual Injection (CFVI), an inference-time method for mitigating hallucinations in LVLMs. Figure 2 presents the overall framework.

4.1 Correlation-Flip Token Selection

CFVI first identifies visual tokens that change from supporting the current generation context to suppressing it during cross-layer propagation. Based on the observation above, we define the set of correlation-flip tokens as

$$\mathcal{F}_t = \{i \in \mathcal{V} \mid C^a_{i,t} > \tau_{\text{pos}},\ C^L_{i,t} < -\tau_{\text{neg}}\}, \quad (5)$$

where τ_{pos} and τ_{neg} control the positive correlation strength at the anchor layer and the negative correlation strength at the final layer, respectively. A token is selected only when it satisfies both conditions. The first condition ensures that the token is not arbitrary noise but is indeed relevant to the current generation context at an earlier stage. The second condition ensures that this originally useful visual evidence becomes inconsistent with the final generation state. The set $\mathcal{F}_t$ therefore captures visual tokens that are likely suppressed by deep language priors.

This selection mechanism differs from conventional token-importance estimation. CFVI does not select tokens merely according to high similarity, high attention weight, or large activation magnitude. Instead, it selects tokens according to a cross-layer contradiction. The selected tokens are especially valuable for hallucination mitigation because they reveal where the final decision deviates from earlier visual grounding.

4.2 Flip-Aware Visual Enhancement

After obtaining $\mathcal{F}_t$, CFVI constructs a compact visual correction vector from these tokens. We assign each flip token a weight according to both its anchor-layer support strength and its final-layer suppression strength. For each $i \in \mathcal{F}_t$, we define an unnormalized flip score as

$$q_{i,t} = \text{ReLU}(C^a_{i,t}) \cdot \text{ReLU}(-C^L_{i,t}). \quad (6)$$

This score has an intuitive interpretation: a token receives a high score only when it is both originally relevant and finally suppressed. We use multiplication rather than subtraction or a ratio because the desired flip strength should reflect the simultaneous satisfaction of both conditions, rather than assigning a high score when only one side is strong. Multiplication also emphasizes tokens that are strong on both sides. We normalize these scores with a softmax:

$$\alpha_{i,t} = \frac{\exp(q_{i,t}/T_w)}{\sum_{j\in\mathcal{F}_t}\exp(q_{j,t}/T_w) + \epsilon}, \quad (7)$$

where T_w is the weighting temperature and ϵ is a small constant for numerical stability. CFVI then directly injects the weighted anchor-layer visual evidence into the final-layer generation state:

$$\tilde{\mathbf{h}}^L_t = \mathbf{h}^L_t + \lambda \sum_{i\in\mathcal{F}_t} \alpha_{i,t}\mathbf{v}^{\ell_a}_i, \quad (8)$$

