TEXT-CONDITIONED VISUAL STEERING FOR HALLUCINATION MITIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) are becoming a general interface for visual question answering, captioning and image-grounded reasoning. Their fluency makes them useful in practice, while also making unsupported objects and relations difficult to detect. We evaluate our method on POPE, CHAIR and complementary multimodal benchmarks. Recent inference-time methods use contrastive decoding, attention intervention, visual-token reinjection or latent steering. A representative class of **visual-steering methods** contrasts image-conditioned and text-only activations, extracting a reusable direction for generation. Yet this direction typically compresses the whole visual sequence into one global shift, leaving the text-dependent organization of visual evidence implicit. As a result, a correction that is useful for one question can be poorly aligned with another question about the same image, while methods that preserve token-level structure often add decoding-time computation. To address this gap, we introduce **TVS** (Text-Conditioned Visual Steering), which combines **global image–text contrast** with **text-conditioned visual-token routing**. TVS uses the attention weights from the current text query to estimate which visual tokens should receive more or less emphasis, reweights them around a uniform reference, and pools their hidden states with signed zero-sum coefficients. The resulting correction is added to the inherited global direction, so the method remains training-free, preserves the steering interface and caches one direction for ordinary decoding. Across POPE, CHAIR and general multimodal evaluations, TVS improves the visual-steering baseline on Qwen and on LLaVA MSCOCO, with smaller but positive average gains on LLaVA A-OKVQA and GQA and variation across individual splits. Component ablations verify that the global and token-level terms are complementary, while efficiency experiments show that the cached implementation retains reference decoding accuracy with lower overhead. Code is available at `https://anonymous.4open.science/r/TVS-E8F0/`.

1 INTRODUCTION

Multimodal large language models (MLLMs) Liu et al. (2023; 2024a); Dai et al. (2023); Bai et al. (2025) combine visual encoders with large language decoders and now support visual question answering, captioning and multimodal reasoning. They are increasingly used to assist people with search, accessibility, education and analysis, where a plausible answer can be more harmful than an obvious failure. Hallucinated objects and relations Li et al. (2023b); Rohrbach et al. (2018) weaken trust in these systems and motivate methods that improve visual grounding while preserving the capabilities of the underlying model.

Recent training-free approaches intervene during inference through contrastive decoding, attention modification, visual-token reinjection or latent steering Leng et al. (2024); Huang et al. (2024); Zou et al. (2024); Li et al. (2025). These methods avoid retraining and can be applied to frozen MLLMs, but they operate at different points in the generation pipeline. Contrastive decoders such as VCD Leng et al. (2024) alter the output distribution through a reference branch, attention-based methods such as OPERA Huang et al. (2024) modify internal aggregation during decoding, and reinjection methods such as MemVR Zou et al. (2024) repeatedly restore visual evidence. Latent steering is attractive because a direction can be extracted once and reused for all generated tokens. VISTA Li et al. (2025) uses this idea to contrast image-conditioned and text-only prefills, but its

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.

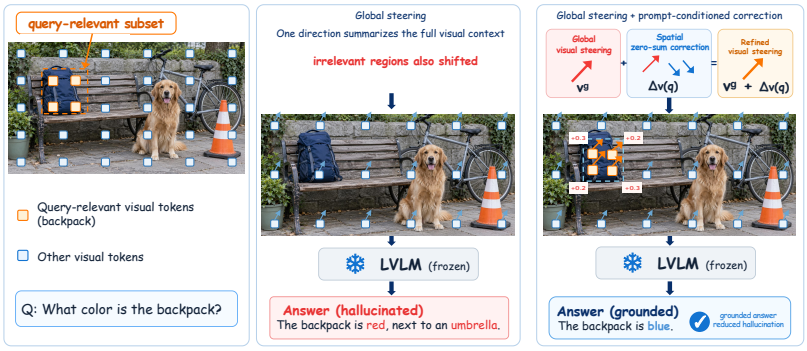

Figure 1: **Motivation.** Given an image and a text query, a cached global visual direction applies broad visual guidance, while query-conditioned token routing can selectively refine the visual evidence used for the answer.

visual-steering component still compresses the entire visual sequence into one layer-wise vector. As illustrated in Figure 1, the text may ask about a particular object or relation, while the correction itself contains no explicit text-dependent organization of visual tokens. Thus, **efficient global guidance needs a text-dependent spatial refinement**.

