Inject to Heal: Alleviating hallucination in LVLMs via Context Embedding Injection

Anonymous ACL submission

Abstract

Hallucinations—generating responses inconsistent with the visual input—remain a critical limitation of large vision-language models (LVLMs), especially in open-ended tasks such as image captioning and visual reasoning. In this work, we probe the layer-wise generation dynamics that drive hallucinations and propose a training-free mitigation strategy. Employing the Logit Lens, we examine how LVLMs construct next-token distributions across decoder layers, uncovering a pronounced *commitment-depth gap*: truthful tokens accumulate probability mass on their final candidates earlier than hallucinatory ones. Drawing on this discovery, we introduce Context Embedding Injection (CEI), a lightweight method that harnesses the hidden state of the last input token—the *context embedding*—as a grounding signal to maintain visual fidelity throughout decoding and curb hallucinations. Evaluated on the CHAIR, AMBER, and MMHal-Bench benchmarks (with a maximum token length of 512), CEI outperforms state-of-the-art baselines across three LVLMs, with its dynamic variant yielding the lowest overall hallucination rates. By integrating novel mechanistic insights with a scalable intervention, this work advances the mitigation of hallucinations in LVLMs. Data and code are available at `https://anonymous.4open.science/r/CEI-HallucinationMitigation-2026/`.

1 Introduction

Large vision–language models (LVLMs) (Bai et al., 2023; Chen et al., 2023b; Liu et al., 2023; Chen et al., 2024; Dai et al., 2023; Ye et al., 2024) have rapidly advanced general-purpose multimodal understanding by pairing strong vision encoders (e.g., CLIP (Radford et al., 2021)) with large language models (e.g., LLaMA (Touvron et al., 2023)). They excel at image captioning, visual question answering, and medical report generation (Chen et al., 2023a; Wu et al., 2023; Hartsock and Rasool, 2024). Recent systems further scale in capability and robustness across diverse image resolutions and inputs, expanding their application (Wang et al., 2024b; Liu et al., 2024b). In spite of these strengths, LVLMs remain susceptible to *hallucination*—producing text that deviates from the image—undermining reliability in applications that demand factuality and faithfulness, such as autonomous driving and medical report generation (Keskar et al.; Hartsock and Rasool, 2024).

Extensive research has sought to analyze hallucinations in LVLMs. Prior work attributes them to interacting factors, with analyses predominantly focusing on attention mechanisms. Studies highlight biased cross-attention to a few image patches (Woo et al., 2024; Zhu et al., 2025), attention drift during long generations (Fazli et al., 2025; Zhou et al., 2024), and attention aggregation patterns such as "anchor/summary" tokens (Huang et al.; Zhang et al., 2024) as key contributors. Despite these findings, *the layer-wise progression of token prediction—how models accrue probability mass on their final decision set—remains unexplored.*

Meanwhile, various mitigation strategies have been proposed, including data optimization (e.g., negative samples), architectural enhancements (e.g., improved alignment modules), and decoding optimizations (e.g., contrastive schemes) (Huang et al., 2025). Inference-time interventions, such as contrastive decoding (Leng et al., 2023; Favero et al., 2024; Zhu et al., 2024) and attention calibration (Fazli et al., 2025; Liu et al., 2024c), have gained traction for their low cost and generalizability. However, *these methods often achieve inconsistent performance in open-ended generative scenarios, such as image captioning with performance degrading as sequence length increases due to accumulated errors and drift from visual grounding (Fazli et al., 2025).*

To address these gaps, instead of characterizing hallucinations via post-hoc cues such as raw atten-

tion maps or dataset-level correlations, we apply Logit Lens (nostalgebraist, 2020)—a mechanistic interpretability technique—to directly track how token preferences form across decoder layers during decoding. This analysis unveils a profound commitment-depth gap: truthful tokens stabilize probability mass on their final candidates earlier across decoder layers than hallucinatory ones, offering mechanistic insights into the roots of hallucination in LVLMs. Building on this discovery, we introduce Context Embedding Injection (CEI), a novel model-agnostic, training-free method that continually aligns the hidden states of generated tokens with a visual grounding signal—the context embedding—extracted from the initial input processing, thereby sustaining fidelity in open-ended generation. Our contributions are threefold:

- An empirical analysis, via the Logit Lens, demonstrating that truthful tokens consolidate predictions earlier across decoder layers than hallucinatory ones.
- A novel mitigation method, Context Embedding Injection (CEI), for sustained visual alignment in open-ended generation.
- Consistent improvements over baselines on generative hallucination benchmarks such as CHAIR, AMBER, and MMHal-Bench (GPT4-evaluated).

2 Related Work

2.1 Large Vision-Language Models

Large vision-language models (LVLMs) extend LLMs with visual inputs via a vision encoder (e.g., CLIP (Radford et al., 2021), ViT (Dosovitskiy et al., 2021)), an alignment module (e.g., linear projection (Liu et al., 2023, 2024b) or Q-former (Dai et al., 2023; Zhu et al., 2023)), and an autoregressive LLM backbone (e.g., LLaMA (Touvron et al., 2023), Vicuna (Zheng et al., 2023)). Recent families (Liu et al., 2024b; Wang et al., 2024b; Ye et al., 2024) scale data/model size and improve visual tokenization and positional fusion (e.g., multi-scale inputs and M-RoPE (Liu et al., 2024b; Wang et al., 2024b)), yielding broad gains across captioning and VQA. Despite these advances, LVLMs remain prone to hallucination, limiting reliability in safety-critical settings (Bai et al., 2025).

2.2 Hallucination Mitigation in LVLMs

LVLM hallucinations are outputs that deviate from the image (e.g., fabricated objects/attributes/relations), commonly linked to linguistic priors, dataset bias, and modality misalignment (Bai et al., 2025; Li et al., 2023b; Liu et al., 2024a). Existing mitigation methods span fine-tuning (Gunjal et al., 2024; Jiang et al., 2024; Kim et al., 2023), post-hoc correction (Yin et al., 2023; Zhou et al., 2024), and decoding-time approaches (Leng et al., 2023; Huang et al., 2024; Suo et al., 2025; Fazli et al., 2025; An et al., 2025; Yang et al., 2025). Decoding-time methods are especially attractive since they require no retraining; prominent examples include contrastive decoding variants that compare original vs. perturbed inputs to promote visual grounding (e.g., VCD (Leng et al., 2023), M3ID (Favero et al., 2024), IBD (Zhu et al., 2024)) as well as attention-centric calibration of cross-modal interactions (e.g., AGLA (An et al., 2025), CAAC (Fazli et al., 2025)). In contrast, our method intervenes in the embedding space, continually aligning token representations with a visual grounding signal. For a more comprehensive discussion of related work and additional references, see Appendix D.

3 Preliminary Insights

To effectively mitigate hallucinations in large vision-language models (LVLMs), a nuanced understanding of the underlying generation dynamics is essential. A key question is: *what internal signals and structural patterns distinguish truthful generations from hallucinatory ones?* Motivated by recent advances in mechanistic interpretability, we apply targeted mechanistic probes to decoder layers to address this question, revealing the internal signals responsible for divergence from visual information. In Section 3.1, we identify a reusable grounding signal embedded in the initial decoding step, and in Section 3.2, we uncover a systematic commitment-depth gap between truthful and hallucinatory tokens—insights that directly inform our subsequent intervention strategy.

3.1 Initial Decoding Step as a Grounding Signal

We hypothesize that the hidden state of the last input token—the *context embedding*—encodes query-aligned visual information, steering the logit distribution of the first generated token toward image-grounded content while serving as a reusable an-

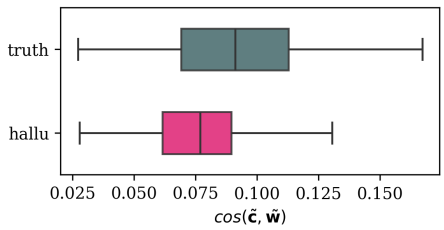


Figure 1: **Context embedding semantic alignment.** Box plot of centered cosine similarities $\cos(\tilde{\mathbf{c}}, \tilde{\mathbf{w}})$ between the centered context embedding $\tilde{\mathbf{c}}$ and centered target token embedding $\tilde{\mathbf{w}}$. Blue and orange boxes denote truthful and hallucinatory tokens respectively.

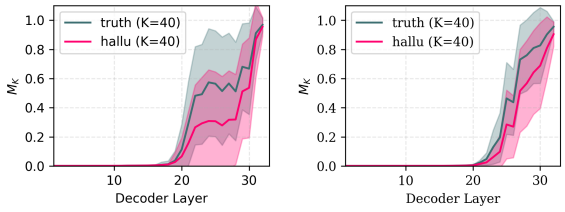


Figure 2: **Layer-wise commitment curves.** Top-K probability mass, $M_K(\ell)$, of truthful (blue) and hallucinatory (red) tokens across decoder layers, with shaded bands indicating one standard deviation: (a) InstructBLIP and (b) LLaVA-1.5.

