MITIGATING MULTIMODAL LLMS HALLUCINATIONS VIA RELEVANCE PROPAGATION AT INFERENCE TIME

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) achieve strong performance on vision- and audio-language tasks, yet can generate responses that conflict with the given visual or auditory inputs, a problem known as multimodal hallucinations. Prior work suggests that this occurs when models rely more on textual cues and learned language patterns than on evidence from the perceptual input. To obtain a more direct account of this imbalance, we apply Layer-wise Relevance Propagation (LRP), which attributes predictions to individual input tokens, and use the resulting relevance scores to analyze and mitigate hallucinations. First, we examine whether this imbalance leads to multimodal hallucinations. We find that hallucinations often arise when the model relies less on perceptual inputs, and that changing this reliance affects its predictions. We further leverage LRP and propose a training-free framework that shifts relevance toward perceptual tokens by optimizing key-value representations during decoding, without modifying model parameters or requiring training data. We call this method Learning Inference-time Modality Enhancement (LIME). Despite using no spatial or temporal supervision, LIME concentrates relevance on query-relevant regions. We evaluate LIME across multiple multimodal benchmarks in both vision and audio domains, demonstrating consistent reductions in hallucinations and enhanced grounding while preserving generation quality.

1 INTRODUCTION

Multimodal Large Language Models (MLLMs) have recently demonstrated remarkable capabilities in integrating language with perceptual modalities such as images (Liu et al., 2023; Bai et al., 2023b; 2025) and audio (Chu et al., 2024; Tang et al., 2024), enabling strong performance on tasks including visual question answering, audio reasoning, and multimodal dialogue. By coupling powerful large language models (LLMs) with perceptual-specific encoders, these systems are able to generate fluent and context-aware responses grounded in diverse sensory inputs. Despite these advances, MLLMs remain prone to hallucinations, producing outputs that are inconsistent with or unsupported by the provided multimodal evidence (Liu et al., 2024a; Huang et al., 2024b; Zheng et al., 2025).

A growing body of empirical evidence suggests that multimodal hallucinations are related to an imbalance in how models utilize different input modalities during inference (Liu et al., 2024b; Zou et al., 2025; Hsu et al., 2025). In many settings, text tokens dominate generation, while visual or auditory inputs are under-utilized. Asadi et al. (2026) indicated that the apparent visual reasoning capabilities of multimodal models may, in part, arise from strong textual priors, with models producing plausible visual descriptions even without access to the corresponding inputs. However, this imbalance is often inferred through indirect signals, leaving unclear whether reduced reliance on perceptual information actually contributes to hallucination or merely accompanies it.

To investigate this question, we employ Layer-wise Relevance Propagation (LRP) (Bach et al., 2015), an interpretability method that decomposes model predictions into the additive contributions of individual input tokens. LRP provides fine-grained *relevance* scores that quantify how strongly each token contributes to a given output. Our analysis shows that LRP uncovers a systematic imbalance in MLLMs: textual tokens dominate the relevance assigned to the input, and what perceptual tokens receive is spread diffusely rather than concentrated on the query-relevant evidence, as illustrated in Figure 1(b) and (e). We further observe that both patterns tend to be most pronounced on

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



Figure 1: Examples of multimodal hallucinations and their mitigation with LIME. Panels (a, d) show the inputs and questions. Under standard decoding (b, e), the model fails to groundits answers in the visual evidence, while with LIME (c, f) it localizes the relevant regionsand predicts correctly.

samples the model gets wrong, and that intervening on this distribution affects whether the model hallucinates.

Motivated by these findings, this paper focused on mitigating this imbalance at inference time, without additional training. Prior training-free methods pursue this goal indirectly, through raw attention weights or output shifts under perturbed inputs, without measuring whether perceptual reliance actually changes (Hsu et al., 2025; Leng et al., 2024; Liu et al., 2024b; Wang et al., 2024). We instead use LRP relevance itself as the guiding signal, as it directly quantifies each token's contribution rather than relying on proxies such as attention weights. To the best of our knowledge, this is the first work to use LRP, traditionally a post-hoc explanation method, as an inference-time guidance signal in MLLMs.

Specifically, we turn the model's relevance signal into an optimization objective that shifts relevance toward perceptual tokens by modifying key-value representations during decoding. The optimization is performed locally at each decoding step while the underlying model parameters remain fixed. We term this method *Learning Inference-time Modality Enhancement (LIME)*. Despite using no spatial or temporal supervision, LIME increases perceptual utilization and promotes more focused relevance on query-relevant regions. Further analysis shows that its annotation-free objective approaches the localization performance of a region-supervised variant. We evaluate LIME across multiple MLLMs and tasks in both vision and audio domains, demonstrating consistent reductions in hallucination and improved localization.

In summary, our contributions are the following: (i) We provide an LRP-based analysis showing that hallucinations are associated with reduced perceptual relevance, and demonstrate through targeted interventions that altering this relevance systematically changes prediction behavior; (ii) We propose a self-supervised, relevance-guided inference-time framework that mitigates hallucinations by increasing the contribution of perceptual tokens, while concentrating relevance on query-relevant regions without spatial or temporal supervision; (iii) We demonstrate the effectiveness and generality of our approach across multiple multimodal models and benchmarks. Implementation details and source code are publicly available [1].

[1] https://anonymous.4open.science/r/lime-326D/README.md

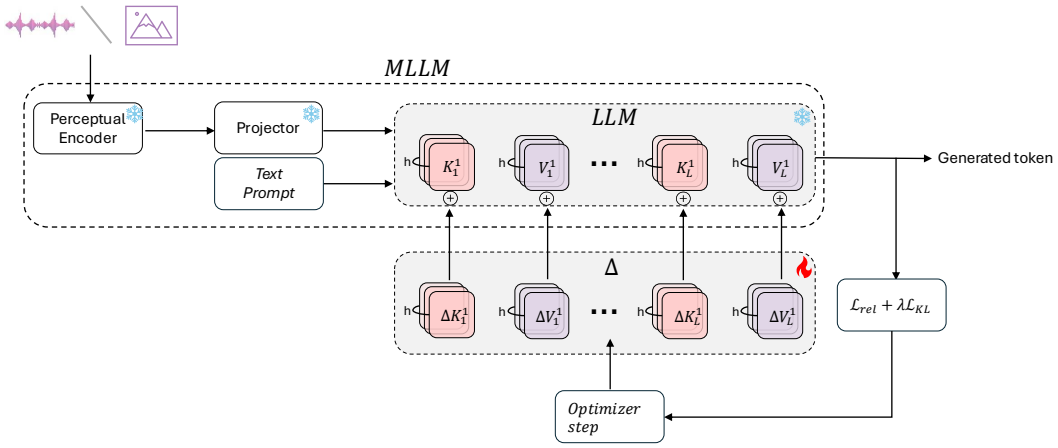


Figure 2: LIME iteratively increases the modality relevance ($\mathcal{L}_{rel}$) while remaining close to the original model distribution ($\mathcal{L}_{KL}$). The base MLLM is frozen; learning occurs only through optimizable Δ. Snowflake and flame symbols denote frozen and optimizable variables, respectively.

2 Related Work

Hallucination in MLLMs refers to the generation of content that is inconsistent with the provided perceptual inputs, such as describing nonexistent objects in images or hallucinating events and entities in audio. This phenomenon has been widely observed across vision- and audio-language models (Liu et al., 2024b; Zou et al., 2025; Zheng et al., 2025; Kuan & Lee, 2025). Existing approaches for mitigating multimodal hallucinations can be broadly categorized into training-based and training-free methods. Training-based approaches aim to improve perceptual utilization by updating model parameters through supervised fine-tuning (Gunjal et al., 2024), reinforcement learning (Yu et al., 2024), or auxiliary modules that revise hallucinated outputs (Zhou et al., 2024). In contrast, training-free methods intervene directly during inference without modifying model weights. A prominent line of training-free work focuses on attention-based interventions. For example, OPERA (Huang et al., 2024a) penalizes dominant tokens to encourage attention toward perceptual inputs. Another class of methods is based on contrastive decoding (CD) (Li et al., 2023a), which suppresses tokens that are insensitive to perceptual information by comparing outputs under perturbed or modality-removed inputs. Representative approaches include PAI (Liu et al., 2024b), Visual Contrastive Decoding (VCD) (Leng et al., 2024), Instruction Contrastive Decoding (ICD) (Wang et al., 2024), and Audio-Aware Decoding (AAD) (Hsu et al., 2025). More recent methods, such as MemVR (Zou et al., 2025) and V-ITI (Sun et al., 2025), modify hidden representations during inference to improve grounding. Despite notable progress, existing training-free methods often rely on heuristic intervention rules or indirect signals, and do not explicitly quantify the contribution of different modalities to the final prediction. This leaves the relationship between modality contribution and hallucination behavior less directly examined.

Concurrently, Elisha et al. (2026) proposed Concept-Guided Fine-Tuning (CFT), which uses LRP-based relevance maps to align model explanations with semantic concepts and reduce spurious correlations. Unlike CFT, which uses relevance to fine-tune model parameters, our method uses relevance both to characterize modality utilization and to directly guide inference-time optimization, without updating model parameters or requiring labeled training data.

3 Method

We introduce an inference-time framework that uses LRP relevance to guide decoding by increasing the contribution of perceptual tokens. At each decoding step, the method iteratively modifies the model's key and value representations based on the relevance signal, while keeping the underlying model parameters fixed. We first introduce the notation and model setup, then formulate token-level

relevance, and finally describe the inference-time optimization. An overview of the framework is provided in Figure 2.

3.1 Notation and Setup

Consider an MLLM composed of a perceptual encoder, a projection module, and an LLM backbone, as shown in Figure 2. Given a multimodal input, the perceptual encoder first transforms the raw perceptual signal (e.g., image or audio) into latent representations, which are then mapped through a projection layer into a sequence of perceptual tokens. Hence, the perceptual tokens are represented in the same space as the embedding of the textual tokens.

Let $\mathbf{x}_i$ denote the embedding vector of the i-th input token, and let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_N)$ denote the sequence of input token embeddings. We partition the input token indices into two disjoint sets: $\mathcal{P}$ for perceptual token and $\mathcal{T}$ for textual tokens. Let $\mathbf{y}_{<t} = (y_1, \ldots, y_{t-1})$ denote the sequence of generated tokens prior to decoding step t. At step t, the model predicts the next-token distribution $p_\theta(y_t \mid \mathbf{X}, \mathbf{y}_{<t})$ with corresponding logit vector $\mathbf{z}_t$. We define $z_t^* = \mathbf{z}_t[y_t]$ as the scalar logit associated with the generated token y_t.

3.2 Relevance and Problem Formulation

We define a relevance function $\Phi(z_t^*)$ that assigns a scalar relevance score to each token in $\mathbf{X}$, with respect to the output logit z_t^*. For brevity, we omit the explicit dependence on z_t^* in the notation of Φ. Let Φ_i denote the relevance assigned to the i-th token in $\mathbf{X}$.