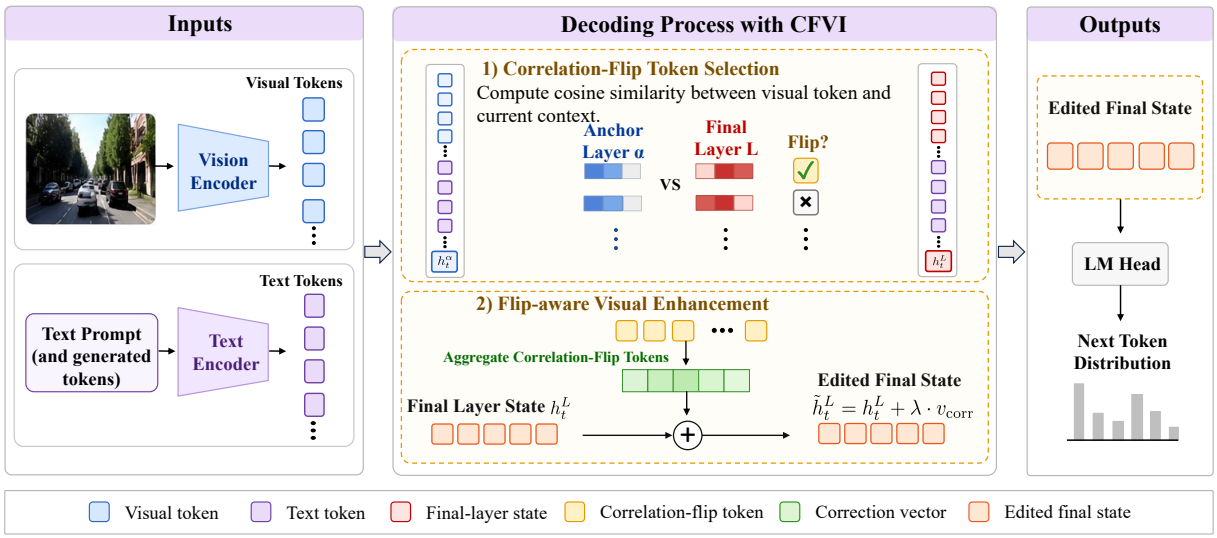

Figure 2: Overview of CFVI. The framework selectively identifies correlation-flip tokens and injects the corresponding visual correction vector into the final-layer hidden state.

where λ controls the correction strength. This additive formulation follows the residual correction paradigm widely adopted in inference-time methods, where hidden states are lightly adjusted to correct biased representations while minimally perturbing the original generation dynamics. Compared with existing visual injection methods, the key distinction of CFVI lies in the introduction of the cross-layer adaptive weight $\alpha_{i,t}$, which dynamically modulates each token's contribution according to its flip magnitude.

The model then predicts the next token with the edited hidden state:

$$P(y_t \mid I, X, y_{<t}) = \text{softmax}(W_o \tilde{\mathbf{h}}_t^L). \quad (9)$$

If no token satisfies the flip condition, CFVI leaves the hidden state unchanged and falls back to standard decoding.

In terms of efficiency, CFVI introduces negligible computational overhead. The additional cost is dominated by computing cosine similarities and applying a softmax over the selected correlation-flip tokens. Importantly, CFVI does not require backpropagation, additional forward passes, or model parameter updates, making it a lightweight inference-time plug-in compatible with LVLMs.

5 Experiments

5.1 Evaluation Setup

LVLMs. To verify the effectiveness of CFVI across model backbones, we evaluate it on representative LVLMs, including LLaVA-1.5-7B, LLaVA-1.5-13B (Liu et al., 2023), MiniGPT-4 (Zhu et al., 2024), and Qwen-VL (Bai et al., 2023). All models use their official configurations. CFVI is training-free and intervenes only at the last layer.

Benchmarks. For general evaluation, MME (Yin et al., 2024) evaluates perception and cognition over 14 subtasks. MM-Vet (Yu et al., 2023) evaluates multidimensional abilities in visual instruction following. For hallucination evaluation, POPE (Li et al., 2023b) formulates hallucination assessment as binary visual question answering. CHAIR (Rohrbach et al., 2018) focuses on the image captioning task. Its core metrics include CHAIR_i, which measures hallucinated objects among all mentioned objects, and CHAIR_s, which measures whether a generated caption contains hallucinated objects:

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

$$\text{CHAIR}_s = \frac{|\text{hallucinated captions}|}{|\text{all captions}|}. \quad (11)$$

Baselines. Contrastive decoding methods include ICD (Wang et al., 2024), which amplifies hallucination-related features with disturbing instructions and subtracts them from the original distribution, and VCD (Leng et al., 2024), which contrasts original and perturbed visual inputs to reduce dependence on language priors. Visual representation enhancement methods include DeCo (Wang et al., 2025), which incorporates shallow-layer visual knowledge into the final layer to correct out-

Method	Existence	Count	Position	Color	Total
LLaVA-1.5	181.67	118.33	104.44	152.78	557.22
VCD (Leng et al., 2024)	172.14	117.14	103.33	119.52	512.14
Nullu (Yang et al., 2025)	190.00	121.11	105.56	156.67	573.34
DeCo (Wang et al., 2025)	175.00	128.33	98.33	125.00	526.66
HulluEdit (Lin et al., 2026)	**195.00**	105.00	126.67	160.00	586.67
CFVI (Ours)	190.00	**155.00**	**128.00**	**170.00**	**643.00**

Table 2: MME benchmark results on LLaVA-1.5. CFVI achieves the highest total score, with clear improvements on perception tasks while maintaining stable performance on cognition-related abilities.