We address this gap with **Text-Conditioned Visual Steering (TVS)**. The method first obtains the usual global image–text difference from the two prefills. It then uses the current text query's attention weights over the visual tokens to construct a second, token-level correction: attention is normalized over the image, reweighted around uniform coefficients, and compared with a layer-specific reference. The resulting **signed zero-sum coefficients** emphasize differences between visual tokens rather than adding a shared offset. Adding this correction to the global direction preserves the original steering interface, requires no training, and allows the combined direction to be cached before generation. The design therefore separates **global image–text contrast** from **text-conditioned token routing** while retaining one ordinary model forward for each generated token.

We evaluate our method on POPE across MSCOCO, A-OKVQA and GQA, and on CHAIR, MME, MM-Vet and LLaVA-Bench. The method improves the visual-steering baseline broadly, with the strongest and most consistent gains on Qwen and LLaVA MSCOCO. Component ablations show that global and aligned token-level corrections are complementary; shuffled token correspondence weakens the gain. Efficiency experiments show that the fused implementation retains the reference accuracy with lower decoding overhead. Our contributions are summarized as follows:

- We derive a **text-conditioned visual-token contrast** from frozen decoder attention using one additional reweighting coefficient beyond the inherited steering gain.
- We add the contrast through **zero-sum token coefficients**, cache the combined direction, and retain one model forward per generated token without training.
- Extensive experiments on POPE (MSCOCO, A-OKVQA, and GQA), CHAIR, MME, MM-Vet, and LLaVA-Bench demonstrate the effectiveness of our method across both Qwen- and LLaVA-based architectures, achieving state-of-the-art performance on multiple benchmarks.

2 Related Work

2.1 Training-Free Hallucination Mitigation in LVLMs

Because retraining LVLMs is expensive, a large family of methods intervenes purely at inference time. One branch is **contrastive decoding**, which reshapes the output distribution using a reference branch. VCD Leng et al. (2024) contrasts predictions on the original and a diffusion-noised image; ICD Wang et al. (2024) uses a perturbed instruction; DoLa Chuang et al. (2024) contrasts logits between an early and the final layer; SID Huo et al. (2024) constructs the reference through self-introspective vision-token pruning. These methods reduce the contribution of unsupported language priors, with different trade-offs in computation and generation quality Yin et al. (2025).

A second branch **reallocates attention or reinjects visual evidence** inside the network. OPERA Huang et al. (2024) penalizes over-concentrated knowledge-aggregation columns during beam search; PAI Liu et al. (2024c) amplifies attention to image tokens; ClearSight/VAF Yin et al. (2025) strengthens visual fusion signals in intermediate attention layers; FarSight Tang et al. (2025) modifies the causal attention pattern; and MemVR Zou et al. (2024) re-injects visual tokens through feed-forward layers when uncertainty is high. LURE Zhou et al. (2024) instead revises a generated caption. These approaches intervene on logits, attention computation, memory access or generated text. Our work studies a complementary intervention point: the construction of the latent direction used for steering.

2.2 Latent Steering and Visual-Attention Structure

Latent steering. Steering a frozen model by adding a direction to its residual stream originated in the LLM interpretability literature: ITI Li et al. (2023a) shifts activations along a truthfulness direction found by probing, ActAdd Turner et al. (2023) contrasts activations of paired prompts, representation engineering Zou et al. (2023) generalizes this into a framework for reading and controlling representations, and in-context vectors Liu et al. (2024b) compress demonstrations into a latent shift. VISTA Li et al. (2025) combines a visual steering vector with self-logit augmentation in the multimodal setting.