Given the output logit z_t^*, relevance is propagated backward from the output to the input tokens. Following the basic LRP formulation (Bach et al., 2015), the relevance score Φ_j^ℓ of unit j at layer ℓ is computed recursively from the relevance $\Phi_i^{\ell+1}$ of units i at the next layer $\ell + 1$

$$\Phi_j^\ell = \sum_i \frac{a_j^\ell W_{ji}^\ell}{\sum_k a_k^\ell W_{ki}^\ell} \Phi_i^{\ell+1},$$

where $\mathbf{a}^\ell$ denotes the input to the ℓ-layer, and $\mathbf{W}^\ell$ denotes the weights associated with layer ℓ. The denominator ensures proper normalization of relevance. Relevance at lower layers is obtained by recursively applying this rule. The full transformer-specific relevance formulation, including the rules used for attention and other architectural components, is provided in Appendix A.

In our setting, relevance propagation is performed through the LLM backbone to obtain token-level relevance scores over $\mathbf{X}$. We define the total relevance assigned to perceptual and textual tokens as:

$$\Phi_\mathcal{P} = \sum_{i \in \mathcal{P}} \Phi_i \quad \Phi_\mathcal{T} = \sum_{j \in \mathcal{T}} \Phi_j. \tag{1}$$

These quantities capture the overall contribution of perceptual and textual tokens to the prediction. We use the aggregated quantities $\Phi_\mathcal{P}$ and $\Phi_\mathcal{T}$ to measure the relative contribution of perceptual and textual tokens. Our objective is therefore to increase the contribution of the *relevant* perceptual tokens during decoding by shifting relevance toward $\mathcal{P}$, while keeping the model close to its original behavior. We further analyze how this relevance redistribution relates to hallucination behavior in Section 4.3.

3.3 Inference-Time Optimization

We define $\Delta = \{\Delta\mathbf{K}, \Delta\mathbf{V}\}$ as additive updates to the key and value tensors across all transformer layers during inference. The updates have the same dimensionality as the corresponding key and value tensors and are applied across the full sequence rather than to specific token positions. These updates are computed independently at each decoding step and reset afterward.

To motivate our unsupervised relevance objective, consider first the *supervised* setting in which the query-relevant perceptual tokens are known. Let $\mathcal{G} \subseteq \mathcal{P}$ denote this subset, corresponding, for example, to image patches within the queried object's region or audio tokens within the relevant temporal segment. This set be considered as the *true label* our model should enhance. We formulate

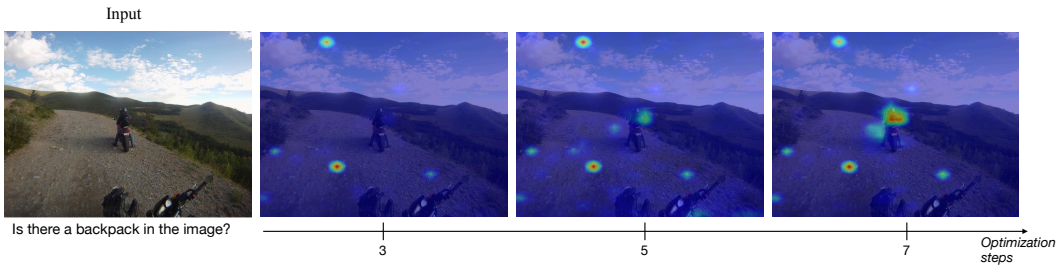


Figure 3: Evolution of visual relevance during inference-time. Relevance over the image is shown across LIME optimization steps. Relevance is initially diffuse and becomes progressively concentrated on relevant regions, indicating improved grounding.

the supervised relevance loss as a cross-entropy objective:

$$\mathcal{L}_{rel}^{sup}(\Delta) = -\frac{1}{|\mathcal{G}|}\sum_{i\in\mathcal{P}} 1_{\{i\in\mathcal{G}\}} \log \frac{\exp(\Phi_{i,\Delta}/\tau)}{\sum_{k\in\mathcal{P}\cup\mathcal{T}}\exp(\Phi_{k,\Delta}/\tau)}, \tag{2}$$

where $\Phi_{i,\Delta}$ denotes the relevance of the i-th token under the key and value updates, $\tau > 0$ is a temperature parameter, and $1_{\{\cdot\}}$ is an indicator function. Minimizing this objective increases the relative relevance assigned to the query-relevant perceptual tokens $\mathcal{G}$.

However, identifying $\mathcal{G}$ requires spatial or temporal annotations, which are unavailable in the inference setting considered here. We therefore relax the supervised objective by replacing $\mathcal{G}$ with the full set of perceptual tokens $\mathcal{P}$, treating each perceptual token as potential relevant evidence. This yields our annotation-free relevance objective

$$\mathcal{L}_{rel}(\Delta) = -\frac{1}{|\mathcal{P}|}\sum_{i\in\mathcal{P}} \log \frac{\exp(\Phi_{i,\Delta}/\tau)}{\sum_{k\in\mathcal{P}\cup\mathcal{T}}\exp(\Phi_{k,\Delta}/\tau)}. \tag{3}$$

The resulting objective is self-supervised: its guidance is derived entirely from the model's own relevance decomposition and requires no knowledge of which perceptual tokens are relevant to the query. Despite this coarser supervision, we find that the induced relevance becomes increasingly concentrated on query-relevant regions, approaching the localization achieved by the supervised formulation, as analyzed in Section 4.3. To explain why this localization emerges, we provide a local analysis in Appendix C.2, showing how the frozen model's sensitivity structure filters the broadly distributed objective, resulting in relevance changes that tend to align with query-relevant evidence.

To constrain deviations from the pretrained model, we use a KL-divergence regularizer between the next-token distributions produced with and without the key-value updates, denoted by $p_{\theta,\Delta}(y_t \mid \mathbf{X}, \mathbf{y}_{<t})$ and $p_\theta(y_t \mid \mathbf{X}, \mathbf{y}_{<t})$, respectively. We define

$$\mathcal{L}_{KL}(\Delta) = D_{KL}(p_{\theta,\Delta}(y_t \mid \mathbf{X}, \mathbf{y}_{<t}) \,||\, p_\theta(y_t \mid \mathbf{X}, \mathbf{y}_{<t})). \tag{4}$$

This term constrains modifications to the key and value representations to preserve the pretrained model's linguistic prior and decoding stability.

Combining both loss terms (Eq. (3) and Eq. (4)), the updates Δ of $\mathbf{KV}$ are optimized at each decoding step by minimizing the following composite loss:

$$\arg\min_{\Delta} \mathcal{L}_{rel}(\Delta) + \lambda\mathcal{L}_{KL}(\Delta), \tag{5}$$

where $\lambda > 0$ controls the trade-off between promoting multimodal relevance and constraining deviations from the reference model. Minimizing this objective aims to promote key and value representations that amplify the explanatory responsibility of perceptual tokens while maintaining linguistic coherence and decoding stability. The key and value perturbations Δ are initialized to zero and updated using a small number of gradient-based optimization steps. Since the relevance objective depends on LRP relevance, optimizing Δ requires second-order derivatives through the relevance computation. Importantly, these updates are discarded after each step, ensuring that control remains local and does not accumulate across time.

Table 1: POPE results with LLaVA-1.5-7B on MSCOCO. Accuracy (Acc) and F1 are reported for the random, popular, and adversarial splits and their average. Best and second-best results are in **bold** and underlined, respectively.

Dataset	Methods	Random (%) ↑		Popular (%) ↑		Adversarial (%) ↑		Average (%) ↑	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1
MSCOCO	LLaVA-1.5-7B	83.49	82.28	79.98	79.34	76.03	76.26	79.83	79.29
	+ OPERA (Huang et al., 2024a)	87.53	86.45	84.21	83.5	80.88	80.69	84.21	83.55
	+ VCD (Leng et al., 2024)	86.84	86.83	82.65	83.37	77.31	79.28	82.27	83.16
	+ ICD (Wang et al., 2024)	84.87	83.27	82.93	81.45	81.07	79.96	82.96	81.56
	+ MemVR (Zou et al., 2025)	88.5	87.34	87.1	86.01	85.2	84.28	86.93	85.88
	+ V-ITI (Sun et al., 2025)	89.74	87.72	84.96	84.77	**86.31**	82.44	87	84.98
	+ LIME (ours)	**90.27**	**89.75**	**87.91**	**87.85**	85.51	**84.52**	**87.89**	**87.37**

4 EXPERIMENTS

We next present the empirical evaluation of our approach. We first report results on vision and audio benchmarks, then examine the relationship between modality relevance and hallucination behavior, and finally provide an ablation study. Details on datasets, evaluation protocols, and metrics are deferred to Appendix B.1. Relevance scores are computed using Attention-Aware Layer-wise Relevance Propagation (AttnLRP) (Achtibat et al., 2024), which adapts LRP to transformer architectures through attention-specific propagation rules. Implementation details and hyperparameter settings are given in Appendix B.2. All experiments were conducted on 8 NVIDIA A100 GPUs.

4.1 VISION BENCHMARKS

The five MLLMs we used for vision evaluation were LLaVA-1.5-7B (Liu et al., 2023), Qwen-VL-Chat (Bai et al., 2023b), Qwen2.5-VL-7B-Instruct (Bai et al., 2025), Qwen2.5-Omni-7B (Xu et al., 2025), and MiniCPM-o 4.5 (Cui et al., 2026). LLaVA-1.5-7B and the Qwen-based models employ CLIP (Radford et al., 2021) as visual encoder, paired with LLaMA (Touvron et al., 2023) and Qwen (Bai et al., 2023a) LLM backbones respectively, while MiniCPM-o 4.5 uses LLaVA-UHD (Guo et al., 2024) with a Qwen3-8B backbone. We used their pretrained, instruction-tuned variants without additional fine-tuning, adhering to a strictly zero-shot setting. We compared LIME against the training-free baselines OPERA (Huang et al., 2024a), VCD (Leng et al., 2024), ICD (Wang et al., 2024), MemVR (Zou et al., 2025), and V-ITI (Sun et al., 2025), reporting comparisons where reproducible, since some are not applicable across all architectures.

The first benchmark we considered was POPE (Li et al., 2023b), which frames hallucination detection as a binary yes/no question-answering task constructed from the MSCOCO (Lin et al., 2014) and A-OKVQA (Schwenk et al., 2022) datasets, using three splits: random, popular, and adversarial (see Appendix B.1 for more details). Table 1 reports accuracy and F1 for LLaVA-1.5-7B across all baselines on MSCOCO images for each split. Our proposed approached, LIME, outperforms all baselines on most metrics, achieving the highest average accuracy and F1, with the largest improvements over the base model on the adversarial split, indicating increased robustness to challenging negative samples. A similar trend holds for other models and for A-OKVQA (Appendix B.3).

The second vision benchmark we considered was CHAIR (Rohrbach et al., 2018), which quantifies hallucinations by evaluating object mentions in generated captions on MSCOCO. Performance is measured using CHAIR_S, the percentage of sentences containing at least one hallucinated object, and CHAIR_I, the fraction of hallucinated mentions among all generated object instances (lower is better for both). As reported in Table 2, our method reduces hallucination rates across all evaluated models, indicating improved utilization of visual tokens. While some methods attain higher recall, they often do so at the expense of increased hallucinations; in contrast, our approach achieves a more favorable trade-off between faithfulness and coverage.

Finally, we evaluate LIME on HallusionBench (Guan et al., 2024), extending our analysis beyond object presence and captioning to diverse image-context reasoning scenarios. As reported in Table A.6, LIME improves both LLaVA-1.5 and Qwen2.5-VL-7B-Instruct across all reported metrics. Overall, these results demonstrate that our method effectively mitigates visual hallucinations while maintaining strong performance across models and datasets.