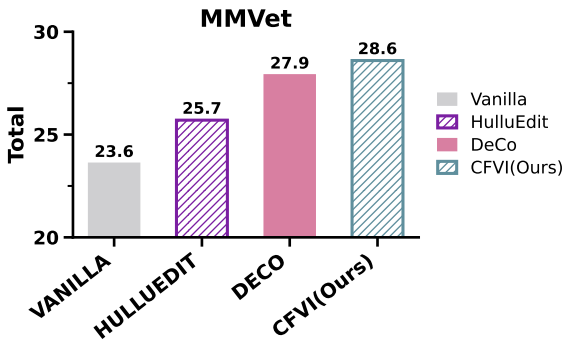

Figure 3: MM-Vet benchmark results on LLaVA-1.5. CFVI achieves the highest overall score of 28.6, outperforming Vanilla, HulluEdit, and DeCo.

puts; VGA (Zhao et al., 2025), which uses visual confidence to guide attention toward relevant regions; and HulluEdit (Lin et al., 2026), which edits hallucination priors through orthogonal subspace decomposition without disrupting visual representations. All baselines use the hyperparameters recommended by their authors and are evaluated with the same prompts and decoding configurations.

Implementation Details. For CFVI, we set the injection weight λ to 1.5, and set τ_{pos} and τ_{neg} to 0.0. For 7B models, we select layer 22 as the anchor layer and apply hidden-state editing to the last Transformer layer. All experiments are conducted on a single NVIDIA RTX 4090 GPU.

5.2 Benchmark Results

MME Results. MME results show that CFVI not only mitigates object hallucination but also improves general multimodal ability. As shown in Table 2, the gains on perception tasks, such as color, count, and position, are especially significant. These results confirm that CFVI is a precise and training-free hallucination mitigation method: it performs targeted correction only when visual evidence is missing or flipped, without harming language reasoning ability.

MM-Vet Results. As shown in Figure 3, CFVI achieves the highest overall score of 28.6. While mitigating hallucinations, CFVI also brings further improvements in more complex visual question answering scenarios. Unlike methods based on decoding adjustment or coarse-grained visual enhancement, CFVI selectively compensates for visual tokens that undergo cross-layer correlation flip, thereby restoring visual evidence suppressed by language priors more precisely.

CHAIR Results. As shown in Table 3, CFVI achieves the best overall performance across the LLaVA-1.5, MiniGPT-4, and Qwen-VL LVLMs compared with other advanced hallucination mitigation methods. CFVI further reduces object hallucination on key metrics, indicating strong generalizability and stability across different model architectures. This improvement is consistent with the ability of CFVI to model hallucination-related correlation-flip signals more precisely. CFVI explicitly identifies visual tokens whose correlations flip between the anchor layer and the final layer. By accurately restoring this missing visual information at the final generation stage, the model relies more on real visual content when generating image captions. The stability of this method across different backbones indicates that cross-layer correlation flipping is a general phenomenon rather than an artifact of specific model design.

Backbone	Method	$\text{CHAIR}_i \downarrow$	$\text{CHAIR}_s \downarrow$
LLaVA-1.5	base	5.2	9.8
	ICD (Wang et al., 2024)	5.1	10.0
	VAF (Yin et al., 2025)	5.1	9.8
	DeCo (Wang et al., 2025)	4.5	8.2
	VGA (Zhao et al., 2025)	4.8	9.4
	HulluEdit (Lin et al., 2026)	4.8	9.4
	CFVI (Ours)	**2.8**	**5.8**
MiniGPT-4	base	7.8	14.4
	HulluEdit	7.9	13.4
	CFVI (Ours)	**5.8**	**11.6**
Qwen-VL	base	6.6	15.0
	HulluEdit	4.1	9.6
	CFVI (Ours)	**4.0**	**9.0**

Table 3: CHAIR benchmark results across several LVLM backbones. CFVI consistently achieves the best performance in reducing object hallucinations, demonstrating strong generalizability across different model architectures.

POPE Results. As shown in Table 4, CFVI achieves the best overall performance on two different backbone sizes of LLaVA-1.5, showing good model scalability. These results validate the effectiveness of identifying and correcting correlation-flip visual tokens through cross-layer comparison.