Visual-attention structure. Transformer attention can be highly concentrated and is not, by itself, a semantic explanation. StreamingLLM Xiao et al. (2024) identifies positional attention sinks, while vision transformers can repurpose low-information patches as registers Darcet et al. (2024). In LVLMs, DAMRO Gong et al. (2024) relates high-attention visual outliers to hallucination, Zhang et al. (2024) studies individual heads that retain visual evidence, and Luo et al. (2025) analyzes alternative visual-information pathways through sink tokens. TVS does not alter attention computation or treat attention as ground-truth relevance; it uses normalized within-image attention weights as a text-dependent routing signal for an additive steering correction.

3 Method

Let $\pi = (t_0, \ldots, t_{T-1})$ denote the token sequence after model-specific multimodal expansion. Visual tokens occupy the exclusive-end interval $[v_s, v_e)$, with $N = v_e - v_s$ (N=576 for LLaVA-1.5-7B). At decoder block $\ell \in \{1, \ldots, L\}$, $H^{\mathrm{POS},(\ell)} \in \mathbb{R}^{T\times d}$ denotes the image-conditioned block-output states and $h^{(\ell)}_{-1,\mathrm{POS}} \equiv H^{\mathrm{POS},(\ell)}_{T-1}$. We write $A^{(\ell)} \in \mathbb{R}^{n_h\times T\times T}$ for image-conditioned post-softmax attention, $\bar{A}^{(\ell)} = \frac{1}{n_h}\sum_h A^{(\ell)}_h$, and $A^{(\ell)}_{-1,j} \equiv A^{(\ell)}_{T-1,j}$. For LLaVA-1.5-7B, L=32, n_h=32 and d=4096.

We use a cached global visual direction as the common steering interface. We instantiate the cached global direction with the VSV extractor from VISTA Li et al. (2025). It runs two prefills: **POS** on the image-conditioned prompt and **NEG** on the same textual context without visual tokens. Their layer-wise last-token difference is

$$v^{(\ell)}_{\mathrm{global}} = h^{(\ell)}_{-1,\mathrm{POS}} - h^{(\ell)}_{-1,\mathrm{NEG}} \in \mathbb{R}^d. \tag{1}$$

At each generation step, the same direction is applied using the injection operator shared by the global-direction baseline and our method. In our implementation this operator is

$$\tilde{h}^{(\ell)}_t = h^{(\ell)}_t + \lambda v^{(\ell)}, \tag{2}$$

where λ controls intervention strength. Self-logit augmentation is disabled, so the baseline and TVS differ only in the construction of the latent direction.

3.1 Attention-Derived Token Contrast

Figure 2 summarizes the complete extraction and injection pipeline. Equation equation 1 provides one global direction per layer. TVS augments it with a post-hoc contrast of image-conditioned visual-token states. For each target layer ℓ, we first define the last-text-token attention weights

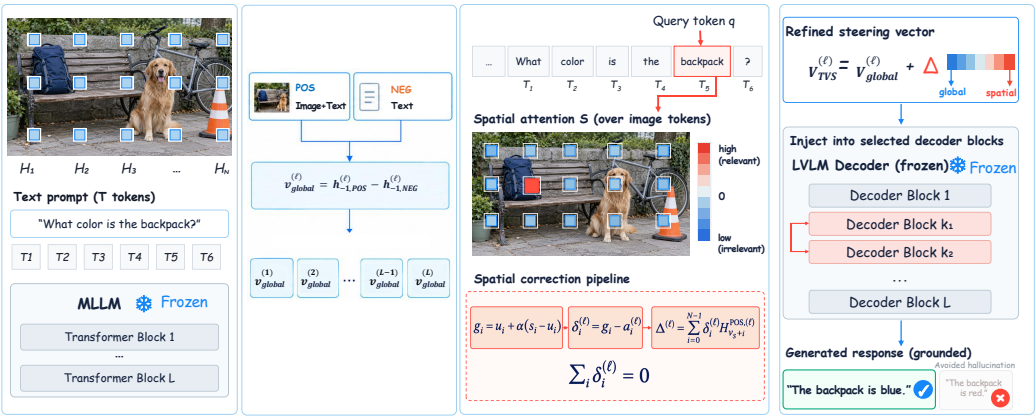

Figure 2: **Overview of our method.** A frozen LVLM provides a cached global visual direction and text-conditioned attention weights. TVS forms signed zero-sum coefficients, pools visual-token states into $\Delta^{(\ell)}$, and adds this correction before decoding-time injection. The global-direction baseline provides global visual guidance, while TVS adds a prompt-conditioned, spatially selective correction with zero-sum token coefficients.