Table 2: CHAIR evaluation results. CHAIR_S, CHAIR_I, their average, and Recall are reported. Best and second-best results are in **bold** and underlined, respectively.

Methods	$\text{CHAIR}_S \downarrow$	$\text{CHAIR}_I \downarrow$	Average $\downarrow$	Recall $\uparrow$
LLaVA-1.5-7B	52	15.8	32.7	75.2
+ OPERA	47.8	14.6	31.8	77.3
+ VCD	48.6	14.9	31.2	76.8
+ ICD	56.2	16.3	36.3	16.3
+ MemVR	46.6	**13**	29.8	**80.8**
+ V-ITI	46.1	13.5	29.8	80.4
+ LIME (ours)	**42.7**	**13**	**27.85**	72
Qwen-VL-Chat	46	12.5	29.3	64.3
+ VCD	46.8	12.3	29.6	67.9
+ ICD	45	14.3	29.7	47.6
+ V-ITI	**44.2**	12.5	28.4	66.4
+ LIME (ours)	44.5	**12**	**28.25**	**68.7**
Qwen2.5-VL-7B-Instruct	25.6	9.1	17.35	55.1
+ LIME (ours)	**21.2**	**8.2**	**14.7**	**56.5**

Table 3: Accuracy and F1 scores for the Audio Hallucination QA and Air-Bench benchmarks.

Methods \ Dataset	AHQA (%) ↑						Air-Bench (%) ↑		
	Random		Popular		Adversarial		Speech	Sound	Music
	Acc	F1	Acc	F1	Acc	F1			
SALMONN-7B	53.91	23.37	49.32	18.27	50.31	20.01	37.51	33.58	31.05
+ AAD (Hsu et al., 2025)	**57.22**	36.74	48.71	18.78	48.04	17.42	42.62	34.56	30.38
+ LIME (ours)	56.88	**36.85**	**53.12**	**25.76**	**54.35**	**26**	**45.2**	**36.9**	**31.95**
Qwen2-Audio-7B-Instruct	56.19	26.78	51.34	20.5	50.13	20.24	57.56	60.86	55.89
+ AAD (Hsu et al., 2025)	59.5	31.62	51.31	12.09	51.29	11.43	60	61.9	**57.43**
+ LIME (ours)	**63.36**	**50.27**	**57.53**	**46.43**	**53.1**	**37.08**	**66.1**	**66.41**	56.25

4.2 AUDIO BENCHMARKS

The three MLLMs we used for audio evaluation were SALMONN-7B (Tang et al., 2024), Qwen2-Audio-7B-Instruct (Chu et al., 2024), and Qwen2.5-Omni-7B. SALMONN-7B employs a Whisper (Radford et al., 2023) encoder large v2 paired with Vicuna (Chiang et al., 2023) LLM backbone. Both Qwen2-Audio-7B-Instruct and Qwen2.5-Omni-7B also use Whisper encoder large v3 paired with Qwen LLM backbone. We used the default configurations from the original papers for all methods. Our proposed model is compared against AAD (Hsu et al., 2025), the only hallucination mitigation training-free method we found. The results reported for this model are based on our own implementation, as the original authors have not released their code.

The first audio benchmark, Audio Hallucination QA (AHQA) (Kuan et al., 2024), formulates hallucination detection as a binary question-answering task: given an audio clip, models are asked yes/no questions of the form "Is there a sound of [object] in the audio?". The second audio benchmark was AIR-Bench (Yang et al., 2024), which evaluates generative audio understanding across three domains: speech, sound, and music. In this setting, questions are presented in a multiple-choice format, but models must generate answers freely, making the evaluation fully generative.

Results summarized in Table 3 and Appendix B.3. On AHQA, our method improves both accuracy and F1 score across all sampling strategies. For Qwen2-Audio-7B-Instruct, we observe substantial gains over prior methods, particularly in F1 score. Similar improvements are observed for SALMONN-7B and Qwen2.5-Omni-7B. On AIR-Bench, our method improves performance across all domains, with consistent gains in speech and sound understanding and competitive performance in music tasks. These results suggest that LIME not only reduces hallucinations but also enhances general multimodal reasoning in audio settings.

Table 4: Perceptual utilization and relevance localization across models and configurations.

	Perceptual Utilization ↑		Relevance Localization ↑						
			Baseline	LIME			LIME-Sup		
Model	Baseline	LIME	–	3	5	7	3	5	7
LLaVA-1.5-7B	0.10	0.17	0.27	0.29	0.315	0.351	0.32	0.361	0.37
Qwen2-Audio-7B	0.34	0.42	0.31	0.4	0.5	0.57	0.491	0.56	0.585

4.3 Modality Utilization Analysis

To better understand how modality utilization changes under LIME, we conduct a relevance-based analysis using LRP. Since LRP conserves relevance (Bach et al., 2015), LIME cannot change how much relevance exists, only how it is distributed among tokens. Here, we ask two questions about this distribution. First, whether relevance shifts from textual to perceptual tokens relative to standard decoding. Second, how the increased perceptual relevance is distributed within the perceptual input: since Eq. (3) treats all perceptual tokens symmetrically and uses no spatial or temporal supervision, the shift could in principle be spread uniformly rather than concentrated on the evidence relevant to the query. Separately, Appendix C.1 examines whether these relevance shifts are linked to hallucination behavior. It first compares relevance changes across samples with different prediction outcomes, and then tests how behavior changes when relevance is increased or decreased for different token groups.

In the vision setting, we follow the POPE benchmark protocol and randomly sample 100 examples with a positive label, using the bounding box of the queried object as the annotated relevant region. For audio, we analogously sample 100 positive examples from DCASE 2019 Task 4 (Turpault et al., 2019), generate queries of the form "Is [sound event] present in the audio?", and use the annotated temporal segment of the queried event as the relevant region. For each sample, we generate responses using both the baseline and our method and compute token-level relevance scores with respect to the predicted answer, normalized to $[0, 1]$ across samples.

Following the notation in Eq. (1), let $\Phi_{\mathcal{P}}$ and $\Phi_{\mathcal{T}}$ denote the total relevance assigned to perceptual and textual tokens, respectively. Using the query-relevant perceptual region $\mathcal{G}$ defined in Section 3.3, we define *relevance localization* as the fraction of perceptual relevance localized within the ground-truth region, given by $\sum_{i\in\mathcal{G}} \Phi_i/\Phi_{\mathcal{P}}$. Since this quantity is normalized by $\Phi_{\mathcal{P}}$, it is invariant to a uniform rescaling of relevance across perceptual tokens. In addition, we define *perceptual utilization* as $\Phi_{\mathcal{P}}/(\Phi_{\mathcal{P}} + \Phi_{\mathcal{T}})$, measuring the relative contribution of perceptual evidence to the prediction. For relevance localization, we report two configurations of the full inference-time objective in Eq. (5): one using our annotation-free relevance loss in Eq. (3), and one using the supervised relevance loss in Eq. (2), which we denote *LIME-Sup*. In both cases, the same KL regularization is applied. We evaluate each at three, five, and seven optimization steps to observe how the redistribution evolves. As stated in Section 3.3, LIME does not assume knowledge of $\mathcal{G}$, which is used here for analysis only, and since LIME-Sup targets the same region over which relevance localization is measured, it serves as an upper bound rather than a deployable configuration.

The results are presented in Table 4. We observe that our approach consistently improves both relevance localization and perceptual utilization across vision and audio models. In particular, the increase in perceptual utilization indicates that the model assigns greater explanatory responsibility to perceptual tokens, while the improvement in relevance localization suggests that this relevance becomes more concentrated on semantically relevant regions. The redistribution within $\mathcal{P}$ is therefore selective, favoring the query-relevant region even though the objective does not identify it. Comparing the two configurations quantifies how selective it is, LIME-Sup is stronger at every budget, but the gap narrows as optimization proceeds: by seven steps our objective attains 81% and 95% of its improvement over vanilla decoding on the two models. The annotation-free objective therefore converges toward the region-supervised one rather than plateauing at a coarser solution, without access to $\mathcal{G}$. Together, these results show that LIME not only shifts relevance from textual to perceptual tokens, but also selectively concentrates perceptual relevance within query-relevant regions, despite having no access to $\mathcal{G}$. As noted earlier, Appendix C.2 provides a local analysis of why this selective localization can emerge despite the objective having no access to the query-relevant regions.

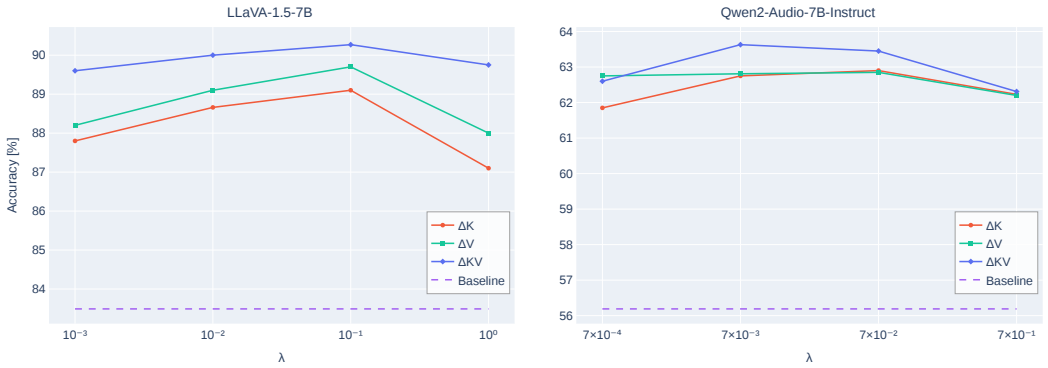


Figure 4: Effect of KL regularization λ under different editing strategies. Here, $\Delta\mathbf{K}$ and $\Delta\mathbf{V}$ denote optimizing only the key or value perturbations, respectively, while $\Delta\mathbf{KV}$ denotes their joint optimization and corresponds to Δ in the main text. Results are shown for LLaVA-1.5-7B (left) and Qwen2-Audio-7B-Instruct (right).

4.4 Ablation Studies

We next examine the main design choices of LIME by studying the roles of $\Delta\mathbf{K}$ and $\Delta\mathbf{V}$, and the sensitivity to the KL-regularization weight λ. Efficiency ablations varying the number of optimization steps and the layers to which Δ is applied, together with latency and memory measurements, are reported in Appendix D.1. Unless otherwise stated, the vision experiments use LLaVA-1.5-7B on the random split of POPE-MSCOCO, while the audio experiments use Qwen2-Audio-7B-Instruct on the random split of AHQA.

Key–value updates and KL regularization. We first evaluate three variants of LIME: optimizing only the keys $\Delta\mathbf{K}$, only the values $\Delta\mathbf{V}$, or both jointly Δ. For each variant, we sweep the KL-regularization weight λ, while fixing the number of optimization steps to seven. The results are presented in Figure 4. All variants outperform the baseline, indicating that key and value updates independently contribute to improving multimodal grounding. Among them, jointly modifying both consistently achieves the best performance, indicating that these components play complementary roles in controlling attention and information aggregation. Smaller values of λ enable stronger updates to the model's internal representations, while larger values constrain the optimization and reduce its effect. Across both domains, intermediate values of λ provide the best trade-off between enhancing modality utilization and preserving the model's original behavior.