Efficiency Analysis. In the efficiency evaluation,

Method	LLaVA-1.5-7B		LLaVA-1.5-13B	
	Accuracy ↑	F1 ↑	Accuracy ↑	F1 ↑
Greedy	82.63	83.37	82.70	83.67
VCD (Leng et al., 2024)	83.20	83.80	83.60	84.20
ICD (Wang et al., 2024)	82.90	83.47	83.37	84.00
VAF (Yin et al., 2025)	84.73	85.07	85.33	86.00
DeCo (Wang et al., 2025)	84.00	85.57	77.73	81.50
HulluEdit (Lin et al., 2026)	86.41	87.00	86.29	87.11
VGA (Zhao et al., 2025)	83.27	84.67	83.25	84.95
CFVI (Ours)	**86.67**	**87.21**	**86.53**	**87.36**

Table 4: POPE benchmark results on LLaVA-1.5-7B and LLaVA-1.5-13B. CFVI achieves strong results across both backbones.

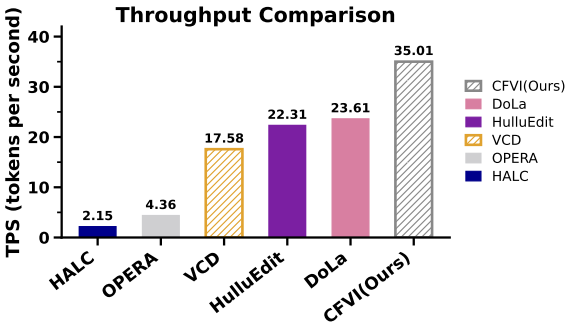

Figure 4: Generation throughput comparison measured under the same LLaVA-1.5-7B setting for all methods. CFVI achieves the highest throughput at 35.01 tokens per second, showing superior efficiency with minimal inference overhead.

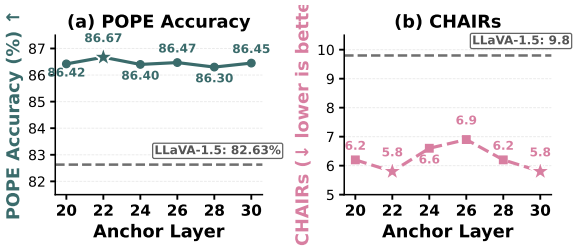

Figure 5: Ablation study of the anchor layer. We select layer 22 as the default anchor layer.

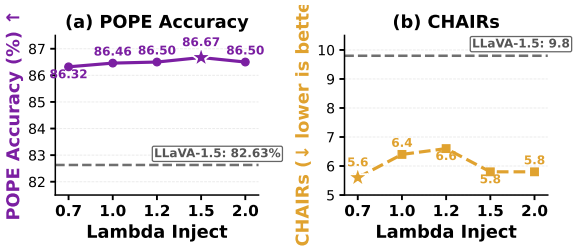

Figure 6: Ablation study of the injection weight λ. We select $\lambda = 1.5$ as the default injection weight.

CFVI achieves the highest generation throughput, reaching 35.01 tokens per second. This is substantially higher than DoLa at 23.61 tokens per second, HulluEdit at 22.31 tokens per second, and VCD at 17.58 tokens per second, and also much higher than OPERA and HALC. As shown in Figure 4, CFVI improves visual faithfulness without introducing additional inference overhead. This efficiency advantage stems from its lightweight design: reusing hidden states from the standard forward pass to compute cross-layer correlations without requiring additional encoders or resampling, and performing simple weighted injection on correlation-flip tokens. This makes CFVI a low-cost, easily deployable solution for hallucination mitigation.

5.3 Ablation Study

We conduct systematic ablation experiments on the CHAIR dataset and POPE dataset. All experiments are conducted with LLaVA-1.5-7B.

Core Parameter Analysis. The choice of anchor layer is crucial for identifying correlation-flip tokens. An ideal anchor layer should preserve strong visual grounding information while already integrating the prompt context. We conduct experiments over layers 20 to 30, and the results are shown in Figure 5. We find that layer 22 achieves the best results on both datasets, indicating that overly shallow layers may not sufficiently integrate semantic information, whereas overly deep layers may already be overly affected by language priors. Based on this observation, we choose layer 22 as the default anchor layer.

The injection weight λ controls the strength of the visual correction vector. We test λ from 0.7 to 2.0, and the results are shown in Figure 6. The results show that $\lambda = 1.5$ achieves a low CHAIR_s score of 5.8 and the highest accuracy. A correction that is too strong may introduce noise, while a correction that is too weak may not sufficiently mitigate hallucination. To prioritize hallucination mitigation, we choose $\lambda = 1.5$ as default.

Overall, our method is not sensitive to hyperparameters. Different hyperparameter choices have only a limited impact on the results, further validating the stability of CFVI.