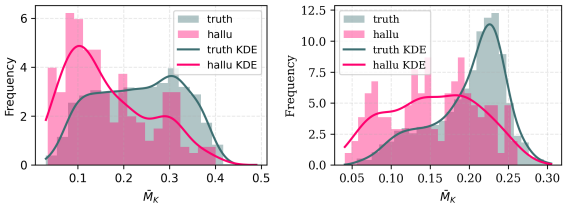


Figure 3: **Average confidence in decision set.** Histogram of mean top-K probability mass, $\bar{M}_K$, of truthful (blue) and hallucinatory (red) tokens for (a) InstructBLIP and (b) LLaVA-1.5 ($K = 40$).

chor to sustain visual fidelity throughout decoding.

To test this, we evaluate the semantic alignment between the context embedding and target tokens using the generative captioning split of the AMBER dataset (Wang et al., 2024a), which provides annotations for truthful and hallucinatory objects. Using InstructBLIP (Dai et al., 2023), we generate captions, extract target tokens, and compute centered cosine similarities $\cos(\tilde{\mathbf{c}}, \tilde{\mathbf{w}})$ between the centered context embedding $\tilde{\mathbf{c}}$ and centered token embeddings $\tilde{\mathbf{w}}$ (following Muennighoff et al., 2023 to account for anisotropy).

As shown in Figure 1, truthful tokens exhibit significantly higher mean cosine similarities than hallucinatory ones (95% CI excluding zero), confirming the context embedding's inclination toward grounded content. This aligns with Zhao et al. (2024), who leverage first-token hidden states for detecting hallucinations via lightweight probes. Together, these findings indicate that *the context embedding is intrinsically more semantically aligned with visually supported tokens than with hallucinated ones.*

3.2 Layer-wise Commitment in Truthful vs. Hallucinatory Tokens

Prior research has shown that language models determine function words within mid-layers, maintaining stable predictions thereafter, while content words involve ongoing adjustments in later layers (Zhu et al., 2024). Motivated by this, we ask if a similar disparity in commitment depth underlies the generation of truthful and hallucinatory tokens. Specifically, we investigate *if truthful tokens accrue probability mass on their final decision set earlier than hallucinatory ones.*

Similar to the experimental pipeline from Section 3.1—which leverages the AMBER generative split (Wang et al., 2024a) for caption generation and target token identification—we perform, for each truthful or hallucinatory token, a forward pass on the prefix up to each target token to extract intermediate hidden states h_ℓ at layer ℓ. We then apply the Logit Lens (nostalgebraist, 2020) to map these states to vocabulary distributions $p_\ell = \sigma(h_\ell E^\top)$, where E is the embedding matrix.

Let S_K be the *final decision set* defined as top-K tokens under the distribution p_L. The layer-wise top-K probability mass is defined as:

$$M_K(\ell) = \sum_{w \in S_K} p_\ell(w), \tag{1}$$

which measures how much probability layer ℓ assigns to the final decision set. To quantify the overall confidence of the model on the final decision set, we compute the layer-wise average top-K mass:

$$\bar{M}_K = \frac{1}{L} \sum_{\ell=1}^{L} M_K(\ell). \tag{2}$$

Across InstructBLIP (Dai et al., 2023), LLaVA (Liu et al., 2023), and LLaVA-NeXT (Liu et al., 2024b), truthful tokens show earlier commitment (higher $M_K(\ell)$ in mid-to-late layers, $\ell \approx 20$–30; see Figure 2), with robust separation for $K \in [20, 80]$ (Appendix B).

Furthermore, the distributions of mean top-K mass $\bar{M}_K$ differ markedly between truthful and hallucinatory tokens (Figure 3): hallucinatory ones exhibit lower values, signaling delayed probability consolidation on final candidates, while truthful tokens accrue confidence earlier across layers, yielding stronger, more stable commitments.

> Truthful tokens commit earlier and accrue higher cumulative confidence on their final decision set, revealing a clear commitment-depth gap between truthful and hallucinatory generations.

4 Proposed Method

We propose CEI, a training-free decoding intervention that reuses the context embedding to maintain outputs visually grounded (Figure 4). CEI extracts an image-conditioned context embedding before generation and re-injects it into the decoder during autoregressive decoding to anchor token predictions to the original visual input. Section 4.1 fixes the inference notation; Section 4.2 introduces *Static CEI*, which injects this fixed context into the last-input hidden state at a chosen layer with a constant weight α. Section 4.3 presents *Dynamic CEI*, where α is modified per token from $\bar{M}_K$ when commitment is weak.

4.1 Inference in Large Vision-Language Models

The inference process in LVLMs comprises two main stages: *input preparation* and *text generation*. These stages facilitate autoregressive text generation conditioned on multimodal inputs.

Input preparation. The processor module transforms the raw inputs into a unified embedding space suitable for the autoregressive language decoder.

A vision encoder E_v extracts features from the input image I, which are then projected into the textual embedding space via an alignment module (e.g., linear projection or Q-former) to yield visual tokens $V = [v_1, \ldots, v_{N_v}] \in \mathbb{R}^{N_v \times d}$, where d is the embedding dimension. The text tokenizer E_t embeds the input sequence $q = [P; y_{<t}] = [q_1, \ldots, q_{N_q}]$, where P is the prompt and $y_{<t}$ are previously generated tokens, producing $Q = E_t(q) \in \mathbb{R}^{N_q \times d}$.

The multimodal input is then formed by concatenation:

$$X = [V; Q] \in \mathbb{R}^{(N_v+N_q)\times d}. \tag{3}$$

Text generation. The decoder, with L layers, processes X autoregressively. At step t, hidden states evolve as $H^{(\ell)} = f^{(\ell)}(H^{(\ell-1)})$ where $\ell = 1, \ldots, L$, $f^{(\ell)}$ being the ℓ-th transformer layer, and $H^{(0)} = X$. The final hidden state at the last position, $h_t^{(L)} \in \mathbb{R}^d$, aggregates contextual information from the entire sequence for predicting the next token. This state is projected back to the vocabulary space via the unembedding matrix $E^\top \in \mathbb{R}^{d\times|\mathcal{V}|}$: to yield logits:

$$z_t = h_t^{(L)} E^\top, \tag{4}$$

followed by a softmax results in a probability distribution over the vocabulary $\mathcal{V}$:

$$p_\theta(y_t \mid x, p, y_{<t}) = \mathrm{softmax}(z_t). \tag{5}$$

A decoding strategy (e.g., greedy selection, nucleus sampling, or beam search) selects the next token $\hat{y}_t$, which is appended to the sequence for the subsequent iteration. This autoregressive process continues until a termination criterion is met.

4.2 Static Context Embedding Injection

Building on the insights from Section 3.1, we propose Context Embedding Injection (CEI), a training-free method to enhance visual grounding in LVLMs during autoregressive generation. We define *context embedding* as the hidden state at the last prompt token in the final decoder layer, which directly yields the logits for the first generated token. This representation preserves the full information of the initial logit distribution while being amenable to blending with subsequent hidden states. CEI leverages context embedding in subsequent decoding steps to reinforce visual alignment throughout the generation process.

Let the prompt be tokenized into N_p tokens. The multimodal input for the initial forward pass (before generation) is then $X_0 = [V; P] \in \mathbb{R}^{(N_v+N_p)\times d}$. Prior to generation, we perform an initial forward pass on the multimodal input X to extract the context embedding $\mathbf{c}$ from the last decoder layer at the last input position Q:

$$\mathbf{c} = h_{N_v+N_p}^{(L)}. \tag{6}$$

During decoding, we inject this fixed signal at a chosen layer $\ell_{\text{inj}} \in 1, \ldots, L$ using a constant mixing weight $\alpha \in [0, 1]$. For each step t,

$$\tilde{h}_t^{(\ell_{\text{inj}})} = (1-\alpha)\, h_t^{(\ell_{\text{inj}})} + \alpha\, \mathbf{c}. \tag{7}$$