5 Conclusion

We studied multimodal hallucinations through the lens of modality utilization at inference time. Using LRP, we directly characterize the contribution of perceptual and textual tokens, showing that hallucinations are associated with reduced perceptual relevance and that intervening on this relevance systematically changes prediction behavior. Building on this analysis, we introduced LIME, a training-free framework that uses the model's own relevance signal to optimize key-value representations during decoding without modifying model parameters or requiring training data. Across vision and audio benchmarks, LIME consistently reduces hallucinations while preserving generation quality. Moreover, despite receiving no spatial or temporal supervision, LIME increases perceptual utilization and concentrates relevance on query-relevant regions, demonstrating that relevance provides an effective inference-time signal for controlling multimodal generation.

LIME incurs additional computational overhead due to per-step optimization during decoding, as discussed in detail in Appendix D.2. Nevertheless, the optimization budget and intervention scope can be adjusted to obtain lower-cost variants while retaining meaningful gains. The method also requires hyperparameter selection, which may limit its applicability in latency-sensitive or resource-constrained settings.

AI USE STATEMENT

In this work, we used generative AI tools solely to improve the presentation of the manuscript, including grammar, readability, and clarity. All research ideas, methodological developments, experiments, and analyses were conceived and conducted by the authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. AttnLRP: Attention-aware layer-wise relevance propagation for transformers. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 135–168. PMLR, 21–27 Jul 2024.

Mohammad Asadi, Jack W O'Sullivan, Fang Cao, Tahoura Nedaee, Kamyar Fardi, Fei-Fei Li, Ehsan Adeli, and Euan Ashley. Mirage the illusion of visual understanding. *arXiv preprint arXiv:2603.21687*, 2026.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023a.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023b.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023. URL `https://lmsys.org/blog/2023-03-30-vicuna/`.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, et al. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction. *arXiv preprint arXiv:2604.27393*, 2026.

Yehonatan Elisha, Oren Barkan, and Noam Koenigstein. Concept-guided fine-tuning: Steering vits away from spurious correlations to improve robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17035–17045, June 2026.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14375–14385, 2024.

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

Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. In *European Conference on Computer Vision*, pp. 390–406. Springer, 2024.

Tzu-wen Hsu, Ke-Han Lu, Cheng-Han Chiang, and Hung-yi Lee. Reducing object hallucination in large audio-language models via audio-aware decoding. *arXiv preprint arXiv:2506.07233*, 2025.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427, 2024a.

Wen Huang, Hongbin Liu, Minxin Guo, and Neil Gong. Visual hallucinations of multi-modal large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 9614–9631, 2024b.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Chun-Yi Kuan and Hung-yi Lee. Can large audio-language models truly hear? tackling hallucinations with multi-task assessment and stepwise audio reasoning. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Chun-Yi Kuan, Wei-Ping Huang, and Hung-yi Lee. Understanding sounds, missing the questions: The challenge of object hallucination in large audio-language models. In *Proc of International Speech Communication Association (INTERSPEECH)*, 2024.

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

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori B Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 12286–12312, 2023a.

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024a.

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

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, pp. 125–140. Springer, 2024b.

Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern recognition*, 65:211–222, 2017.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1437. URL https://aclanthology.org/D18-1437/.

Nikunj Saunshi, Jordan Ash, Surbhi Goel, Dipendra Misra, Cyril Zhang, Sanjeev Arora, Sham Kakade, and Akshay Krishnamurthy. Understanding contrastive learning requires incorporating inductive biases. In *International Conference on Machine Learning*, pp. 19250–19286. PMLR, 2022.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pp. 146–162. Springer, 2022.

Nan Sun, Zhenyu Zhang, Xixun Lin, Kun Wang, Yanmin Shang, Naibin Gu, Shuohuan Wang, Yu Sun, Hua Wu, Haifeng Wang, et al. V-iti: Mitigating hallucinations in multimodal large language models via visual inference-time intervention. *arXiv preprint arXiv:2512.03542*, 2025.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=14rn7HpKVk.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Nicolas Turpault, Romain Serizel, Ankit Parag Shah, and Justin Salamon. Sound event detection in domestic environments with weakly labeled data and soundscape synthesis. In *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2019.

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

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, et al. Air-bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1979–1998, 2024.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13807–13816, 2024.

Kening Zheng, Junkai Chen, Yibo Yan, Xin Zou, Huiyu Zhou, and Xuming Hu. Reefknot: A comprehensive benchmark for relation hallucination evaluation, analysis and mitigation in multimodal large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 6193–6212, 2025.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *International Conference on Learning Representations*, volume 2024, pp. 56969–56998, 2024.

Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Kening Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, and Xuming Hu. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. *Forty-second International Conference on Machine Learning (ICML)*, 2025.

APPENDIX

A LAYER-WISE RELEVANCE PROPAGATION FOR TRANSFORMER MODELS

This appendix provides additional details on the Layer-wise Relevance Propagation (LRP) framework used to compute token-level relevance scores in our method. We briefly review the relevance decomposition principle, the conservation property, and the transformer-specific propagation rules used throughout this work. Our formulation follows the Attention-Aware Layer-wise Relevance Propagation (Achtibat et al., 2024) (AttnLRP) framework, from which the transformer-specific propagation rules presented below are adopted.

LRP is a class of attribution methods that explains a scalar model output by decomposing it into additive contributions of intermediate representations and input features. Consider scalar function f_j with input vector $\mathbf{x}$ of size N. LRP assigns relevance values $\{\Phi_{i\leftarrow j}\}_{i=0}^{N-1}$, where $\Phi_{i\leftarrow j}$ reflects the amount of the output f_j that is attributable to input i. These relevance values are defined such that their sum is proportional to the function output:

$$f_j(\mathbf{x}) \propto \Phi_j = \sum_i \Phi_{i\leftarrow j}. \tag{6}$$

When an input i contributes to multiple neurons j, its total relevance, Φ_i, is obtained by aggregating its contributions across all connected outputs:

$$\Phi_i = \sum_j \Phi_{i\leftarrow j}. \tag{7}$$

A central property of LRP is the *conservation law*. When relevance is propagated backward through a layered network, the total relevance is preserved across layers. Denoting by Φ_j^ℓ the relevance assigned to neuron j at layer ℓ, relevance propagation satisfies the following *conservation law*:

$$\Phi^{\ell-1} = \sum_i \Phi_i^{\ell-1} = \sum_j \Phi_j^\ell = \Phi^\ell. \tag{8}$$

This property ensures that relevance is neither created nor destroyed as it flows from the output layer toward the input. As a result, relevance values remain comparable across layers and retain a direct relationship to the original scalar output being explained.

Unlike gradient-based attribution methods, whose magnitudes may shrink or explode through depth, LRP maintains a relevance-preserving decomposition. To derive propagation rules, LRP is commonly formulated through Deep Taylor Decomposition (DTD) (Montavon et al., 2017). The key idea is to locally approximate each neuron's computation by a first-order Taylor expansion around a reference point, and to interpret the resulting decomposition as additive relevance contributions. Consider a neuron j represented as a scalar function $f_j(\mathbf{x})$ where $\mathbf{x} = (x_0, ..., x_{N-1})$ denotes its input activations. DTD approximates this function locally by a first-order Taylor expansion around a reference point $\mathbf{x}^0 = (x_0^0, ..., x_{N-1}^0)$

$$f_j(\mathbf{x}) \approx f_j(\mathbf{x}^0) + \sum_i \frac{\partial f_j}{\partial x_i}(\mathbf{x}^0)(x_i - x_i^0).$$

Rearranging terms yields an affine approximation of the form

$$f_j(\mathbf{x}) \approx \sum_i \mathbf{J}_{ji} x_i + b_j^0,$$

where $\mathbf{J}$ denotes the Jacobian evaluated at the reference point, and b_j^0 collects the bias and higher-order approximation error.

Under the additive relevance assumption (Achtibat et al., 2024), relevance is taken to be proportional to the neuron's output. Scaling the affine approximation by a constant factor $c \in \mathbb{R}$ with $f_j(\mathbf{x}) \neq 0$. Following Eq (6) yields

$$\Phi_j = c f_j(\mathbf{x}) = \sum_i \underbrace{\mathbf{J}_{ji} x_i \frac{\Phi_j}{f_j(\mathbf{x})}}_{\Phi_{i \leftarrow j}} + \underbrace{b_j^0 \frac{\Phi_j}{f_j(\mathbf{x})}}_{\Phi_{b \leftarrow j}}. \tag{9}$$

Comparing with Eq (6), the first term in the decomposition corresponds to the relevance assigned to the input variables, while the second term captures the contribution of bias and higher-order approximation error. Following the AttnLRP formulation, we focus on propagating input-dependent relevance and omit the bias-related term, since it does not encode input-specific information and can be modeled as a constant contribution. This simplification preserves the conservation property (8) and does not affect the interpretation of token-level relevance. Applying Eq (7) and Eq (9) yields

$$\Phi_i = \sum_j \Phi_{i \leftarrow j} = \sum_j \mathbf{J}_{ji} x_i \frac{\Phi_j}{f_j(\mathbf{x})}.$$

The expression above provides a general relevance decomposition derived from local linearization. For neural network layers, it can be written in a more practical form using activations and layer weights, resulting in the standard LRP-z rule:

$$\Phi_j^\ell = \sum_i \frac{x_j^\ell W_{ji}^\ell}{\sum_k x_k^\ell W_{ki}^\ell} \Phi_i^{\ell+1},$$

where $\mathbf{x}^\ell$ denotes the activations at layer ℓ, and $\mathbf{W}^\ell$ denotes the associated weights. The denominator normalizes the redistribution such that relevance conservation is preserved across layers. This rule forms the basis for propagating relevance through linear transformations.

While the LRP-z rule is suitable for many linear operations, transformer architectures contain additional components whose computations do not naturally follow simple linear redistribution. Consequently, specialized propagation rules are required to maintain stable and faithful attribution.

LRP-ϵ rule. The first extension is the LRP-ϵ rule, which stabilizes propagation when denominators become numerically small. In deep networks, the normalization term in the z-rule may approach zero, producing unstable relevance assignments. To address this, a small stabilizing term is introduced:

$$\Phi_j^\ell = \sum_i \frac{x_j^\ell W_{ji}^\ell}{\sum_k x_k^\ell W_{ki}^\ell + \epsilon \,\mathrm{sign}\big(\sum_m x_m^\ell W_{mi}^\ell\big)} \Phi_i^{\ell+1},$$

where $\epsilon > 0$ is a small constant. This stabilization prevents division by values close to zero while preserving the overall relevance decomposition. As a result, relevance propagation becomes more robust in deep transformer architectures where activations may vary substantially across layers.

LRP for softmax normalization. Beyond linear mappings, transformer models also include normalization operations that redistribute activations across dimensions. In particular, the softmax function introduces competition between neurons through normalization. Let

$$a_j^\ell = \mathrm{softmax}(\mathbf{x}^\ell)_j = \frac{\exp(x_j^\ell)}{\sum_k \exp(x_k^\ell)}.$$

Relevance propagation through softmax cannot be treated as an identity operation because each output depends on all input dimensions. Instead, relevance is redistributed according to

$$\Phi_j^{\ell-1} = x_j^\ell \left(\Phi_j^\ell - a_j^\ell \sum_i \Phi_i^\ell \right).$$

This formulation captures the competitive normalization behavior of softmax while enabling stable relevance propagation across dimensions.