Core Component Analysis. We further conduct component ablations, as shown in Table 5. The results lead to two main observations. First, replacing correlation-flip tokens with an equal number of random tokens, high-attention tokens, high-activation tokens, or high-similarity tokens consistently degrades performance. This indicates that conventional single-layer saliency signals, including attention, activation magnitude, and similarity, are insufficient for locating hallucination-related

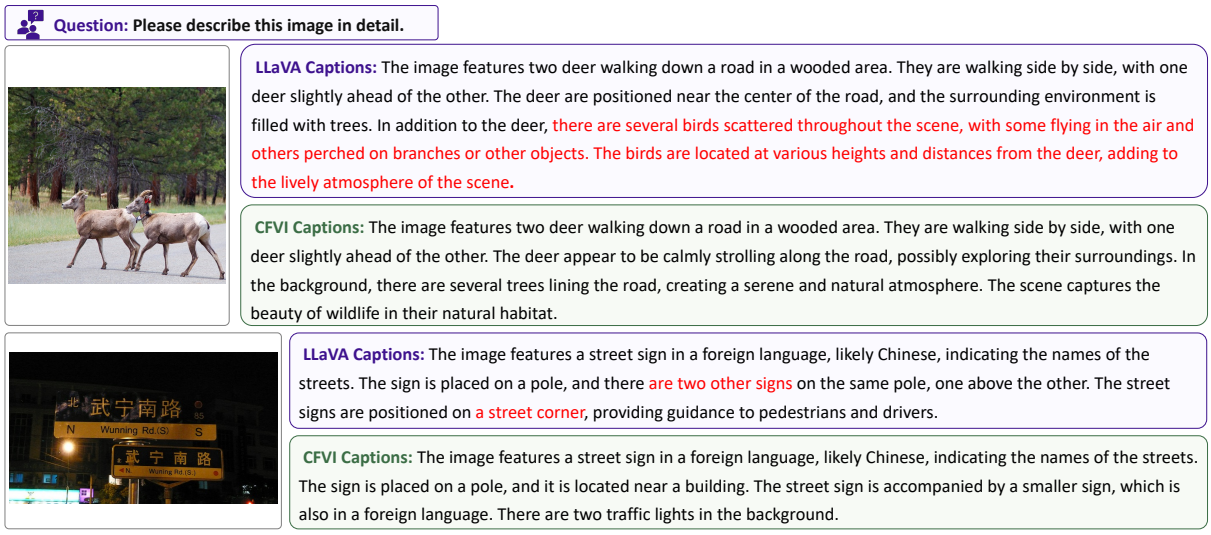

Figure 7: Qualitative comparison of object hallucination mitigation. Example images are paired with captions generated by the baseline model and CFVI.

Method	$\text{CHAIR}_s \downarrow$	$\text{CHAIR}_i \downarrow$	POPE Acc ↑
Full model	**5.8**	**2.8**	**86.67**
High-attention tokens	7.0	3.2	86.09
High-activation tokens	6.2	3.1	86.20
High-similarity tokens	6.6	3.2	86.21
Random tokens	6.7	3.2	86.17
All visual tokens	6.4	3.2	86.26
Without weighting	6.0	2.9	86.51

Table 5: Component ablation results of CFVI. Replacing correlation-flip tokens with conventional token-selection strategies consistently degrades performance, validating the effectiveness of cross-layer correlation-flip modeling and adaptive weighting.

visual evidence. In contrast, cross-layer correlation flips more accurately identify visual tokens that are suppressed by deep language priors. Injecting all visual tokens also causes clear performance degradation, suggesting that non-selective enhancement introduces noise and dilutes useful visual signals. Second, weighted aggregation provides consistent gains. Removing the weighting mechanism degrades performance, which shows that weighting selected flip tokens according to correlation strength helps construct a more accurate correction direction and further reduces hallucinations while preserving visual recall.

Overall, the full model achieves the best performance, demonstrating that CFVI provides a better balance between hallucination suppression and visual-information preservation. The random-control experiment further supports that the gains are tied to meaningful flip-token discovery rather than random perturbation.

6 Case Study

We present qualitative examples to demonstrate the effectiveness of CFVI. As shown in Figure 7, the baseline model, such as LLaVA-1.5, incorrectly indicates that birds are present in the first image. In the second image, it incorrectly describes both the number and location of street signs, including a location that is not a street corner. In contrast, CFVI suppresses object hallucination more precisely while preserving the ability of the model to answer general questions about scene description, counting, and spatial recognition. These qualitative examples further demonstrate the effectiveness of our method in mitigating object hallucination.