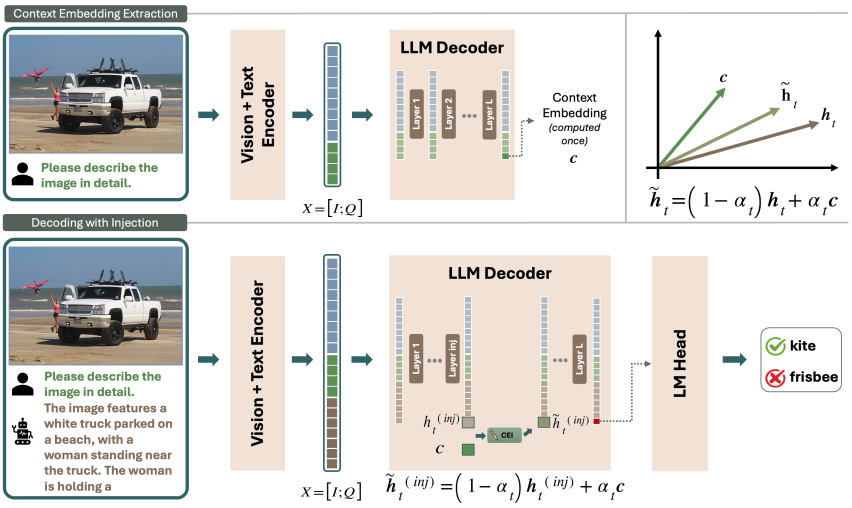


Figure 4: **CEI overview.** An initial forward pass over the image–prompt input extracts a fixed *context embedding* $\mathbf{c}$ from the final decoder layer at the last prompt position (top). During autoregressive decoding, this pre-computed signal is injected at a chosen decoder layer via a weighted average mechanism, guiding generation toward the original image-conditioned context and improving visual grounding (bottom). Blue, green, and brown squares denote the embeddings of image tokens, prompt tokens, and previously generated tokens, respectively.

The blended hidden state $\tilde{h}_t^{(\ell_{\text{inj}})}$ is thereby aligned more closely with the context embedding. This alignment propagates through subsequent layers, influencing the next-token distribution and promoting fidelity to the input image. Static CEI computes $\mathbf{c}$ once (before decoding) and applies uniform injection across all tokens, incurring only a lightweight vector blending operation without requiring additional training. The hyperparameters α and ℓ_{inj} are tuned for each LVLM.

While Static CEI effectively reinforces visual grounding, it applies a uniform injection across all decoding steps. In practice, however, not all tokens are equally image-dependent: function tokens primarily support linguistic fluency, whereas content tokens (e.g., objects, attributes, actions) are more tightly coupled to the visual input. This mismatch motivates a dynamic formulation that adapts the injection strength at the token level.

4.3 Dynamic Context Embedding Injection

The analysis in Section 3.2 reveals that hallucinatory tokens exhibit lower mean top-K probability mass $\bar{M}_K$ compared to truthful ones, indicating delayed commitment and a higher risk of deviation from visual grounding. We leverage this signal to extend static CEI by dynamically modulating the injection weight α_t per generation step t.

At each step t, we first compute $\bar{M}_K$ via a probe forward pass as explained in Section 3.2. We then map it to α_t using a half-cosine schedule:

$$\alpha_t = \min(\alpha_{max} \cos(\frac{\pi}{2} . \frac{\bar{M}_K}{\beta}), 0). \quad (8)$$

This schedule ensures low $\bar{M}_K$ values that signal high hallucination risk, enforce stronger injection (higher α) for better alignment. While risk decreases rapidly as $\bar{M}_K$ rises; β also indicates the cutoff where injection tapers off. The blended hidden state is then formed as in static CEI, but with the adaptive α_t:

$$\tilde{h}_t^{(\ell_{\text{inj}})} = (1 - \alpha_t)\, h_t^{(\ell_{\text{inj}})} + \alpha_t\, \mathbf{c}. \quad (9)$$

A second forward pass is subsequently performed with the adjusted injection weight to predict the next token. Dynamic CEI thus adapts the injection online based on generation dynamics, enhancing robustness without requiring additional training.

5 Experiments and Results

5.1 Experimental Setup

Models. Following the experimental protocols used in recent works on LVLM hallucination, we evaluate the proposed CEI method on three widely used 7B-parameter LVLMs: InstructBLIP (Dai et al., 2023), LLaVA-1.5 (Liu et al., 2023), and LLaVA-NeXT (Liu et al., 2024b). The CEI framework is designed to be model-agnostic, enabling integration with diverse LVLMs. Detailed experimental configurations and implementation specifics are provided in Appendix A.

Benchmarks. We focus on generative benchmarks that assess faithfulness in open-ended generation. Specifically, we adopt the CHAIR (Rohrbach et al., 2018), AMBER (Wang et al., 2024a), and MMHal-Bench (Sun et al., 2023) benchmarks. We set the maximum number of new tokens to 512 for all benchmarks to allow uninterrupted, free-form generations, as hallucinations often occur in the later stages of decoding.

Metrics. For CHAIR, we report CHAIR_i and CHAIR_s, which quantify object hallucinations at the instance and sentence levels, respectively. For AMBER, we report CHAIR and HAL, which measure hallucination frequency and response-level occurrence, along with COVER, which evaluates completeness and informativeness. Lower CHAIR/HAL and higher COVER indicate better grounding. For MMHal-Bench, we report Score and HalRate which measure the quality of the responses and hallucination rates.

Baselines. We compare CEI against five training-free hallucination mitigation techniques. Contrastive decoding methods include VCD (Leng et al., 2023), AvisC (Woo et al., 2024), and M3ID (Favero et al., 2024), which leverage contrastive strategies, alongside OPERA (Huang et al.), a beam-search variant penalizing overconfident tokens for visual grounding, and CAAC (Fazli et al., 2025), an attention-calibration approach.

Implementation Details. Baseline hyperparameters follow settings from their original publications for consistency. For CEI, the critical hyperparameters of dynamic CEI, $\alpha_{\max}$ and β, govern the mapping from $\bar{M}_K$ to α, and are tuned per LVLM, yielding $\alpha_{\max} = 0.4, 0.25, 0.17$ and $\beta = 0.7, 0.55, 0.35$ for InstructBLIP, LLaVA-1.5, and LLaVA-NeXT, respectively. The injection layer is fixed at 10, as experimental analysis indicated that layers 10–15 yield the most substantial improvements in faithfulness. More implementation details in Appendix A. Data and code can be found at `https://anonymous.4open.science/r/CEI-HallucinationMitigation-2026/`.

5.2 Evaluation Results

CHAIR. The Caption Hallucination Assessment with Image Relevance (CHAIR) benchmark (Rohrbach et al., 2018) quantifies object hallucinations in image captioning by comparing objects mentioned in generated captions against ground-truth annotations from the MSCOCO 2014 dataset (Lin et al., 2015). We report two standard metrics: CHAIR_i, the proportion of hallucinated objects among all mentioned objects, and CHAIR_s, the fraction of captions containing at least one hallucinated object, with lower values indicating superior performance. Following established protocols (Huang et al.), we evaluate on 500 randomly selected images from the MSCOCO validation set, prompting models with "Please describe this image in detail." To capture hallucinations that often emerge in extended sequences, we set the maximum new tokens to 512, allowing uninterrupted generation.

As summarized in table 1, both Static CEI and Dynamic CEI reduce object hallucination on CHAIR_i and CHAIR_s for nearly all metrics comparing to the baseline methods while Dynamic CEI attains the lowest *average* CHAIR across models. The overall improvements of Dynamic over Static CEI align with its risk-aware adaptation, which adjusts intervention the commitment depth. Overall, these results indicate that reusing the pre-generation context to steer decoding is effective for curbing object hallucinations in image captioning.

AMBER. AMBER (Wang et al., 2024a) provides an LLM-free evaluation of LVLM hallucination. We use its *generative* split, where models produce free-form descriptions scored by three metrics: CHAIR (frequency of objects mentioned but not present), HAL (fraction of responses containing any hallucination across objects/attributes/relations), and COVER (fraction of image-grounded objects that are mentioned). Lower CHAIR/HAL and higher COVER indicate better faithfulness and completeness. We adopt the official prompts and scoring and set *max new tokens* to 512 to allow uninterrupted generations.