within the visual span:

$$a_i^{(\ell)} = \frac{\bar{A}^{(\ell)}_{-1,\, v_s+i}}{\sum_{j=0}^{N-1} \bar{A}^{(\ell)}_{-1,\, v_s+j}}, \qquad i = 0, \ldots, N-1. \tag{3}$$

Thus $a^{(\ell)} \in \Delta^{N-1}$ is a normalized, layer-specific reference distribution. It defines a post-hoc pooling of block-output token states; it is not the block's value-projected attention output.

3.2 TVS: Text-Conditioned Visual Steering

Text-conditioned attention readout. Let nonempty $\mathcal{S} \subseteq \{1, \ldots, L\}$ and $\mathcal{Q} \subseteq \{0, \ldots, T-1\}$ denote source layers and text-query positions whose total attention to the visual span is positive. We average image-conditioned attention over heads, source layers and queries,

$$\tilde{s}_i = \frac{1}{|\mathcal{S}|\,|\mathcal{Q}|\,n_h} \sum_{k \in \mathcal{S}} \sum_{q \in \mathcal{Q}} \sum_{r=1}^{n_h} A^{(k)}_{r,q,\, v_s+i}, \tag{4}$$

and normalize the result within the visual span, $s_i = \tilde{s}_i / \sum_j \tilde{s}_j$, yielding $s \in \Delta^{N-1}$. Because normalization follows raw-attention averaging, source/query pairs with more total visual attention contribute more to s. Unless stated otherwise, $\mathcal{S} = \{L\}$ and $\mathcal{Q} = \{T-1\}$. The selected query positions follow the visual span in our text inputs; the implementation falls back to the cached global direction if the normalizer is numerically zero.

Centered attention reweighting. Let $u_i = 1/N$ denote uniform coefficients. We define

$$g_i = u_i + \alpha(s_i - u_i) = \alpha s_i + (1-\alpha)\tfrac{1}{N}. \tag{5}$$

This form makes the role of α explicit: $g - u = \alpha(s-u)$. Values in $[0,1]$ move from uniform weights toward the observed attention distribution, whereas $\alpha > 1$ sharpens the deviations from uniform. The coefficients still sum to one but may be negative, so g is a signed reweighting vector rather than a probability distribution.

Centered token-contrast correction. For each target layer, we compare the shared coefficient vector g with the layer-specific last-token reference profile $a^{(\ell)}$:

$$\delta_i^{(\ell)} = g_i - a_i^{(\ell)}, \tag{6}$$

$$\Delta^{(\ell)} = \sum_{i=0}^{N-1} \delta_i^{(\ell)} H^{\mathrm{POS},(\ell)}_{v_s+i}. \tag{7}$$

Algorithm 1 Single-Forward TVS Extraction and Injection

Input: frozen LVLM $\mathcal{M}_\theta$, image I, prompt p, α, $\mathcal{S}$, $\mathcal{Q}$
$(H^{\mathrm{POS}}, A^{\mathrm{POS}}) \leftarrow \mathcal{M}_\theta(I, p)$
$H^{\mathrm{NEG}} \leftarrow \mathcal{M}_\theta(p)$
locate visual span $[v_s, v_e)$; $N \leftarrow v_e - v_s$; $u \leftarrow \mathbf{1}_N/N$
compute normalized source profile s from A^{POS}, $\mathcal{S}$ and $\mathcal{Q}$
$g \leftarrow u + \alpha(s - u)$
for $\ell = 1, \ldots, L$ **do**
    head-average and normalize $A^{\mathrm{POS},(\ell)}_{-1,v_s:v_e}$ to obtain $a^{(\ell)}$
    $\Delta^{(\ell)} \leftarrow \sum_{i=0}^{N-1}(g_i - a_i^{(\ell)})H^{\mathrm{POS},(\ell)}_{v_s+i}$
    $v^{(\ell)}_{\mathrm{tvs}} \leftarrow h^{(\ell)}_{-1,\mathrm{POS}} - h^{(\ell)}_{-1,\mathrm{NEG}} + \Delta^{(\ell)}$
end for
cache $\{v^{(\ell)}_{\mathrm{tvs}}\}_{\ell=1}^{L}$
generate each token with one model forward, injecting the cached direction via Eq. 2