7 Conclusion

This work proposes CFVI, a training-free method for reducing hallucinations in large vision-language models. CFVI identifies visual tokens that change from supportive evidence in an anchor layer to suppressed evidence in the final layer. It then uses these correlation-flip tokens to construct a compact visual correction vector and enhance the final hidden state. This targeted correction allows CFVI to suppress hallucinations without weakening visual grounding. Experiments on multiple benchmarks and model backbones show that CFVI achieves strong performance while maintaining high inference efficiency. These results suggest that correlation flip is a useful internal signal for understanding and mitigating hallucinations in LVLMs.

Limitations

CFVI has several limitations. First, it relies on detectable correlation-flip signals. If hallucination originates from early visual encoding failure or complex reasoning failures, the model may not exhibit a clear intermediate-positive and final-negative pattern, which limits the effect of CFVI. In addition, CFVI requires choices of hyperparameters such as anchor layer and injection strength. The optimal settings may vary across models and tasks. Future work will explore adaptive flip detection and dynamic injection strategies, and will combine CFVI with training-time alignment or external verification to improve robustness in adversarial scenarios and complex reasoning tasks.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2):3.

Xinrong Chen, Xu Chu, Yingmin Qiu, Hengyuan Zhang, Jing Xiong, Shiyu Tang, Shuai Liu, Shaokang Yang, Cheng Yang, Hayden Kwok-Hay So, and 1 others. 2026. Residual decoding: Mitigating hallucinations in large vision-language models via history-aware residual guidance. *arXiv preprint arXiv:2602.01047*.

Hao Fang, Changle Zhou, Jiawei Kong, Kuofeng Gao, Bin Chen, Tao Liang, Guojun Ma, and Shu-Tao Xia. 2026. Grounding language with vision: A conditional mutual information calibrated decoding strategy for reducing hallucinations in lvlms. *Advances in Neural Information Processing Systems*, 38:105233–105258.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. 2024. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14303–14312.

Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. 2024. Self-introspective decoding: Alleviating hallucinations for large vision-language models. *arXiv preprint arXiv:2408.02032*.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, and 1 others. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. 2025. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25004–25014.

Pegah Khayatan, Jayneel Parekh, Arnaud Dapogny, Mustafa Shukor, Alasdair Newson, and Matthieu Cord. 2026. When prompts override vision: Prompt-induced hallucinations in lvlms. *arXiv preprint arXiv:2604.21911*.

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

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Qiming Li, Zekai Ye, Xiaocheng Feng, Weihong Zhong, Weitao Ma, and Xiachong Feng. 2026. Causal tracing of object representations in large vision language models: Mechanistic interpretability and hallucination mitigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 31645–31653.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 292–305.

Yangguang Lin, Quan Fang, Yufei Li, Jiachen Sun, Junyu Gao, and Jitao Sang. 2026. Hulluedit: Single-pass evidence-consistent subspace editing for mitigating hallucinations in large vision-language models. *arXiv preprint arXiv:2602.22727*.

Aixin Liu, Bei Feng, Bin Wang, Bingxuan Wang, Bo Liu, Chenggang Zhao, Chengqi Dengr, Chong Ruan, Damai Dai, Daya Guo, and 1 others. 2024a. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model. *arXiv preprint arXiv:2405.04434*.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024b. A survey on hallucination

in large vision-language models. *arXiv preprint arXiv:2402.00253*.

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

Kyungmin Min, Minbeom Kim, Kang-il Lee, Dongryeol Lee, and Kyomin Jung. 2025. Mitigating hallucinations in large vision-language models via summary-guided decoding. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4183–4198.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045.

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. 2025. Mllm can see? dynamic correction decoding for hallucination mitigation. In *International Conference on Learning Representations*, volume 2025, pages 13712–13736.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15840–15853.

Le Yang, Ziwei Zheng, Boxu Chen, Zhengyu Zhao, Chenhao Lin, and Chao Shen. 2025. Nullu: Mitigating object hallucinations in large vision-language models via halluspace projection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14635–14645.