Table 2 reports results on AMBER's generative task. As shown, *Dynamic CEI* consistently achieves the lowest hallucination rates across models, outperforming all baseline methods on both CHAIR and HAL while maintaining a comparable COVER score, indicating improved faithfulness without sacrificing completeness. Despite its simplicity, *Static CEI* also surpasses most baselines and performs comparably with *CAAC*, the strongest attention-based mitigation method. In contrast, contrastive decoding approaches such as *VCD* and *M3ID* attain higher COVER values—reflecting broader descriptive coverage—but exhibit substan-

Table 1: Performance on CHAIR Benchmark

Method	LLaVA-1.5 $C_s\downarrow$	LLaVA-1.5 $C_i\downarrow$	InstructBLIP $C_s\downarrow$	InstructBLIP $C_i\downarrow$	LLaVA-NeXT $C_s\downarrow$	LLaVA-NeXT $C_i\downarrow$
base	55.2	17.6	55.6	16.6	33.0	9.4
+ OPERA	44.6	12.8	46.4	14.2	39.4	11.8
+ VCD	57.8	16.3	60.8	17.9	41.6	9.9
+ AvisC	60.4	17.2	71.0	20.1	34.8	9.3
+ M3ID	56.2	16.4	72.8	21.1	42.0	12.4
+ CAAC	39.2	**10.4**	37.4	10.8	30.6	8.1
+ Stat. CEI	35.2	13.8	32.4	**8.5**	**26.0**	7.6
+ Dyn. CEI	**34.0**	10.9	**32.2**	8.7	26.4	**7.3**

Table 2: Performance on AMBER Benchmark Across Different LVLMs

	CHAIR↓	HAL↓	COVER↑
InstructBLIP	12.8	53.5	52.7
+ OPERA	9.7	40.5	51.2
+ VCD	10.8	46.6	**53.4**
+ M3ID	10.4	47.3	51.7
+ AvisC	10.1	46.8	51.2
+ CAAC	7.0	30.9	51.9
+ Stat. CEI	6.1	31.7	**53.4**
+ Dyn. CEI	**5.6**	**30**	51.9
LLaVA-1.5	11.3	48.1	50.4
+ OPERA	7.3	29.5	47.5
+ VCD	8.2	37.3	51.9
+ M3ID	7.2	41.4	**57.3**
+ AvisC	11.0	48.0	52.5
+ CAAC	6.0	**25.0**	48.7
+ Stat. CEI	6.4	27.3	48.6
+ Dyn. CEI	**5.9**	**25.0**	48.1
LLaVA-NeXT	9.3	51.3	60.6
+ OPERA	-	-	-
+ VCD	10.5	57.2	**63.5**
+ M3ID	12.4	59.8	61.4
+ AvisC	9.2	50.4	61.1
+ CAAC	8.8	47.5	60.5
+ Stat. CEI	10.0	47.4	57.8
+ Dyn. CEI	**8.6**	**46.6**	60.4

tially elevated hallucination rates, undermining their overall reliability. These results demonstrate that CEI offers a balanced and robust mitigation strategy for open-ended generation tasks.

MMHal-Bench MMHal-Bench (Sun et al., 2023) targets hallucination in multimodal queries through 96 adversarially designed image-question pairs spanning 8 categories (e.g., object attributes, counting, spatial relations, adversarial objects) and 12 COCO meta-categories. Questions elicit detailed responses, evaluated via GPT-4 (augmented with image annotations and human-generated answers for text-only API compatibility), which rates for hallucinations with 94% human agreement. GPT-4 scores the responses according to hallucination-level and informativeness. This judge-based protocol enables nuanced assessment

Table 3: Performance comparison on MMHal-Bench (GPT4-evaluated) across different LVLMs. Scores (Sc ↑) and Hallucination Rates (HR ↓) are reported for InstructBLIP, LLaVA-1.5, and LLaVA-NeXT.

Method	InstructBLIP Sc↑	InstructBLIP HR↓	LLaVA-1.5 Sc↑	LLaVA-1.5 HR↓	LLaVA-NeXT Sc↑	LLaVA-NeXT HR↓
base	1.84	0.64	1.59	0.72	3.08	0.47
+ OPERA	2.10	0.58	2.41	0.57	–	–
+ VCD	1.75	0.64	1.96	0.64	2.82	0.57
+ M3ID	1.70	0.65	2.14	0.61	2.83	0.57
+ AvisC	2.03	0.59	2.19	0.59	3.07	0.48
+ CAAC	**2.25**	0.64	1.67	0.64	2.92	0.53
+ Stat. CEI	2.16	**0.57**	**2.46**	**0.56**	**3.10**	0.49
+ Dyn. CEI	2.21	**0.57**	2.42	0.57	3.07	**0.47**

across diverse error types, at the cost of an additional model in the loop.

As shown in Table 3, across the three LVLMs evaluated on MMHal-Bench, the proposed dynamic CEI consistently achieves the lowest hallucination rates, outperforming all baselines. This reduction in hallucinations is accompanied by competitive informativeness scores, yielding the highest overall performance in terms of balanced hallucination mitigation and response quality. Notably, static CEI ranks as the second-best method overall, demonstrating robust gains over prior techniques while highlighting the incremental benefits of dynamic adaptation in context embedding injection.

5.3 Hyperparameter Study

We analyze the influence of the scheduler and the two key parameters, α and β, on hallucination reduction.

Figure 7a compares our half-cosine mapping against a linear alternative. The half-cosine scheduler decays more gradually for low $\bar{M}_K$ values, thereby sustaining stronger corrections where commitment is weakest. In contrast, the linear scheduler declines too rapidly, attenuating interventions for low-confidence tokens and yielding elevated hallucination rates: for InstructBLIP, it produces 4.6% higher CHAIR_s and 1.7% higher CHAIR_i scores relative to half-cosine, with similar disparities observed across other LVLMs.

For cosine scheduler α controls the maximum injection strength. Small values under-correct, while large values produce over-correction that suppresses truthful tokens. As shown in Figure 7b, $\alpha = 0.17$ yeilds lower hallucination rates for LLaVA-NeXT compared with $\alpha = 0.15$ and $\alpha = 0.20$.

Additionally, we observe that β determines the

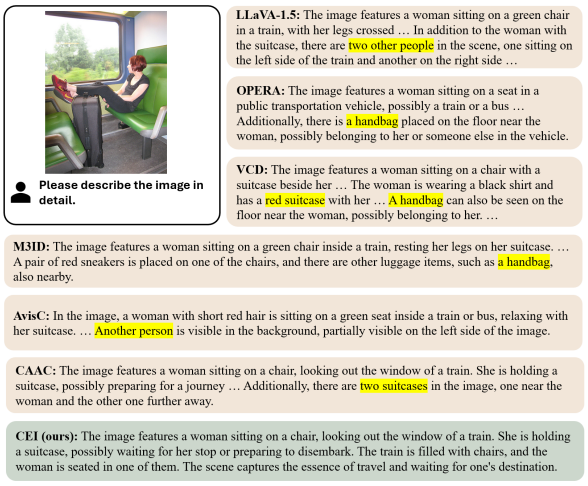


Figure 5: **Image captioning comparison.** Comparing generated captions via the baseline methods and our CEI on a sample image.

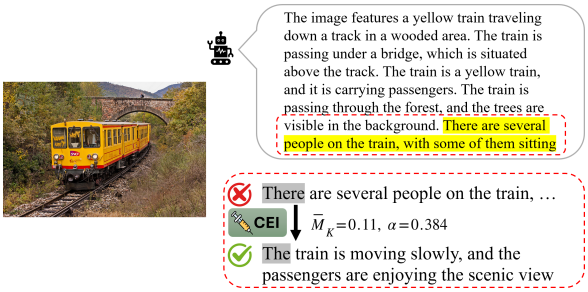


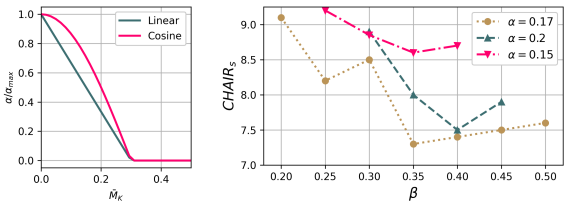
Figure 6: **Branched decoding analysis.** We generate the caption with CEI until CEI changes the top token. Then, we branch into a short greedy decoding (no CEI) to visualize how the caption would have unfolded without intervention. As shown, CEI redirects the hallucinatory continuation to a grounded one.

mean top-K mass at which injection vanishes. Too small a β keeps injection active even during confident steps (leading to over-correction), whereas too large a value disables correction prematurely. For LLaVA-NeXT, $\beta \approx 0.35$–0.40 offers the optimal hallucination reduction (Figure 7b).

5.4 Qualitative Evaluation

To complement the quantitative results, we conduct a qualitative evaluation of our method by comparing its generated captions against those from baseline models on representative samples from the AMBER benchmark. Figure 5 illustrates one such example: while baselines induce more hallucinations (e.g., handbag, red suitcase, two suitcases), our approach suppresses these erroneous concepts and remains faithful grounding to the image.

To better understand how CEI mitigates hallucinations, we conduct an analysis based on branched decoding. During generation under CEI, whenever



(a) (b)

Figure 7: (a) comparison of linear and cosine scheduler. (b) Impact of α and β on CHAIR_s for LLaVA-NeXT.

the intervention alters the model's top-token selection, we temporarily branch into a short greedy decoding (without CEI) to observe how the model would have continued if CEI did not intervene. This allows us to visualize the downstream impact of a single token swap, which is often not immediately obvious at the next step alone. A key observation is that *CEI most often intervenes at the onset of prospective hallucinatory phrases, preemptively redirecting the generation toward grounded continuations*. For instance, as depicted in Figure 6, the low average top-K probability mass ($\bar{M}_K = 0.11$) associated with the token "There" triggers a strong injection weight ($\alpha = 0.38$), swapping it for "The" and thereby sidestepping the ensuing hallucinatory elaboration.