The resulting steering direction is

$$\boxed{v^{(\ell)}_{\mathrm{tvs}} = v^{(\ell)}_{\mathrm{global}} + \Delta^{(\ell)}}. \tag{8}$$

Under the joint default $\mathcal{S} = \{L\}$ and $\mathcal{Q} = \{T-1\}$, $s = a^{(L)}$ and $\delta^{(L)} = (\alpha - 1)(a^{(L)} - u)$. At lower target layers, $\delta^{(\ell)}$ contrasts the shared final-layer distribution with the layer-specific last-token distribution. The query set $\mathcal{Q}$ affects only g; $a^{(\ell)}$ is always read from the last text token. The resulting direction is injected with Eq. 2. Algorithm 1 summarizes the procedure.

3.3 Properties and Interpretation

Zero-sum coefficients. Since g and $a^{(\ell)}$ both sum to one,

$$\sum_i \delta_i^{(\ell)} = 0. \tag{9}$$

This is a coefficient-space property: it centers the post-hoc token pooling, but does not constrain the norm, mean or angle of the resulting vector $\Delta^{(\ell)}$.

Common-component cancellation. For fixed attention-derived coefficients and any $b \in \mathbb{R}^d$ shared by all pooled visual-token states,

$$\sum_i \delta_i^{(\ell)} \big(H^{\mathrm{POS},(\ell)}_{v_s+i} + b\big) = \Delta^{(\ell)}. \tag{10}$$

This property applies to the post-hoc linear pooling with fixed $\delta^{(\ell)}$; it is not an invariance claim about recomputing the full network after perturbing its hidden states. The cached global direction remains an unmodified additive term in Eq. 8.

Computation. **TVS** retains the cached global steering procedure's two one-time prefills: image-conditioned POS and text-only NEG. During POS, visual-span attention and token states are immediately reduced to $\Delta^{(\ell)}$ and combined with $v^{(\ell)}_{\mathrm{global}}$. The resulting direction is cached, so each generated token requires one ordinary model forward with an in-stream residual addition. Thus *single-forward* refers to decoding-time branches, not to the number of prefills.

4 Experiments

Backbones. We evaluate LLaVA-1.5-7B Liu et al. (2024a), which uses a fixed grid of 576 visual tokens, and Qwen2.5-VL-7B-Instruct Bai et al. (2025), which uses a dynamic-resolution visual sequence. Every paired comparison within a backbone uses the same prompt, generation, decoding and evaluation configuration.

Baselines. Vanilla decoding measures the unchanged backbone. VSV Li et al. (2025) is the steering component of VISTA with self-logit augmentation disabled. VSV and TVS share the same

Table 1: **POPE results on Qwen2.5-VL** across MSCOCO, A-OKVQA and GQA. Cells report Accuracy/F1 and the point change from Vanilla.