Hao Yin, Guangzong Si, and Zilei Wang. 2025. Clearsight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14625–14634.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Xiaofeng Zhang, Yuanchao Zhu, Chaochen Gu, Xiaosong Yuan, Qiyan Zhao, Jiawei Cao, Feilong Tang, Sinan Fan, Yaomin Shen, Chen Shen, and 1 others. 2026. Hallucination begins where saliency drops. *arXiv preprint arXiv:2601.20279*.

Jianfei Zhao, Feng Zhang, Xin Sun, Chong Feng, and Zhixing Tan. 2025. Tell model where to look: Mitigating hallucinations in mllms by vision-guided attention. *arXiv preprint arXiv:2511.20032*.

Deyao Zhu, Xiaoqian Shen, Xiang Li, Mohamed Elhoseiny, and 1 others. 2024. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations*, volume 2024, pages 18378–18394.

Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. 2025. Ibd: Alleviating hallucinations in large vision-language models via image-biased decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1624–1633.

Xingyu Zhu, Kesen Zhao, Liang Yi, Shuo Wang, Zhicai Wang, Beier Zhu, and Hanwang Zhang. 2026. Look carefully: Adaptive visual reinforcements in multimodal large language models for hallucination mitigation. *arXiv preprint arXiv:2602.24041*.

Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Kening Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, and 1 others. 2024. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. *arXiv preprint arXiv:2410.03577*.

A Hyperparameter Specifications

This appendix reports the full hyperparameter configuration for CFVI on LLaVA-1.5-7B, as shown in Table 6. All parameter values are determined through systematic empirical validation to optimize the trade-off between hallucination reduction and generation quality. The anchor layer is set to layer 22, which retains strong visual grounding information while not fully collapsing into language-prior-dominated decisions. This provides a reliable reference for identifying correlation-flip tokens. Both positive and negative thresholds are set to zero to maximize recall of flip tokens, while the weighting mechanism continuously adjusts their contributions and avoids information loss from hard thresholding. The injection weight $\lambda = 1.5$ achieves the lowest CHAIR_s in ablations, validating its balance between correcting hallucination and maintaining generation quality. The weighting temperature is set to 1 so that token weights reflect the degree of suppression. The generation temperature is set to 0.05 to encourage deterministic generation and ensure fair evaluation. All generation settings follow the established evaluation protocols.

Parameter	Value
Anchor layer	22
Positive threshold τ_{pos}	0.0
Negative threshold τ_{neg}	0.0
Injection weight λ	1.5
Softmax temperature T_w	1.0
Stability ϵ	1.0×10^{-6}
Top p	1.0
Generation temperature	0.05

Table 6: Hyperparameter configuration for CFVI on LLaVA-1.5-7B.

B Hyperparameter Sensitivity Analysis

We analyze the sensitivity of key hyperparameters. CFVI includes architecture-related parameters such as the anchor layer ℓ_a, the injection weight λ, the softmax temperature for weighted aggregation T_w, and the positive and negative thresholds τ_{pos} and τ_{neg}. The anchor layer is selected from middle and deep layers according to model architecture to ensure stable visual-semantic alignment. The injection weight controls the degree of visual compensation and must balance precise correction against the risk of adding noise. As shown in Table 7, the remaining hyperparameters exhibit low sensitivity within practical ranges.

Parameter range	$\text{CHAIR}_s \downarrow$	Recall $\uparrow$
τ_{pos} (0–0.2)	5.8 ± 0.4	48.62 ± 0.6
τ_{neg} (0–0.2)	6.2 ± 0.2	48.06 ± 0.8
T_w (0.8–1.2)	5.6 ± 0.2	48.18 ± 0.5

Table 7: Hyperparameter sensitivity analysis on LLaVA-1.5-7B.

C Correlation-Flip Tokens and Hallucination

To further verify the relationship between correlation-flip tokens and hallucination, we compare CFVI with two simplified variants that use only one-sided correlation signals. The all-neg variant selects visual tokens that are negatively correlated in the final layer, while the all-pos variant selects visual tokens that are positively correlated in the anchor layer. As shown in Table 8, both simplified variants underperform the full model, especially on CHAIR_i. These results indicate that correlation-flip tokens are more closely associated with hallucination than tokens selected by a single correlation sign. They also validate the effectiveness of the flip-token selection module, which identifies visual evidence that is perceived in intermediate layers but suppressed before prediction.