Find more qualitative examples in appendix C.

6 Conclusion

This work bridges analysis and mitigation of hallucination in large vision–language models. Our layer-wise study revealed that truthful tokens commit earlier across decoder layers than hallucinatory ones, highlighting *commitment depth* as an interpretable signal of visual faithfulness. Leveraging this finding, we proposed *Context Embedding Injection (CEI)*, a lightweight, training-free intervention that continually aligns decoding with an image-conditioned context derived from the first generation step. The dynamic CEI variant further adjusts guidance strength online using the mean Top-K mass, improving robustness for long-form outputs. Across CHAIR, AMBER, and MMHal-Bench benchmarks, CEI achieves consistent reductions in hallucination while maintaining comparable coverage, demonstrating its effectiveness and generality. Future work may extend this framework to multi-image or video grounding, explore integration with contrastive decoding, and investigate commitment-aware steering for other multimodal reasoning tasks.

Limitations

While our study demonstrates the effectiveness of CEI for reducing hallucinations in LVLMs, several limitations remain. We discuss them here to clarify the scope of our claims and to encourage future work.

Computational Overhead. Dynamic CEI introduces additional inference cost by requiring a probe forward pass at each decoding step. This results in two forward passes per token, which may limit its applicability in latency-sensitive settings. However, this overhead is comparable to that of many existing hallucination-mitigation techniques—such as contrastive decoding methods (e.g., VCD, M3ID, AvisC)—which also require duplicate forward passes. Moreover, our static variant of CEI incurs only a single additional forward pass per instance while achieving substantial improvements and outperforming most baselines across benchmarks. We primarily view dynamic CEI as a diagnostic tool for studying adaptive interventions, with runtime optimizations left for future work.

Evaluation Scope and Generalization. Our experiments focus on three widely used, COCO-style hallucination benchmarks, leaving open questions about how CEI generalizes to more specialized domains (e.g., medical imaging, autonomous driving). Although we do not evaluate on domain-specific benchmarks, MMHal-Bench includes adversarial visual reasoning examples that extend beyond simple captioning and require more complex grounding behavior. On this benchmark, CEI consistently outperforms existing mitigation methods, suggesting that the underlying mechanism is not narrowly tailored to captioning alone. Nonetheless, evaluating CEI in domain-specific settings with distinct visual distributions remains an important direction for future work.

Dependence on Token-Level Hallucination Labels for Analysis. Our preliminary insights and mechanistic analyses rely on externally provided truthful and hallucinatory token labels from annotated datasets. This dependency may limit the range of settings in which similar analyses can be performed. Importantly, these annotations are used only for analysis and do not play any role during inference or in the CEI mechanism itself; CEI remains fully unsupervised and model-agnostic at inference time. The use of annotated benchmarks is standard practice in hallucination research, particularly for probing internal model dynamics, but extending these analyses to unannotated or partially annotated settings is a promising avenue.

White-Box Access Requirement. CEI operates by modifying hidden states and requires access to internal representations and the unembedding matrix. This limits direct applicability to closed-source LVLMs accessible only through black-box APIs. However, this requirement is shared by nearly all mechanistic or decoding-time hallucination-mitigation techniques—including attention modification methods and contrastive decoding approaches. Our focus in this work is to advance understanding and control of open LVLMs, which provide the transparency necessary for reproducibility and for deeper investigation of hallucination phenomena.

Ethics Statement

We use publicly available benchmarks and follow their licenses. Our mitigation aims to reduce factual errors; however, faithful yet harmful content remains a societal risk and should be filtered by downstream safety layers.

Potential Risks

While our goal is to reduce hallucinations, the method does not guarantee correctness and could lead to over-trust if used in high-stakes settings. Moreover, improvements in coherence and perceived grounding may be misused to produce more convincing misleading content. We therefore position this work as a research contribution and recommend pairing it with application-specific safeguards (e.g., human oversight and domain validation) in downstream deployments.

AI Assistance Disclosure

We used several AI language models to help edit and paraphrase text for clarity and grammar and as programming assistance. All technical content, experiments, and claims were verified by the authors.

References

Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Ping Chen, Xiaoqin Zhang, and Shijian Lu. 2025. Mitigating Object Hallucinations in Large Vision-Language Models with Assembly of Global and Local Attention. *arXiv preprint*. ArXiv:2406.12718 [cs].

Jinze Bai and 1 others. 2023. Qwen-vl: A versatile vision-language model. *arXiv preprint arXiv:2308.12966*.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2025. Hallucination of Multimodal Large Language Models: A Survey. *arXiv preprint*. ArXiv:2404.18930 [cs].

Fei-Long Chen, Du-Zhen Zhang, Ming-Lun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. 2023a. VLP: A Survey on Vision-language Pre-training. *Machine Intelligence Research*, 20(1):38–56.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. 2023b. Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic. *arXiv preprint*. ArXiv:2306.15195 [cs].

Zhe Chen and 1 others. 2024. Internvl: Scaling up vision foundation models. *arXiv preprint arXiv:2312.14238*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. 2024. DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models. *arXiv preprint*. ArXiv:2309.03883 [cs].

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *arXiv preprint*. ArXiv:2305.06500.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv preprint*. ArXiv:2010.11929 [cs].

Alessandro Favero and 1 others. 2024. Multi-modal hallucination control by visual information grounding. *arXiv preprint arXiv:*.

Mehrdad Fazli, Bowen Wei, Ahmet Sari, and Ziwei Zhu. 2025. Mitigating Hallucination in Large Vision-Language Models via Adaptive Attention Calibration. *arXiv preprint*. ArXiv:2505.21472 [cs].

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18135–18143. Issue: 16.

Iryna Hartsock and Ghulam Rasool. 2024. Vision-language models for medical report generation and visual question answering: a review. *Frontiers in Artificial Intelligence*, 7:1430984.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: Alleviating Hallucination in Multi-Modal Large Language Models via Over-Trust Penalty and Retrospection-Allocation.

Qidong Huang and 1 others. 2024. Opera: Over-trust penalty and retrospection-allocation. *arXiv preprint arXiv:*.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2024. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27036–27046.

Aryan Keskar, Srinivasa Perisetla, and Ross Greer. Evaluating Multimodal Vision-Language Model Prompting Strategies for Visual Question Answering in Road Scene Understanding.

Jae Myung Kim, A. Koepke, Cordelia Schmid, and Zeynep Akata. 2023. Exposing and mitigating spurious correlations for cross-modal retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2585–2595.

Sicong Leng and 1 others. 2023. Mitigating object hallucinations via visual contrastive decoding. *arXiv preprint arXiv:*.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023a. Contrastive Decoding: Open-ended Text Generation as Optimization. *arXiv preprint*. ArXiv:2210.15097 [cs].

Yifan Li and 1 others. 2023b. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft COCO: Common Objects in Context. *arXiv preprint*. ArXiv:1405.0312 [cs].

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024a. A Survey on Hallucination in Large Vision-Language Models. *arXiv preprint*. ArXiv:2402.00253 [cs].

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

Haotian Liu and 1 others. 2023. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

Shi Liu, Kecheng Zheng, and Wei Chen. 2024c. Paying More Attention to Image: A Training-Free Method for Alleviating Hallucination in LVLMs. *arXiv preprint*. ArXiv:2407.21771 [cs].

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive Text Embedding Benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

nostalgebraist. 2020. interpreting gpt: the logit lens.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. *arXiv preprint*. ArXiv:2103.00020 [cs].

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *EMNLP*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2023. Aligning Large Multimodal Models with Factually Augmented RLHF. *arXiv preprint*. ArXiv:2309.14525 [cs].

Wei Suo, Lijun Zhang, Mengyang Sun, Lin Yuanbo Wu, Peng Wang, and Yanning Zhang. 2025. Octopus: Alleviating Hallucination via Dynamic Contrastive Decoding. *arXiv preprint*. ArXiv:2503.00361 [cs].

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint*. ArXiv:2302.13971 [cs].

Junyang Wang and 1 others. 2024a. Amber: An llm-free multidimensional benchmark for mllm hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024b. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint*. ArXiv:2409.12191 [cs].

Sangmin Woo and 1 others. 2024. Don't miss the forest for the trees: Attentional vision calibration for lvlms. *arXiv preprint arXiv:*.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Philip S. Yu. 2023. Multimodal Large Language Models: A Survey. In *2023 IEEE International Conference on Big Data (BigData)*, pages 2247–2256.

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. *arXiv preprint*. ArXiv:2412.13817 [cs].