LRP for matrix decomposition in attention. For the attention-value interaction, relevance must be propagated through the matrix product between attention weights and value representations. Let

$$\mathbf{O}^\ell = \mathbf{A}^\ell \mathbf{V}^\ell,$$

where $\mathbf{A}^\ell$ denotes the attention matrix and $\mathbf{V}^\ell$ the value tensor of layer ℓ. Relevance propagation through this bilinear interaction is decomposed across the matrix product. The relevance assigned to attention entry A^ℓ_{ji} is given by

$$\Phi^{\ell-1}_{ji} = \sum_p \frac{A^\ell_{ji} V^\ell_{ip}}{2O^\ell_{jp} + \epsilon} \Phi^\ell_{jp}.$$

An analogous rule is applied to the value branch by summing over the complementary dimension. This decomposition preserves conservation while enabling attribution through attention-value interactions.

LRP for LayerNorm and RMSNorm. Finally, transformer architectures contain normalization layers such as LayerNorm and RMSNorm. These operations primarily rescale activations rather than redistribute information across features. Consequently, relevance propagation through normalization layers is approximated using an identity mapping

$$\Phi^{\ell-1}_i = \Phi^\ell_i.$$

This approximation assumes that normalization preserves feature attribution while only modifying activation scale, allowing relevance to pass unchanged through these layers.

Together, these propagation rules provide a complete relevance decomposition framework for transformer architectures. By combining stabilized linear propagation, normalization-aware redistribution, and attention-specific matrix decomposition, relevance can be propagated consistently from output predictions back to multimodal input tokens.

B Experimental Setup

B.1 Evaluation Benchmarks

We provide additional details on the datasets and evaluation procedures used in our experiments. For each benchmark, we specify the prompting strategy and decoding parameters, including the maximum generation length. Unless stated otherwise, all methods are evaluated under identical prompting and decoding settings.

POPE (Li et al., 2023b) Polling-based Object Probing Evaluation (POPE) is a benchmark designed to measure object hallucination in large vision–language models through a binary visual question answering task. Given an image, models are asked yes/no questions of the form "Is there a [object] in the image?". Positive questions correspond to objects that are present in the scene according to ground-truth annotations, while negative questions query objects that do not appear in the image. Following prior works (Liu et al., 2024b; Zou et al., 2025; Sun et al., 2025), we conduct the evaluation on both the MSCOCO dataset (Lin et al., 2014) and the A-OKVQA dataset (Schwenk et al., 2022), using 500 randomly sampled images for each dataset, with multiple object queries. Similar to the official POPE protocol, negative queries are constructed using three sampling strategies that vary in difficulty; random, which samples objects uniformly from the object vocabulary; popular, which samples from frequently occurring objects in the dataset; and adversarial, which selects objects that commonly co-occur with the ground-truth objects in similar scenes, making them more challenging negatives. For each query, the model predicts whether the object is present in the image, and performance is evaluated using accuracy and F1 score. We report results separately for each sampling strategy, as well as the average across the three settings. To reduce ambiguity in free-form generation, we append the instruction "Please answer yes or no." to each prompt. We use a maximum generation length of 20 tokens. Final predictions are obtained following the standard POPE evaluation protocol.

CHAIR (Rohrbach et al., 2018) Caption Hallucination Assessment with Image Relevance (CHAIR), evaluates object hallucination in image captioning models by measuring whether generated captions mention objects that are not present in the image. Following prior work (Liu et al., 2024b; Zou et al., 2025; Sun et al., 2025), we conduct the evaluation on the MSCOCO (Lin et al., 2014) 2014 validation split and randomly sample 500 images for evaluation. For each image, the model is prompted to generate a descriptive caption using a generic instruction of the form "Please describe this image in detail.", following prior work as well. The generated captions are then analyzed using the CHAIR evaluation procedure. Ground-truth objects are obtained from MSCOCO instance annotations and reference captions, allowing the evaluator to determine which objects are truly present in the scene. CHAIR reports per-instance evaluation CHAIR_I and per-sentence evaluation CHAIR_S, defined as follows:

$$\text{CHAIR}_I = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all objects mentioned}\}|}$$

$$\text{CHAIR}_S = \frac{|\{\text{sentences with hallucinated object}\}|}{|\{\text{all sentences}\}|}$$

We use a maximum generation length of 150 tokens.

HallusionBench (Guan et al., 2024) HallusionBench evaluates hallucination in diverse image-context reasoning scenarios, including OCR, charts, tables, mathematical figures, maps, and visual illusions. Following the official evaluation protocol, we report figure-level accuracy (fACC), question-level accuracy (qACC), and accuracy on the easy (easyA) and hard (hardA) subsets.

AIR-Bench (Yang et al., 2024) The AIR-Bench Foundation benchmark consists of 19 single-task abilities spanning three audio domains: speech, natural sounds, and music, totaling over 19k single-choice questions. Although the questions are presented in a multiple-choice format, models are required to generate answers freely, making the evaluation fully generative. We follow the official AIR-Bench evaluation protocol and use the GPT-4-based evaluator to determine correctness. The evaluator compares the generated hypothesis with the reference answer given the question and associated meta-information, and outputs a binary judgment. We report accuracy, averaged per domain (Speech, Sound, Music). We use a maximum generation length of 50 tokens.

Audio Hallucination QA (Kuan et al., 2024) The Audio Hallucination QA benchmark evaluates object hallucination in large audio-language models by formulating the task as a binary discriminative question-answering problem. Given an audio clip, models are asked yes/no questions of the form "Is there a sound of [object] in the audio?". Positive questions are constructed from ground-truth sound labels, while negative questions are generated using three sampling strategies; random, popular, and adversarial, designed to probe the model's susceptibility to hallucinating non-existent objects. In our evaluation, we report accuracy and F1 score under each of the three sampling strategies. These metrics directly measure the model's tendency to over-predict object presence and provide a focused assessment of hallucination behavior under discriminative settings. To reduce ambiguity in generation, we append the instruction "Please answer yes or no." to each prompt and use a maximum generation length of 20 tokens.

B.2 Baselines and Implementation Details

We provide additional implementation details for the proposed method. Unless stated otherwise, all hyperparameters are reported in Table A.1. At each decoding step, we optimize the key and value updates using a small number of gradient-based updates, following the formulation in Section 3.3. We use the Adam optimizer (Kingma & Ba, 2014) for all inference-time optimization steps. To improve stability and reduce the number of optimization variables, the updates are shared across all attention heads within each layer. After each decoding step, the perturbations are reset and re-initialized for the next token.

B.3 Additional Experiments and Results

We provide additional experimental results and qualitative examples that complement the main findings presented in Section 4. These include extended evaluations across models and benchmarks, as

Table A.1: Hyperparameters of the proposed method for each evaluated model.

Model	Optimization Steps	Learning Rate	KL Weight (λ)	Temperature (τ)
LLaVA-1.5-7B	7	3×10^{-4}	0.1	0.1
Qwen-VL-Chat	7	4×10^{-4}	0.1	0.1
Qwen2.5-VL-7B-Instruct	7	3×10^{-4}	0.1	0.1
SALMONN-7B	7	3×10^{-4}	0.1	0.1
Qwen2-Audio-7B-Instruct	7	5×10^{-4}	7×10^{-3}	0.1
Qwen2.5-Omni-7B	7	4×10^{-4}	0.1	0.1
MiniCPM-o 4.5	7	5×10^{-4}	0.1	0.1

well as visualizations illustrating the effect of our method on multimodal grounding. Together, these results further support the effectiveness of the proposed approach.

Table A.2: Evaluation results on the POPE benchmark with Qwen-VL-Chat and Qwen2.5-VL-7B Instruct.

Dataset	Methods	Random (%) ↑		Popular (%) ↑		Adversarial (%) ↑		Average (%) ↑	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1
MSCOCO	Qwen-VL-Chat	84.73	82.67	84.13	82.06	82.26	80.37	83.7	81.7
	+ VCD	88.63	87.81	87.12	86.4	84.26	83.9	86.67	86.03
	+ LIME (ours)	**89.1**	**88.42**	**88.12**	**87.89**	**84.55**	**84.78**	**87.25**	**87.03**
	Qwen2.5-VL-7B-Instruct	86.6	84.48	85.2	83.12	84.23	82.16	85.34	83.25
	+ LIME (ours)	**88.11**	**86.32**	**85.77**	**84.1**	**85.33**	**83.88**	**86.4**	**84.76**
A-OKVQA	Qwen-VL-Chat	86.67	85.59	85.56	84.63	79.57	79.5	83.93	83.24
	+ VCD	**89.22**	**89.01**	87.85	87.81	81.27	82.38	86.11	**86.4**
	+ LIME (ours)	87.8	87.21	**87.95**	**87.95**	**83.21**	**83.36**	**86.32**	86.17
	Qwen2.5-VL-7B-Instruct	88.53	87.23	86.12	85.07	85.05	84.14	86.56	85.48
	+ LIME (ours)	**89.97**	**88.73**	**87.22**	**86.52**	**86.22**	**86**	**87.8**	**87.08**

Table A.3: POPE benchmark results using LLaVA-1.5-7B on the A-OKVQA dataset. We report Accuracy and F1 for the Random, Popular, and Adversarial splits, as well as their average. Best and second-best results are highlighted in **bold** and underlined, respectively.

Dataset	Methods	Random (%) ↑		Popular (%) ↑		Adversarial (%) ↑		Average (%) ↑	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1
A-OKVQA	LLaVA-1.5-7B	83.45	82.56	79.9	79.59	74.04	75.15	79.13	79.1
	+ OPERA	88.27	87.54	85.17	84.74	79.37	79.97	84.27	84.08
	+ VCD	86.15	86.34	81.85	82.82	74.97	77.73	80.99	82.3
	+ ICD	85.57	85.06	81.93	81.95	77.43	78.99	81.64	82
	+ MemVR	91.1	90.83	87.33	87.43	80.20	81.66	86.21	86.64
	+ V-ITI	91.51	91.04	87.47	86.97	80.34	81.12	86.44	86.37
	+ LIME (ours)	**91.71**	**91.12**	**88.75**	**87.93**	**81.55**	**82.4**	**87.33**	**87.15**

Table A.4: Evaluation results on the random splits of POPE, using the MSCOCO and A-OKVQA datasets, and Audio Hallucination QA. Accuracy and F1 scores are reported for all settings.

Methods	POPE (%) ↑				AHQA (%) ↑	
	MSCOCO		A-OKVQA			
	Acc	F1	Acc	F1	Acc	F1
Qwen2.5-Omni-7B	87.7	86	88.1	87.9	89.25	89.34
+ LIME (ours)	**89.5**	**88.12**	**89.5**	**89.3**	**90.5**	**90.33**

Table A.5: Evaluation results on the random splits using the MSCOCO and A-OKVQA datasets. Accuracy and F1 scores are reported for all settings.