Method	$\text{CHAIR}_s \downarrow$	$\text{CHAIR}_i \downarrow$
Full model	**5.8**	**2.8**
All negative tokens	6.4	3.2
All positive tokens	**5.8**	2.9

Table 8: Ablation study on the relationship between correlation-flip tokens and hallucination. The full flip-token selection module outperforms variants that use only one-sided correlation signals.

D Analysis of Flip Score Formulation

In CFVI, the flip score is defined as

$$q_{i,t} = \text{ReLU}(C^a_{i,t}) \cdot \text{ReLU}(-C^L_{i,t}), \qquad (12)$$

which measures whether a visual token is simultaneously positively correlated in the anchor layer and negatively correlated in the final layer. We further compare different formulations of the flip score, including subtraction and ratio forms, as shown in Table 9.

The multiplicative formulation achieves the best performance for several reasons. First, it is naturally consistent with the AND structure of the flip condition: a token receives a high score only when

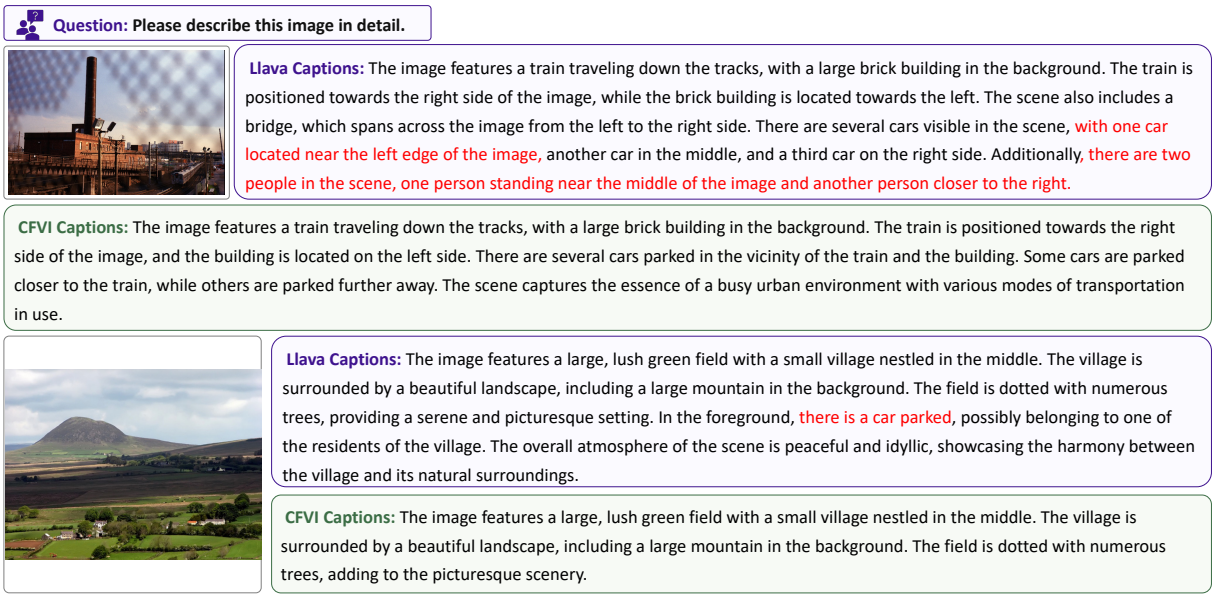

Figure 8: Additional qualitative comparisons of object hallucination mitigation.

Flip score formulation	$\text{CHAIR}_s \downarrow$	$\text{CHAIR}_i \downarrow$
Multiplication	**5.8**	**2.8**
Subtraction	6.2	3.0
Ratio	6.3	3.1

Table 9: Ablation study on different flip score formulations.

both positive anchor-layer correlation and negative final-layer correlation are simultaneously strong. This avoids assigning high scores to pseudo-flip tokens that are only weakly correlated, only decreasing across layers, or only negative in the final layer. Second, the multiplicative form automatically prioritizes visual tokens with strong evidence on both sides of the flip, which better matches the intended semantics of "flip intensity". Third, the output is always non-negative, making it naturally compatible with the subsequent softmax weighting process. In contrast, subtraction or ratio formulations are less stable and may amplify noisy or numerically unstable correlations, leading to inferior hallucination mitigation performance.

E Qualitative Case Studies

Figure 8 provides additional qualitative examples showing that CFVI restores missing visual information at the final generation stage without damaging valid visual evidence, thereby mitigating object hallucinations.