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2024. mPLUG-Owl: Modularization Empowers Large Language Models with Multimodality. *arXiv preprint*. ArXiv:2304.14178 [cs].

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2023. Woodpecker: Hallucination Correction for Multimodal Large Language Models. *arXiv preprint*. ArXiv:2310.16045 [cs].

Xiaofeng Zhang, Yihao Quan, Chaochen Gu, Chen Shen, Xiaosong Yuan, Shaotian Yan, Hao Cheng, Kaijie Wu, and Jieping Ye. 2024. Seeing Clearly by Layer Two: Enhancing Attention Heads to Alleviate Hallucination in LVLMs. *arXiv preprint*. ArXiv:2411.09968 [cs].

Qinyu Zhao, Ming Xu, Kartik Gupta, Akshay Asthana, Liang Zheng, and Stephen Gould. 2024. The first to know: How token distributions reveal hidden knowledge in large vision-language models? In *European Conference on Computer Vision*, pages 127–142. Springer.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *arXiv preprint*. ArXiv:2306.05685 [cs].

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. Analyzing and Mitigating Object Hallucination in Large Vision-Language Models. *arXiv preprint*. ArXiv:2310.00754 [cs].

Deyao Zhu and 1 others. 2023. Minigpt-4: Enhancing vision-language understanding. *arXiv preprint arXiv:2304.10592*.

Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. 2024. IBD: Alleviating Hallucinations in Large Vision-Language Models via Image-Biased Decoding. *arXiv preprint*. ArXiv:2402.18476 [cs].

Younan Zhu, Linwei Tao, Minjing Dong, and Chang Xu. 2025. Mitigating Object Hallucinations in Large

Vision-Language Models via Attention Calibration. *arXiv preprint*. ArXiv:2502.01969 [cs].

A Additional Experimental Details

Models. We evaluate CEI on three widely used 7B-parameter LVLMs: InstructBLIP (Dai et al., 2023), LLaVA-1.5 (Liu et al., 2023), and LLaVA-NeXT (Liu et al., 2024b). All models are loaded via `transformers` 4.47 in 16-bit floating-point precision with no additional finetuning. CEI is applied in a model-agnostic manner by modifying hidden states at a chosen decoder layer and passing them through the original output head. For each LVLM, the injection layer and CEI hyperparameters are selected via a small grid search on held-out images.

Benchmarks and Metrics. **CHAIR.** Following Rohrbach et al. (2018), we use the COCO-based CHAIR setup to evaluate object hallucinations in image captioning. Ground-truth object annotations are derived from COCO labels, and generated captions are mapped to object names via lemmatization and synonym lists. We report the standard instance-level

$$\mathrm{CHAIR}_i = \frac{|\mathcal{O}_{\text{hall}}|}{|\mathcal{O}_{\text{all}}|}$$

where $\mathcal{O}_{\text{hall}}$ and $\mathcal{O}_{\text{all}}$ are hallucinated and all mentioned objects, and sentence-level

$$\mathrm{CHAIR}_s = \frac{|\mathcal{S}_{\text{hall}}|}{|\mathcal{S}_{\text{all}}|}$$

where $\mathcal{S}_{\text{hall}}$ is the set of captions with at least one hallucination. Lower values indicate fewer hallucinations.

AMBER. AMBER (Wang et al., 2024a) is an LLM-free hallucination benchmark with dense human annotations covering object existence, hallucination targets, and salient objects. In the generative setting, models are asked to describe the image; noun phrases are extracted and matched to annotated objects and hallucination targets using the AMBER ontology. We report three metrics: (i) CHAIR, defined as the average fraction of hallucinated object mentions per response; (ii) Hal, the proportion of responses with any hallucination; and (iii) Cover, which measures response informativeness as the fraction of annotated objects correctly mentioned in each caption, averaged over all examples. Thus, lower CHAIR/Hal and higher Cover indicate better grounding.

MMHal-Bench. MMHal-Bench (Sun et al., 2023) consists of adversarial visual reasoning questions designed to elicit hallucinations under challenging prompts. Responses are scored on a discrete 0–5 scale by a factually augmented evaluation pipeline. We follow the official protocol and report

$$\text{Score} = \frac{1}{N}\sum_{n=1}^{N} s_n$$

and hallucination rate

$$\text{HalRate} = \frac{1}{N}\sum_{n=1}^{N} \mathbb{I}[s_n < 3],$$

where s_n is the score for question n and N is the total number of questions. Higher Score and lower HalRate correspond to better performance. For all benchmarks, we allow up to 512 new tokens to avoid truncating long generations, as hallucinations frequently occur in later decoding steps.

Baseline Methods. We compare CEI with five training-free hallucination mitigation methods:

- **OPERA** (Huang et al.) modifies beam search by penalizing over-confident beams and reallocating probability mass based on a retrospection signal. This encourages generations that remain consistent with visual evidence rather than language priors. We use the official implementation with the authors' recommended hyperparameters for each LVLM.

- **VCD** (Leng et al., 2023) (Visual Contrastive Decoding) constructs contrastive candidate paths by perturbing the visual input and comparing logits between original and perturbed runs. Tokens whose probability is not robust to visual changes are down-weighted during decoding, reducing visually unsupported continuations. We adopt the configuration provided in the AvisC repository.

- **AvisC** (Woo et al., 2024) extends contrastive decoding by introducing adaptive penalties that focus on visually grounded paths. It dynamically adjusts the contrastive strength during decoding to discourage hallucinations while preserving fluency. We use the authors' recommended hyperparameters for each model.

- **M3ID** (Favero et al., 2024) uses multi-modal contrastive signals to identify and suppress tokens that are weakly supported by visual features. By contrasting predictions under different visual conditions, it targets hallucinations that arise from over-reliance on language priors. We follow the default settings in the official codebase.

- **CAAC** (Fazli et al., 2025) (Confidence-Aware Attention Calibration) calibrates self-attention maps to re-balance attention to image tokens based on token-level confidence. It aims to counteract the skew where the decoder ignores visual features in favor of language statistics.

All baselines are evaluated using their official implementations, and we retain the decoding parameters (e.g., temperature, top-p) recommended in the original papers to ensure fair comparison.

CEI Hyperparameters and Grid Search. For dynamic CEI, we tune three hyperparameters per LVLM: the maximum injection weight $\alpha_{\max}$, the cutoff parameter β in the mean top-K mass scheduler, and the injection layer. We perform a coarse grid search over a small set of layers and a few candidate values of $\alpha_{\max}$ and β spanning low, medium, and high intervention regimes. The final settings are:

- **InstructBLIP**: $\alpha_{\max} = 0.40$, $\beta = 0.70$, injection layer $= 10$.

- **LLaVA-1.5**: $\alpha_{\max} = 0.25$, $\beta = 0.55$, injection layer $= 10$.

- **LLaVA-NeXT**: $\alpha_{\max} = 0.17$, $\beta = 0.35$, injection layer $= 10$.

Static CEI uses the same injection layer but replaces the dynamic schedule with a fixed injection weight derived from the average effective α in the dynamic setting.

Hardware and Runtime. All experiments are run on a single server with 4×NVIDIA H100 40 GB GPUs and 512 GB system RAM. We parallelize evaluation across images and models. A full AMBER run (512-token limit) takes roughly 12 hours for InstructBLIP and 10 hours for LLaVA-1.5, while CHAIR and MMHal-Bench evaluations are somewhat faster due to smaller dataset sizes. Runtime differences are mainly driven by model architecture and average caption length.

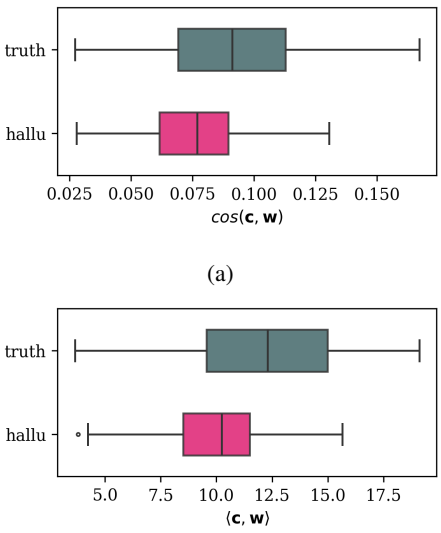


Figure 8: Box plot of (a) cosine similarities $\cos(\mathbf{c}, \mathbf{w})$, and (b) dot product $\langle \mathbf{c}, \mathbf{w} \rangle$ between the context embedding $\mathbf{c}$ and target token embedding $\mathbf{w}$. Blue and orange boxes denote truthful and hallucinatory tokens respectively.

Reproducibility. We fix random seeds, use consistent preprocessing and prompt templates across models, and share decoding parameters across baselines and CEI variants. We plan to release code, evaluation scripts, and configuration files (including all CEI hyperparameters and grid-search ranges) to facilitate reproducibility.