Method	MSCOCO		A-OKVQA	
	Acc.	F1	Acc.	F1
MiniCPM-o 4.5	90.4	89.5	92.2	91.7
+ LIME (ours)	**92.2**	**91.9**	**93.85**	**93.6**

Table A.6: Evaluation results on HallusionBench for LLaVA-1.5 and Qwen2.5-VL-7B-Instruct. Best and second-best results are highlighted in **bold** and underlined, respectively.

Methods	**fACC ↑**	**qACC ↑**	**easyA ↑**	**hardA ↑**
LLaVA-1.5	17.9	8.13	36	36.7
+ OPERA	16.2	5.49	37.6	35.4
+ VCD	13.9	11.4	33	34.7
+ ICD	13.9	8.35	36.9	33.5
+ MemVR	17.9	9.01	36.9	**37.7**
+ V-ITI	17.9	**10.27**	36.6	37
+ LIME (ours)	**19.24**	9.85	**37.62**	**37.7**
Qwen2.5-VL-7B-Instruct	19.29	16.4	46.11	37.9
+ LIME (ours)	**21.06**	**18.61**	**50.37**	**39.1**

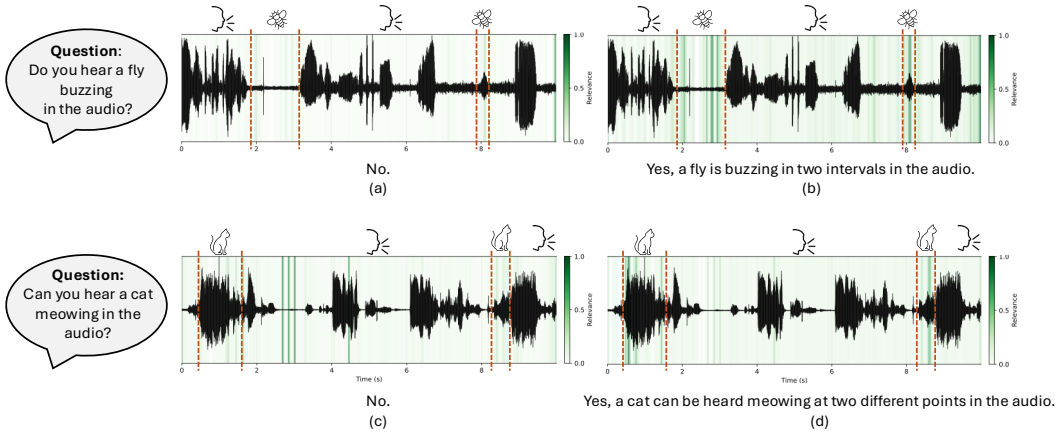


Figure 5: Qualitative examples of multimodal hallucination in the audio domain and its mitigation with LIME. The audio waveform is overlaid with relevance scores (green), and dashed vertical lines indicate ground-truth event regions. Under standard decoding (a,c) relevance is diffuse and poorly aligned with relevant audio segments, leading to incorrect predictions. With LIME (b,d), relevance concentrates on the correct temporal regions, resulting in grounded and accurate predictions.

Input



Describe the image in detail.

LLaVA-1.5-7B: The image features a woman standing on a sidewalk, flying a kite in a park. She is holding the kite string, and the kite is soaring in the sky. The park is filled with various potted plants, creating a lively and colorful atmosphere. *In addition to the woman and the kite, there are several other people scattered throughout the scene, possibly enjoying the park or engaging in other activities.* The overall scene is a pleasant and active outdoor setting.

LLaVA-1.5-7B + LIME: The image features a woman standing on a sidewalk, flying a kite. The kite is soaring high in the sky, and the woman appears to be enjoying the activity. The scene is set in a park, with a potted plant visible in the background. The woman is positioned near the center of the image, and the kite can be seen in the upper part of the scene.

Input

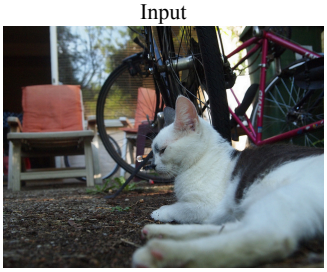


Describe the image in detail.

LLaVA-1.5-7B: The image depicts a group of people gathered in a living room, playing a video game together. There are four people in the scene, with two of them standing closer to the left side of the room, and the other two standing more towards the right side. The living room is furnished with a couch *and a chair*, both located in the middle of the room. *A TV is mounted on the wall*, and a remote control can be seen in the hands of one of the players.

LLaVA-1.5-7B + LIME: The image depicts a group of people standing in a living room, playing a video game together. There are four people in the scene, with two of them standing closer to the left side of the room, and the other two standing closer to the right side. They are all holding Wii remotes, which are white in color. The living room is furnished with a couch located in the middle of the room. There is also a bottle visible on the right side of the room, possibly a beverage for the players to enjoy during their gaming session.

Input



Describe the image in detail.

LLaVA-1.5-7B: The image features a white and black cat lying on the ground, possibly on a patio. The cat is positioned near a bicycle, which is parked in the background. There are also two chairs in the scene, one located closer to the cat and the other further away. *In addition to the cat and bicycle, there are two people in the image. One person is standing near the left side of the scene, while the other person is standing on the right side.*

LLaVA-1.5-7B + LIME: The image features a white and black cat lying on the ground, appearing to be sleeping or relaxed. The cat is positioned in the center of the scene. In the background, there is a bicycle parked, and an orange chair is placed nearby. The chair is located on the left side of the image, while the bicycle is situated on the right side.

Figure 6: Qualitative examples of hallucination reduction with LIME on LlaVA-1.5-7B. Hallucinated predictions are indicated in *italics* and highlighted in red.

Input



Describe the image in detail.

Qwen-VL-Chat: The image features a man sitting on the ground *in front of a table, which has various fruits on it*, including several bunches of bananas. The man is sitting near a basket of bananas, and there are multiple apples displayed in other locations around him. A child can be seen standing in the background, near the right side of the scene. There are several other people present in the image, some of them closer to the man and others scattered throughout the scene.

Qwen-VL-Chat + LIME: The image features a man and a child sitting on the ground, surrounded by several baskets of fruit, specifically bananas. There are multiple apples in the scene, with some placed near the man and others spread throughout the area. The bananas are stored in baskets, with two main ones and a few smaller ones. In addition to the man, the child can be seen in the background, and they both appear to be selling the fruit as they are sitting among several baskets of fruit for sale. The scene takes place in an outdoor market or a street market.

Input



Describe the image in detail.

Qwen-VL-Chat: The image features a clean and well-lit bathroom with a vanity, toilet, and a large window. The bathroom has a blue counter, providing a nice contrast to the white walls. The window, located above the counter, allows natural light to fill the space and offers a pleasant view for those using the bathroom. There are several small objects on the counter, including *three teddy bears* and a flower, adding a touch of personality to the space. *A shower can be seen in the background, indicating that this bathroom is part of a larger living area.*

Qwen-VL-Chat + LIME: The image features a bathroom with a large window, letting in plenty of natural light. There is a blue countertop and a teddy bear placed on the counter. In the background, a toilet is clearly visible, along with a potted plant that adds a touch of liveliness to the space. The bathroom appears to be well-decorated and clean, providing a comfortable atmosphere.

Input



Describe the image in detail.

Qwen-VL-Chat: The image features a red stop sign that is covered in graffiti, with the words "driving" written beneath it. The stop sign is located at the intersection of Virginia Street and 1360, with power lines running above it. There are several houses in the background, *as well as a couple of cars parked nearby*. The graffiti on the stop sign adds an element of urban art to the scene.

Qwen-VL-Chat + LIME: The image features a stop sign that is painted bright red with the words "STOP DRIVING" written on it in white paint. The stop sign is located at an intersection, with a street sign attached to the top of it. The street sign points in two directions, one at the left and the other to the right. There are several power lines visible in the image, with one set of power lines clearly defined and another set less prominent set above the stop sign. The contrast between the red stop sign and the graffiti message it bears makes this scene an interesting and unusual sight.

Figure 7: Qualitative examples of hallucination reduction with LIME on Qwen-VL-Chat. Hallucinated predictions are indicated in *italics* and highlighted in red.

Input



Describe the image in detail.

Qwen2.5-VL-7B-Instruct: The image shows a small, tiled restroom stall with a modern design. The focal point is a toilet with an open lid and seat, positioned centrally in the frame. The toilet appears to be a wall-mounted model, as there there's no visible base or tank. Above the toilet, there is a white flush panel with buttons for flushing. To the right of the toilet, there is a toilet paper holder mounted on the wall, *and below it, a small trash bin is also visible.* The walls and floor of the stall are are covered in light-colored tiles, giving the space space a clean and minimalistic appearance.

Qwen2.5-VL-7B-Instruct + LIME: The image shows an overhead view of a toilet situated within a small, tiled bathroom stall. The walls and floor appear to be covered with white tiles. The toilet is a standard design with a raised seat that is currently open. There's a flush mechanism mounted on the wall above the toilet tank, which includes buttons or levers for flushing.

Input



Describe the image in detail.

Qwen2.5-VL-7B-Instruct: The image shows a fire hydrant situated on a grassy area beside a sidewalk, with a traffic cone placed in front of it. The cone is brightly colored with reflective stripes and a pink base. The scene appears to be edited or filtered with a color effect that gives everything a purple hue, giving it an artistic or surreal appearance. In the background, there is a quiet street with a few houses, trees, *and a parked motorcycle visible.* The overall tone of the image is somewhat dreamlike due to the color manipulation.

Qwen2.5-VL-7B-Instruct + LIME: The image depicts an urban street scene with a focus on a fire hydrant and a traffic cone placed near it. The photo has been edited to give a surreal, almost dreamlike quality. The entire image is tinted with a purple hue, giving it a slightly unnatural appearance. The fire hydrant appears silver and metallic, while the traffic cone is pink and white. The background includes houses, trees, a car and a road, which are all appear normal in color contrast, creating the foreground elements stand more prominent.

Input



Describe the image in detail.

Qwen2.5-VL-7B-Instruct: The image depicts a baseball game in progress, with a pitcher mid-action on the mound. The pitcher is wearing a maroon jersey, white pants with maroon stripes, and a white cap. He appears to be be in the middle of delivering a pitch, with his right leg extended forward and his left leg lifted behind him for balance. *The ball is visible in the air above his head, indicating that he it has just been released.*

Qwen2.5-VL-7B-Instruct + LIME: The image depicts an outdoor baseball game on what appears to be a sunny day, as indicated by the shadows and lighting. The scene is set at a small stadium or park with spectators seated along bleachers in the background. At the right side of the image, the stands are filled with people watching, some wearing casual clothing while others appear dressed for sports. In the foreground, the center of the image, a baseball player is captured mid-action during his pitch. The pitcher is wearing a maroon jersey, white pants with red stripes, and black cleats.

Figure 8: Qualitative examples of hallucination reduction with LIME on Qwen2.5-VL-7B-Instruct. Hallucinated predictions are indicated in *italics* and highlighted in red.

Table A.7: Per-sample relevance shift under LIME relative to vanilla decoding, grouped by prediction outcome (LLaVA-1.5-7B, POPE). Mean ± SD are reported.