B Context Embedding as a Grounding Signal

To substantiate the hypothesis that the hidden state of the final prompt token—hereafter the *context embedding*—serves as a semantic grounding signal, we conduct an intrinsic analysis measuring its alignment with truthful and hallucinatory tokens. This embedding corresponds to the final-layer hidden state of the last input position (image + query) in the decoder, which is directly unembedded to produce the logits of the first generated token. The experiment aims to determine whether this representation is inherently more aligned with truthful words that describe the image faithfully, thereby justifying its use as a grounding signal in our Context Embedding Injection (CEI) framework.

Experimental setup. We perform the analysis using captions generated by InstructBLIP on the AMBER benchmark (Wang et al., 2024a), which provides human-annotated word-level labels distin-

guishing truthful and hallucinatory objects. Each generated caption is tokenized, and the annotations are aligned to the corresponding subword tokens. For multi-token words, we average the representations across their constituent sub-tokens. The representation of a token $\mathbf{w}$ is taken from the unembedding matrix of the language head (W_U), i.e., the column vector used to map hidden states to logits for token y. This choice ensures that all similarity measures are computed in the same space the model uses for generation, where the context embedding $\mathbf{c}$ naturally resides.

Metrics. We report three complementary measures of alignment: (1) the raw dot product $\langle \mathbf{c}, \mathbf{w} \rangle$, which reflects the model's true logit-space affinity for each token; (2) the raw cosine similarity $\cos(\mathbf{c}, \mathbf{w}) = \frac{\mathbf{c}^\top \mathbf{w}}{\|\mathbf{c}\| \, \|\mathbf{w}\|}$, capturing purely directional alignment; and (3) the centered cosine similarity $\cos(\mathbf{c} - \mu, \mathbf{w} - \mu)$, where μ is the mean token embedding vector across the vocabulary. The latter mitigates anisotropy in transformer embedding spaces, following the post-processing approach of Muennighoff et al. (2023). These complementary metrics allow us to disentangle whether truthful tokens align with $\mathbf{c}$ due to directional consistency, overall magnitude, or both.

Results. Across all metrics, truthful tokens exhibit stronger alignment with the context embedding than hallucinatory tokens. As shown in Figure 1, the centered cosine similarities of truthful tokens are substantially higher, and the mean difference between the two classes is significantly positive (95% CI excluding zero). Similar trends hold for the raw cosine and dot-product measures (Figure 8), demonstrating that the context embedding both points toward and activates truthful token directions in the model's output space. These findings provide direct empirical evidence that the context embedding is semantically grounded and not an arbitrary choice: it naturally encodes visually faithful information, aligning closely with truthful tokens. This supports its role as a grounding signal for CEI and aligns with prior observations that early decoding representations encode truthfulness cues (Zhao et al., 2024).

C Additional Qualitative Examples

In this section, we provide an extended set of qualitative examples to further illustrate the behavior of our method across a diverse range of images from the AMBER benchmark. For each example, we display the input image alongside captions generated by the baseline models and by our approach. Hallucinatory tokens are highlighted for ease of comparison.

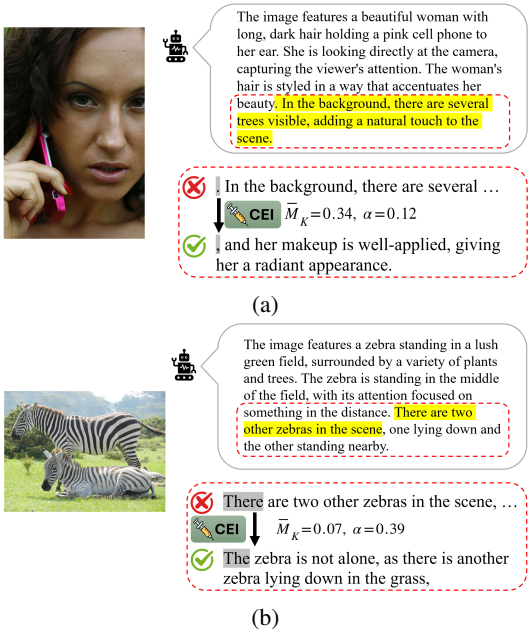


Figure 9: Examples showing how injection eliminates a hallucinatory statement and realigns the generated sequence with the input image. The captions are generated with InstructBLIP and once a token swap occurs we branch into a greedy decoding for 20 tokens to visualize how the sequence would have unfolded hadn't CEI intervened.

Across examples, we consistently observe the same pattern noted in the main text: baseline models often fail to mitigate hallucinations in long generated sequences and sometimes even introduce objects, attributes, or relations that are not visually grounded. In contrast, our method reliably suppresses these hallucinations while preserving the correct visual details.

We include a variety of challenging scenarios—cluttered scenes, small objects, atypical viewpoints, and visually ambiguous contexts (Figure 10, Figure 11).

To further elucidate the intervention mechanism of CEI, we also include examples of branched decoding analyses here (Figure 9). As described in the main text, these analyses involve generating captions under active CEI influence until a top-token swap occurs (identified by contrasting the first forward pass with no injection and second forward pass), at which point we temporarily suspend the intervention to pursue a short "greedy" decod-

ing path (without CEI) for 5–10 subsequent tokens. This branching reveals counterfactual trajectories, highlighting how CEI redirects generation away from hallucinatory continuations toward faithful descriptions. A recurring observation is that interventions typically occur at the onset of prospective hallucinatory phrases, preemptively averting downstream errors.

D Related Work

D.1 Large Vision-Language Models

Large vision-language models (LVLMs) represent a significant advancement in multimodal AI, extending the capabilities of large language models (LLMs) by integrating visual processing. Typically, LVLMs consist of a vision encoder (e.g., CLIP (Radford et al., 2021), ViT (Dosovitskiy et al., 2021)) to extract image features, an alignment module such as a linear projection (Liu et al., 2023, 2024b) or Q-former (Dai et al., 2023; Zhu et al., 2023) to map these features into the language model's embedding space, and an LLM backbone (e.g., LLaMA (Touvron et al., 2023), Vicuna (Zheng et al., 2023)) for autoregressive generation conditioned on both visual and textual inputs. This architecture enables LVLMs to handle diverse tasks, including image captioning, visual question answering (VQA), object detection, and multi-modal reasoning (Chen et al., 2023a; Wu et al., 2023), by concatenating visual tokens with text embeddings for unified processing. Recent families (Liu et al., 2024b; Wang et al., 2024b; Ye et al., 2024) scale model and data while improving visual tokenization (dynamic resolution, multi-scale inputs (Liu et al., 2024b; Wang et al., 2024b)) and positional fusion (e.g., M-RoPE (Wang et al., 2024b)), enabling broad capability gains across captioning, VQA, and free-form assistance. However, LVLMs are prone to hallucination which undermines their reliability in safety critical applications (Bai et al., 2025).

D.2 Hallucination Mitigation in LVLMs

Hallucinations in large vision-language models (LVLMs) refer to the generation of content that deviates from the visual input, such as fabricating non-existent objects, attributes, or relations. They are often attributed to over-reliance on linguistic priors, training data biases, or modality misalignment (Bai et al., 2025; Li et al., 2023b; Liu et al., 2024a). Mitigation approaches mainly fall into three categories: fine-tuning (Gunjal et al., 2024; Jiang et al., 2024; Kim et al., 2023), post-hoc correction (Yin et al., 2023; Zhou et al., 2024), and decoding-time techniques (Leng et al., 2023; Huang et al., 2024; Suo et al., 2025; Fazli et al., 2025; An et al., 2025; Yang et al., 2025).

Decoding-time methods stand out for their efficiency and applicability without retraining. Contrastive decoding (CD) (Li et al., 2023a; Chuang et al., 2024) is prominent, contrasting original and perturbed inputs to favor visual fidelity. Examples include Visual Contrastive Decoding (VCD) (Leng et al., 2023) using image noise, Multi-Modal Mutual Information Decoding (M3ID) (Favero et al., 2024) via query masking, and Image-Biased Decoding (IBD) (Zhu et al., 2024) adjusting probabilities from image-biased models. Attention-centric approaches calibrate cross-modal interactions—e.g., Assembly of Global and Local Attention (AGLA) (An et al., 2025) combines attention maps to reduce biases, and Confidence-Aware Attention Calibration (CAAC) (Fazli et al., 2025) adaptively amplifies attention to image tokens. In contrast, our method operates directly in the embedding space by continually aligning the token representations with a visual grounding signal.

Please describe the image in detail.