Outcome group	n	δ Perceptual Utilization	δ Relevance Localization
correct → hallucinated	2	$+0.019 \pm 0.033$	-0.008 ± 0.041
hallucinated → hallucinated	11	$+0.031 \pm 0.045$	$+0.026 \pm 0.058$
correct → correct	64	$+0.058 \pm 0.041$	$+0.071 \pm 0.049$
hallucinated → correct	23	$\mathbf{+0.121 \pm 0.038}$	$\mathbf{+0.147 \pm 0.052}$

C Analysis of the Relevance Objective

The relevance objective in Eq. (3) is defined over the aggregate relevance assigned to perceptual and textual tokens, without supervision over which perceptual tokens should gain relevance and without access to prediction correctness. As noted in Section 3.3, this is a deliberate trade-off for remaining training- and external-model-free: LIME requires no auxiliary verifier, localization module, or labeled data, and can be applied to an MLLM directly at inference time.

This design raises two questions. First, whether the relevance shifts induced by LIME are actually related to changes in hallucination behavior. Appendix C.1 addresses this through per-sample analysis and controlled interventions on the target and direction of the relevance objective. Second, given that the objective treats all perceptual tokens symmetrically, why does the resulting relevance become concentrated on query-relevant regions? Appendix C.2 provides a mechanistic account of how this redistribution is constrained.

C.1 Interventional Analysis of the Relevance-Hallucination Link

Section 4.3 shows that LIME increases perceptual utilization and concentrates relevance on query-relevant regions. We next examine whether these relevance shifts are linked to prediction behavior rather than merely accompanying improved performance. We perform two complementary analyses. First, we test whether samples whose predictions are corrected exhibit larger relevance shifts than those that remain hallucinated. Second, we vary which tokens the relevance objective targets and whether their relevance is increased or decreased, while keeping the remaining optimization settings fixed.

Per-sample relevance shift and prediction outcome. We use the same 100 POPE samples as in Section 4.3 and compute, for each sample, the change δ in perceptual utilization $\Phi_{\mathcal{P}}/(\Phi_{\mathcal{P}}+\Phi_{\mathcal{T}})$ and in relevance localization $\sum_{i\in\mathcal{G}} \Phi_i/\Phi_{\mathcal{P}}$ between vanilla decoding and LIME. Samples are then partitioned by prediction outcome under the two decoding schemes. Results are reported in Table A.7, where we give the mean and standard deviation within each group. Corrected predictions exhibit substantially larger relevance shifts than those that remain hallucinated: $+0.121$ versus $+0.031$ in perceptual utilization and $+0.147$ versus $+0.026$ in relevance localization. Samples that were already correct show intermediate shifts, consistent with the intervention reinforcing grounding that the model had partially established. The magnitude of the shift therefore tracks the outcome of individual predictions rather than reflecting only a co-movement of two group averages. We do not interpret the correct → hallucinated group, which contains two samples, we report it to show that LIME rarely degrades predictions that were already correct.

Sweeping the target and direction of the relevance objective. The relevance objective in Eq. (3) is defined by a positive set ($\mathcal{P}$) and by a negative sign ($-$). We vary both while holding everything else fixed: the KL regularizer and its weight, the layer scope, the optimization budget, the perturbation family, the learning rate, and the temperature are identical across all variants, so the configurations differ only in which tokens the objective pushes relevance toward, and in which direction. We evaluate five variants:

- **LIME**: the method as proposed,

$$\mathcal{L}_{rel}(\Delta) = -\frac{1}{|\mathcal{P}|}\sum_{i\in\mathcal{P}} \log \frac{\exp(\Phi_{i,\Delta}/\tau)}{\sum_{k\in\mathcal{P}\cup\mathcal{T}} \exp(\Phi_{k,\Delta}/\tau)}$$

- **Reverse-LIME**: the LIME objective with the sign flipped,

$$\bar{\mathcal{L}}_{rel}(\Delta) = -\mathcal{L}_{rel}(\Delta)$$

- **Random**: the same objective with the outer sum taken over a random subset $\mathcal{R} \subset \mathcal{P} \cup \mathcal{T}$ with $|\mathcal{R}| = |\mathcal{P}|$, resampled at each decoding step,

$$\mathcal{L}_{rel}^{rand}(\Delta) = -\frac{1}{|\mathcal{R}|} \sum_{i \in \mathcal{R}} \log \frac{\exp(\Phi_{i,\Delta}/\tau)}{\sum_{k \in \mathcal{P} \cup \mathcal{T}} \exp(\Phi_{k,\Delta}/\tau)}$$

- **LIME-Sup**: the supervised variant introduced in Eq. (2), in which the outer sum is taken only over the ground-truth query-relevant region $\mathcal{G}$,

$$\mathcal{L}_{rel}^{sup}(\Delta) = -\frac{1}{|\mathcal{G}|} \sum_{i \in \mathcal{P}} 1_{\{i \in \mathcal{G}\}} \log \frac{\exp(\Phi_{i,\Delta}/\tau)}{\sum_{k \in \mathcal{P} \cup \mathcal{T}} \exp(\Phi_{k,\Delta}/\tau)}.$$

- **Reverse-LIME-Sup**: LIME-Sup with the sign flipped,

$$\bar{\mathcal{L}}_{rel}^{sup}(\Delta) = -\mathcal{L}_{rel}^{sup}(\Delta)$$

As in Section 4.3, $\mathcal{G}$ is available for this analysis only, which allows us to use it as an intervention target here; as stated in Section 3.3, LIME itself does not assume this prior knowledge. Since the softmax in Eq. (3) normalizes over $\mathcal{P} \cup \mathcal{T}$, the resulting relevance shares sum to one, so increasing the share assigned to $\mathcal{P}$ necessarily decreases the share assigned to $\mathcal{T}$. Amplifying $\mathcal{P}$ and suppressing $\mathcal{T}$ are therefore near-equivalent interventions, and we report one variant per pair. We evaluate on the positive split of POPE-MSCOCO with LLaVA-1.5-7B, since $\mathcal{G}$ is defined only for objects that are present, using 5 optimization steps and Δ on all layers.

Table A.8 gives the ordering: Reverse-LIME-Sup $<$ Reverse-LIME $<$ Baseline $\approx$ Random $<$ LIME $<$ LIME-Sup. Random perturbs with comparable magnitude under an identical objective form yet performs at baseline (83.35 vs. 83.24), so the identity of the amplified tokens, not the presence of an inference-time perturbation, produces the gain. The ordering is signed and monotone about the baseline: amplifying perceptual relevance helps and suppressing it degrades accuracy below the unmodified model. Finally, the effect is asymmetric in $\mathcal{G}$: suppressing relevance on the query-relevant region alone is more damaging (-3.42) than suppressing it across all perceptual tokens (-0.59), and LIME-Sup improves on LIME ($+1.68$), indicating that the relevance that determines the prediction resides on the query-relevant tokens, rather than being spread across perceptual tokens in general. As in Section 4.3, LIME-Sup bounds what the label-free objective could achieve with perfect knowledge of the region, and the account in Appendix C.2 is consistent with our objective recovering much of that gain without access to $\mathcal{G}$. Together, these results show that the reduction in hallucination follows from which tokens gain relevance and which lose it, under otherwise identical optimization.

Table A.8: Sweep over the target and direction of the relevance objective (LLaVA-1.5-7B, positive split of POPE-MSCOCO). All arms share the KL regularizer, layer scope, optimization budget, and remaining hyperparameters, and differ only in the positive set and the sign of the objective.

Method	**Objective target**	**Direction**	**Acc (%)** $\uparrow$	δ **Acc**	**F1** $\uparrow$
Baseline	—	—	83.24	0.00	83.12
Random	random, $\lvert\mathcal{P}\rvert$	amplify	83.35	+0.11	83.20
LIME	$\mathcal{P}$	amplify	90.44	+7.20	89.88
Reverse-LIME	$\mathcal{P}$	suppress	82.65	-0.59	82.14
Reverse-LIME-Sup	$\mathcal{G}$	suppress	79.82	-3.42	79.45
LIME-Sup	$\mathcal{G}$	amplify	92.12	+8.88	91.74

C.2 Mechanism Behind Relevance Localization

Having established that the relevance shift is associated with prediction behavior, we next examine why LIME concentrates relevance on query-relevant perceptual regions despite receiving no spatial

or temporal supervision. Section 4.3 shows that LIME recovers much of the relevance-localization improvement achieved by LIME-Sup, which explicitly targets the annotated region. This behavior is not directly implied by Eq. (3), since the objective treats all perceptual tokens symmetrically and does not identify a spatial or temporal target.

We develop the explanation in three steps. First, we show that the optimization primarily redistributes relevance and characterize the uniform distribution favored by $\mathcal{L}_{rel}$. Second, we derive the local geometry imposed by the KL constraint. Finally, we show how the frozen model filters the objective's broadly distributed direction, concentrating the realized change along existing perceptual pathways.

Relevance is redistributed, not created. LRP conservation (Eq. (8)) ties the total input relevance to the logit being explained, $\Phi_{\mathcal{P}} + \Phi_{\mathcal{T}} \approx z_t^*$, while the KL term (Eq. (4)) discourages changes to the next-token distribution. Together, these constraints limit changes in the total relevance, so the optimization acts primarily by redistributing relevance among tokens: increasing Φ_i for some tokens generally requires decreasing it for others. Perceptual utilization measures how this relevance is divided between $\mathcal{P}$ and $\mathcal{T}$, whereas relevance localization measures how the perceptual relevance is distributed within $\mathcal{P}$, specifically how much falls inside the query-relevant subset $\mathcal{G}$.

The objective favors a uniform spread. To understand the intrinsic preference of $\mathcal{L}_{rel}$, we first analyze it directly as a function of the relevance scores Φ, abstracting away from how these scores are produced by the model. Let $\mathbf{s} = \mathrm{softmax}(\Phi/\tau)$ denote the temperature-scaled relevance distribution, and let

$$u_i = \begin{cases} \frac{1}{|\mathcal{P}|} & i \in \mathcal{P}, \\ 0 & i \in \mathcal{T}, \end{cases}$$

denote the distribution that places equal mass on all perceptual tokens and none on textual tokens. Differentiating Eq. (3) with respect to Φ gives

$$\mathbf{g}_\Phi = -\nabla_\Phi \mathcal{L}_{rel} \;=\; \tfrac{1}{\tau}\,(\mathbf{u} - \mathbf{s}), \qquad \text{so that} \qquad (g_\Phi)_i = \begin{cases} \frac{1}{\tau}\left(\frac{1}{|\mathcal{P}|} - s_i\right) & i \in \mathcal{P}, \\ -\frac{1}{\tau} s_i & i \in \mathcal{T}. \end{cases} \tag{10}$$

The sign and magnitude of each entry in $\mathbf{g}_\Phi$ indicate how the loss favors changing the relevance of the corresponding token. For textual tokens, $(g_\Phi)_i < 0$, so the loss favors decreasing their relevance. For perceptual tokens, relevance is favored to increase when the current share s_i is below $\frac{1}{|\mathcal{P}|}$, and decreased when it is above this value. Thus, within $\mathcal{P}$, the objective favors equalizing the relevance shares rather than concentrating them on a particular subset.

Remark 1. *If the relevance scores were optimized directly under $\mathcal{L}_{rel}$, a stationary point would satisfy $\mathbf{s} = \mathbf{u}$, corresponding to equal relevance share across perceptual tokens and zero share on textual tokens. At this point, the fraction of perceptual relevance inside $\mathcal{G}$ would be $\frac{|\mathcal{G}|}{|\mathcal{P}|}$. Therefore, considered on its own, $\mathcal{L}_{rel}$ favors a uniform distribution over $\mathcal{P}$ and does not directly encourage localization to the query-relevant subset $\mathcal{G}$. This mirrors a known pattern in contrastive learning, where the objective alone does not determine what is learned and the architecture and optimizer form part of the account (Saunshi et al., 2022).*

The role of the KL term. In LIME, the relevance is not optimized directly. Instead, the optimization variables are the key-value perturbations Δ (Section 3.3), viewed here as a single vector containing the perturbations across layers. Consequently, the desired change in relevance, represented by $\mathbf{g}_\Phi$, cannot be applied directly. It can only be induced by changing Δ, and propagating its effect through the frozen model, while the KL term penalizes the resulting change in the next-token distribution. We characterize this penalty locally around $\Delta = 0$, which is the natural expansion point because LIME reinitializes the perturbations to zero at every decoding step (Section 3.3).

Lemma 1. *Around $\Delta = 0$*

$$\mathcal{L}_{KL}(\Delta) = \tfrac{1}{2}\Delta^\top \mathbf{F} \Delta + O(\|\Delta\|^3), \tag{11}$$

where $\mathbf{F} \in \mathbb{R}^{D\times D}$ is the Fisher information of the next-token distribution with respect to Δ, evaluated at $\Delta = 0$, and D is the total number of scalar elements in the key-value perturbation vector Δ.

$$\mathbf{F} = \mathbb{E}_{y\sim p_\theta}[\nabla_\Delta \log p_{\theta,\Delta}\, \nabla_\Delta \log p_{\theta,\Delta}^\top]|_{\Delta=0}. \tag{12}$$

Proof. We expand around $\Delta = 0$, where $p_{\theta,\Delta} = p_\theta$. At this point, $\mathcal{L}_{KL}(0) = 0$. Its gradient also vanishes:

$$\nabla_\Delta \mathcal{L}_{KL}|_{\Delta=0} = \nabla_\Delta \sum_y p_{\theta,\Delta}(y) = \nabla_\Delta 1 = 0.$$

Differentiating once more and using

$$\nabla_\Delta p_{\theta,\Delta} = p_{\theta,\Delta} \nabla_\Delta \log p_{\theta,\Delta},$$

the terms involving

$$\sum_y \nabla^2_\Delta p_{\theta,\Delta}(y) = \nabla^2_\Delta 1 = 0$$

vanish. The remaining term is

$$\mathbb{E}_{y\sim p_\theta}[\nabla_\Delta \log p_{\theta,\Delta} \, \nabla_\Delta \log p^\top_{\theta,\Delta}]|_{\Delta=0} = \mathbf{F}.$$

The result follows from the second-order Taylor expansion of $\mathcal{L}_{KL}$ around $\Delta = 0$ □

Since both the value and the gradient vanish at $\Delta = 0$, the KL term penalizes the perturbation quadratically in the Fisher metric rather than linearly. The Fisher matrix assigns a higher cost to perturbation directions that more strongly change the next-token distribution.

One optimization step. Since relevance depends on the KV perturbation through the frozen model, we write it as $\Phi(\Delta)$. Let $\mathbf{J} = \frac{\partial \Phi(\Delta)}{\partial \Delta}\Big|_{\Delta=0} \in \mathbb{R}^{N\times D}$ be its Jacobian at $\Delta = 0$, where $N = |\mathcal{P}| + |\mathcal{T}|$ is the number of input tokens. Locally, $\Phi(\Delta) \approx \Phi(0) + \mathbf{J}\Delta$. Using $\mathbf{g}_\Phi = -\nabla_\Phi \mathcal{L}_{rel}$, the corresponding first-order approximation of the relevance loss is

$$\mathcal{L}_{rel}(\Phi(\Delta)) \approx \mathcal{L}_{rel}(\Phi(0)) - \mathbf{g}_\Phi^\top \mathbf{J}\Delta.$$

Combining this approximation with Lemma 1 and dropping terms independent of Δ gives

$$\min_\Delta \left\{ -\mathbf{g}_\Phi^\top \mathbf{J}\Delta + \tfrac{\lambda}{2}\Delta^\top \mathbf{F}\Delta \right\}.$$

On the support of $\mathbf{F}$, its minimum-norm solution is

$$\Delta^\star = \tfrac{1}{\lambda}\mathbf{F}^+ \mathbf{J}^\top \mathbf{g}_\Phi,$$

where $\mathbf{F}^+$ denotes the pseudoinverse of $\mathbf{F}$. Therefore, the induced change in relevance is

$$\delta\Phi = \Phi(\Delta^\star) - \Phi(0) \approx \mathbf{J}\Delta^\star = \tfrac{1}{\lambda} \underbrace{\mathbf{J}\mathbf{F}^+\mathbf{J}^\top}_{\mathbf{S}} \mathbf{g}_\Phi = \tfrac{1}{\lambda}\mathbf{S}\mathbf{g}_\Phi. \tag{13}$$

Thus, the realized relevance change is obtained by filtering the direction favored by the relevance loss, $\mathbf{g}_\Phi$, through the model-dependent matrix $\mathbf{S}$.

Why the update concentrates. Eq. (13) shows that the realized relevance change is not $\mathbf{g}_\Phi$ itself, but $\mathbf{S}\mathbf{g}_\Phi$. Although $\mathbf{g}_\Phi$ distributes the desired increase broadly across perceptual tokens, $\mathbf{S}$ restricts this change according to the sensitivity structure of the frozen model. In particular

$$S_{ij} = \mathbf{J}_{i,:}\mathbf{F}^+\mathbf{J}^\top_{j,:},$$

where $\mathbf{J}_{i,:}$ describes how the relevance of token i responds to changes in Δ. Thus, tokens tend to undergo similar relevance changes when they are sensitive to the same perturbation directions allowed by the KL constraint. The broad direction favored by the relevance loss is therefore concentrated along relevance patterns to which the frozen model is already sensitive. The results in Section 4.3 provide empirical evidence that the directions favored by the frozen model align with the query-relevant subset $\mathcal{G}$, as the realized relevance increase concentrates within $\mathcal{G}$ despite the objective having no access to this subset.

D Efficiency Analysis and Limitations

Section 4.4 examines LIME's main design choices, including the roles of key and value updates and the effect of the KL-regularization weight λ. Here, we extend this analysis to computational efficiency by varying the optimization budget and intervention scope, and by measuring latency and memory overhead. We conclude by discussing the practical limitations of the method.

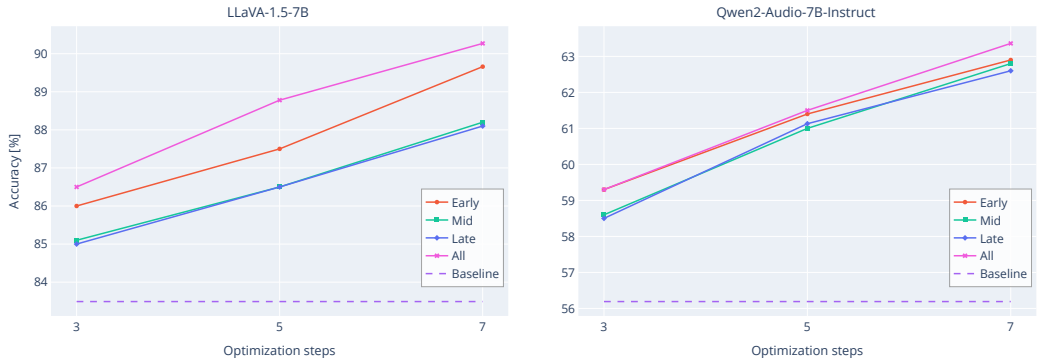


Figure 9: Effect of optimization steps when applying Δ to early (red), middle (green), late (blue), or all (pink) transformer layers.

Table A.9: Computational overhead comparison between vanilla decoding, representative baselines, and LIME across optimization budgets.

Methods	**Tokens/sec** $\uparrow$	**Slowdown** $\downarrow$	**Peak Memory (GB)** $\downarrow$	**Accuracy (%)** $\uparrow$
LLaVA-1.5-7B	3.02	1.0×	26.24	83.49
+ OPERA	0.82	3.66×	39.09	87.53
+ MemVR	2.9	1.04×	26.51	88.5
+ LIME (5 steps)	0.45	6.73×	34.75	88.78
+ LIME (7 steps)	0.32	9.43×	34.75	90.27
Qwen2-Audio-7B	2.76	1.0×	30.84	56.19
+ AAD	1.38	2.0×	32.7	59.5
+ LIME (5 steps)	0.42	6.57×	33.84	61.5
+ LIME (7 steps)	0.3	9.2×	33.84	63.36

D.1 Efficiency Analysis

Layer scope and optimization budget. We investigate whether the Δ updates must be applied throughout the entire LLM backbone. We consider four intervention scopes: early layers (0-8), middle layers (12-20), late layers (24-32), and all transformer layers. For each scope, we perform three, five, or seven optimization steps. The results are shown in Figure 9. Increasing the optimization budget consistently improves accuracy across both models and all intervention scopes. Applying the updates to all layers achieves the strongest performance at every budget. Nevertheless, three optimization steps already provide substantial improvements over the base models, and interventions restricted to selected layer ranges remain effective. For LLaVA-1.5-7B, early-layer updates consistently outperform middle- and late-layer updates, whereas the differences among the restricted scopes are smaller for Qwen2-Audio-7B-Instruct. Overall, these results show that the optimization budget and layer scope provide controllable lower-cost operating points, although the highest accuracy is obtained by applying Δ throughout the network.

Computational overhead. We additionally report the computational overhead introduced by LIME relative to vanilla decoding and to several representative decoding methods. Table A.9 summarizes throughput, relative slowdown, and peak GPU memory usage measured under the same decoding configuration used in the main experiments. For LIME, we evaluate five and seven optimization steps with Δ applied to all transformer layers, corresponding to the strongest intervention scope in the preceding analysis. LIME's overhead increases with the optimization budget: for LLaVA-1.5-7B, slowdown increases from $6.73\times$ at five steps to $9.43\times$ at seven step, with a similar trend for Qwen2-Audio-7B. In contrast, peak memory remains essentially unchanged across optimization budgets, indicating that the additional cost is primarily in latency rather than memory. The five-step configuration provides a lower-cost operating point, reaching 88.78% accuracy on LLaVA-1.5-7B compared with 90.27% at seven steps, and 61.5% on Qwen2-Audio-7B compared with 63.36%.

Unlike the representative baselines, which operate with a fixed inference configuration, LIME exposes an explicit optimization budget that allows computational cost to be adjusted: increasing the number of optimization steps improves performance at the cost of higher latency, while fewer steps provide a lower-cost operating point. Together with the layer-scope results above, this makes the efficiency–performance trade-off explicitly controllable.

D.2 LIMITATIONS

LIME has several practical limitations. First, although it requires no training, its per-step relevance-guided optimization introduces non-negligible inference-time latency. As shown in Appendix D.1, this cost can be reduced by decreasing the optimization budget or restricting the intervention scope, but at the expense of some performance. Second, the method requires the selection of inference-time hyperparameters, including the KL-regularization weight and optimization budget, whose optimal settings may vary across models and tasks. These factors may limit the applicability of LIME in latency-sensitive or resource-constrained settings.