InstructBLIP: The image displays a very neat, tiled bathroom with a stand-up shower, toilet, basin, and a sink area … The toilet is a white color with a small towel draped next to it, and the sink is also close by. … Additionally, a mirror is placed closer …

OPERA: The image depicts a small bathroom with a toilet, sink, and bathtub … Additionally, there are two sinks in the bathroom, one on the left side and the other on the right side …

VCD: The image showcases a bathroom with a unique toilet setup, … A large mirror is situated next to the sink, … a hand towel is hanging from the side of the sink, adding to …

M3ID: The image showcases a bathroom with a white toilet situated on a tan tile floor. The bathroom is open to a doorway, providing a clear view of the toilet from the outside … There are multiple sink handles throughout the scene, possibly providing convenience and easy access to the sink for bathroom needs.

AvisC: The scene showcases a bathroom with a unique floor layout … There are multiple sinks spread across the bathroom, with one near the sink opening and two other sinks situated closer to the edge of the room …

CAAC: The image features a small bathroom with a toilet and a sink. … There is a sink located near the toilet, and a second sink is positioned towards the left … A third sink can be seen in the background, placed closer to the door …

CEI (ours): The image features a small bathroom with a toilet and a sink. The toilet is situated in the middle of the room, with a sink located next to it. The sink is placed on the right side of the toilet, and there is a door leading to the bathroom on the left side. The bathroom has a tiled floor, with the toilet bowl and sink area being tiled as well. The sink is positioned in the middle of the room, with the toilet bowl located on the left side. The bathroom has a small size, making it a cozy and intimate space.

(a)

Please describe the image in detail.

InstructBLIP: In this image, a beautiful calico cat is laying down on the ground, … There are a few flies buzzing around the cat, likely following its relaxed and peaceful presence. The cat seems to be enjoying its down time, …

OPERA: The image features a white and brown cat lying on a brick sidewalk. The cat appears to be relaxed and comfortable as it lounges on its back, exposing its belly. The cat is positioned in the center of the scene, making it the focal point of the image.

VCD: In the image, a beautiful cat is lying on its back, taking a relaxing break on a brick walkway. … its eyes closed, seemingly enjoying the sunshine … The surrounding area includes a few people, suggesting that the cat's relaxation is not exclusive to it …

M3ID: The image showcases a white and brown cat sleeping on a brick walkway. The cat is stretched out on the ground, … A second cat can also be spotted nearby, positioned closer to the center of the scene, …

AvisC: In the scene, a white and brown cat lies comfortably on the brick street, exposing its entire body. … A second cat can also be observed resting near the main cat, suggesting that it might be another cat or simply sharing the same area with its friend.

CAAC: The image features a white and brown cat lying on a brick sidewalk. … There are several other cats in the scene, with one cat lying on the ground near the white and brown cat, and another cat …

CEI (ours): A small white and brown cat is lying on a brick sidewalk, stretching out and relaxing. The cat is stretched out on the ground, taking up a significant portion of the image. The cat's body is positioned in the middle of the brick sidewalk, with its head and tail visible. The cat's relaxed pose and the brick surface create a peaceful and serene atmosphere.

(b)

Please describe the image in detail.

InstructBLIP: In the image, a young girl is shown in a field, wearing a blue soccer shirt, kicking a soccer ball … A second ball is visible off to the left side, likely the ball the girl just kicked. Another ball can be seen on the grass, …

OPERA: The image features a young girl wearing a blue soccer jersey and kicking a soccer ball … There are two soccer balls visible in the scene, one close to the girl and another further away from her. The background appears to be a cloudy day, …

VCD: The image showcases a young boy kicking a soccer ball on top of a grassy field. He is leaping into the air as … and there are several other soccer balls spread out across the field, … In addition, there is a person standing nearby …

M3ID: In the image, a young girl is kicking a soccer ball high into the air on a lush, … She appears to be leaping high with the ball in her hands, capturing the essence of the sport … Several soccer balls are present … Additionally, a person can be spotted …

AvisC: In the scene, a young soccer player is leaping into the air, kicking a white soccer ball … There are multiple soccer balls spread throughout the field, … In addition, there is another soccer ball located closer to the left edge of the field.

CAAC: a young girl kicking a soccer ball in the air on a grassy field.

CEI (ours): The image features a young girl kicking a soccer ball in a grassy field. She is wearing a blue shirt and is jumping up in the air, kicking the ball with her foot. The ball is in the air, and the girl is surrounded by the field, with the ball flying high in the sky. The girl's excitement and enthusiasm are evident in the scene.

(c)

Figure 10: Comparison of CEI outputs with baseline methods for the InstructBLIP model. Hallucinations are highlighted in yellow for easy comparison. MaxTokens is set to 512 for all models.

Please describe the image in detail.

LLaVA-1.5: The image features a stop sign located on the side of a road, with a beautiful mountain range in the background. … In addition to the stop sign, there are several cars parked along the road, with some closer to the foreground and …

OPERA: The image features a stop sign situated on the side of a road, overlooking a lush green field with mountains in the background. … In addition to the stop sign, there are several cars parked along the side of the road, …

VCD: The image features a red stop sign placed on the side of a road, with a beautiful mountain range in the background. … In the distance, there are several cars parked or driving along the road, with a total of nine cars visible. …

M3ID: The image features a stop sign located on the side of a road, overlooking a lush green field and mountains in the background. … In the background, numerous cars and trucks can be seen parked, likely enjoying the scenic surroundings. …

AvisC: The scene features a red stop sign prominently placed on the side of a road … There are several parked cars scattered throughout the area, with some located closer to the mountains and others further down the road. …

CAAC: The image features a stop sign on a pole, located in a grassy field. … Additionally, there are several cars parked in the area, with some closer to the foreground and others further back. The cars are scattered throughout the scene …

CEI (ours): The image features a stop sign located on the side of a road, with a beautiful mountain range in the background. The stop sign is positioned near the center of the scene, and the mountains can be seen in the distance, creating a picturesque landscape.

(a)

Please describe the image in detail.

LLaVA-1.5 : The image features a laptop computer sitting on a wooden desk, accompanied by a mouse and a cup of coffee … In addition to the main items, there are two books on the desk, one on the left side … A bowl can also be seen …

OPERA: The image features a wooden desk with a laptop computer, a mouse, and a cup of coffee. … Additionally, there are two books on the desk, one on the left side and the other on the right side. A person can also be seen in the background …

VCD: The image features a white laptop computer sitting on a wooden desk, accompanied by a mouse and a cup of coffee. … In addition to the laptop and mouse, there are a few books scattered around the desk …

M3ID: The image features a wooden desk with a laptop computer and a copper coffee mug. … In addition to the laptop and coffee mug, there are multiple books scattered across the desk, … A vase is also visible in the background …

AvisC: The image features a wooden dining table with various items on it. … Additionally, there are multiple books scattered across the table, … A bowl can also be seen on the right side of the table, further enriching the assortment of objects present in this setting.

CAAC: The image features a wooden desk with a laptop computer and a mouse placed on it. … In addition to the main items, there are two books on the desk, one located near the center … A bowl is also present on the desk, situated towards the right side.

CEI (ours): The image features a laptop computer sitting on a wooden desk. Beside the laptop, there is a mouse and a cup, possibly containing coffee. The scene appears to be a workspace, with the laptop being the primary focus.

(b)

Please describe the image in detail.

LLaVA-1.5: The image features a man riding a motorcycle down a tree-lined path, surrounded by a peaceful and serene environment … There are several other people in the scene, … In addition to the motorcycle, there are a few cars parked …

OPERA: The image depicts a man riding a blue motor scooter down a tree-lined street … There are a few other people in the scene, … In addition to the people, there are two cars parked on the side of the street, …

VCD: The image features an older man riding a motorcycle down a tree-lined path. … In the background, there are several cars parked along the path, … Additionally, there are two other people visible in the scene, …

M3ID: The image features a man riding a motor scooter down a tree-lined path … Several vehicles are parked or parked nearby, including two cars and two trucks, … The path also contains two benches, one on each side, …

AvisC: In the image, an older man is riding a blue and white motor scooter down a tree-lined pathway. … surrounded by several parked cars nearby. … There are multiple traffic lights positioned along the path, … Some additional motor scooters and cars are parked or stationed along the side of the road…

CAAC: The image features a man riding a motorcycle down a tree-lined street. … There are several other people in the scene, some of them walking along the street. In addition to the people, there are a few cars parked along the street, and a truck is visible in the background …

CEI (ours): The image features a man riding a motorcycle down a tree-lined street. He is wearing a white shirt and a hat, and appears to be enjoying his ride. The street is lined with trees, providing a pleasant atmosphere for the man and his motorcycle.

(c)

Figure 11: Comparison of CEI outputs with baseline methods for the LLaVA-1.5 model. Hallucinations are highlighted in yellow for easy comparison. MaxTokens is set to 512 for all models.