Method	Random		Popular		Adversarial		Average	
	Acc.↑	F1↑	Acc.↑	F1↑	Acc.↑	F1↑	Acc.↑	F1↑
MSCOCO								
Vanilla	82.20 ↑0.00	78.40 ↑0.00	82.40 ↑0.00	78.98 ↑0.00	81.27 ↑0.00	77.52 ↑0.00	81.96 ↑0.00	78.30 ↑0.00
ICD	82.57 ↑0.37	80.16 ↑1.76	81.96 ↓0.44	79.68 ↑0.70	82.36 ↑1.09	79.62 ↑2.10	82.30 ↑0.34	79.82 ↑1.52
VCD	82.10 ↓0.10	79.24 ↑0.84	80.06 ↓2.34	79.84 ↑0.86	81.86 ↑0.59	79.54 ↑2.02	81.34 ↓0.62	79.54 ↑1.24
OPERA	82.19 ↓0.01	79.88 ↑1.48	83.40 ↑1.00	80.11 ↑1.13	83.24 ↑1.97	80.07 ↑2.55	82.94 ↑0.98	80.02 ↑1.72
MemVR	83.80 ↑1.60	80.87 ↑2.47	83.01 ↑0.61	80.16 ↑1.18	83.09 ↑1.82	80.58 ↑3.06	83.30 ↑1.34	80.54 ↑2.24
LEAD	84.07 ↑1.87	82.92 ↑4.52	82.47 ↑0.07	81.35 ↑2.37	83.80 ↑2.53	80.11 ↑2.59	83.45 ↑1.49	81.46 ↑3.16
VSV	90.13 ↑7.93	89.82 ↑11.42	88.47 ↑6.07	88.33 ↑9.35	85.10 ↑3.83	85.43 ↑7.91	87.90 ↑5.94	87.86 ↑9.56
TVS (Ours)	**90.63** ↑8.43	**90.00** ↑11.60	**89.10** ↑6.70	**88.53** ↑9.55	**86.17** ↑4.90	**85.89** ↑8.37	**88.63** ↑6.67	**88.14** ↑9.84
A-OKVQA								
Vanilla	87.67 ↑0.00	86.37 ↑0.00	85.17 ↑0.00	83.86 ↑0.00	80.37 ↑0.00	79.81 ↑0.00	84.40 ↑0.00	83.34 ↑0.00
ICD	87.29 ↓0.38	86.67 ↑0.30	86.03 ↑0.86	84.88 ↑1.02	80.02 ↓0.35	80.79 ↑0.98	84.45 ↑0.05	84.11 ↑0.77
VCD	86.73 ↓0.94	85.91 ↓0.46	84.67 ↓0.50	84.16 ↑0.30	79.53 ↓0.84	79.11 ↓0.70	83.64 ↓0.76	83.06 ↓0.28
OPERA	87.59 ↓0.08	86.45 ↑0.08	85.88 ↑0.71	84.67 ↑0.81	78.32 ↓2.05	80.92 ↑1.11	83.93 ↓0.47	84.01 ↑0.67
MemVR	88.07 ↑0.40	87.18 ↑0.81	85.43 ↑0.26	85.49 ↑1.63	80.87 ↑0.50	80.69 ↑0.88	84.79 ↑0.39	84.45 ↑1.11
LEAD	88.63 ↑0.96	87.56 ↑1.19	85.50 ↑0.33	84.66 ↑0.80	80.47 ↑0.10	80.87 ↑1.06	84.87 ↑0.47	84.36 ↑1.02
VSV	90.17 ↑2.50	90.51 ↑4.14	88.83 ↑3.66	89.32 ↑5.46	80.40 ↑0.03	82.67 ↑2.86	86.47 ↑2.07	87.50 ↑4.16
TVS (Ours)	**92.23** ↑4.56	**92.21** ↑5.84	**89.20** ↑4.03	**89.47** ↑5.61	**82.03** ↑1.66	**83.62** ↑3.81	**87.82** ↑3.42	**88.44** ↑5.10
GQA								
Vanilla	80.97 ↑0.00	76.93 ↑0.00	79.53 ↑0.00	75.92 ↑0.00	77.43 ↑0.00	74.11 ↑0.00	79.31 ↑0.00	75.65 ↑0.00
ICD	87.40 ↑6.43	85.32 ↑8.39	83.49 ↑3.96	82.79 ↑6.87	81.66 ↑4.23	79.94 ↑5.83	84.18 ↑4.87	82.68 ↑7.03
VCD	87.49 ↑6.52	86.17 ↑9.24	82.87 ↑3.34	81.53 ↑5.61	79.65 ↑2.22	80.90 ↑6.79	83.34 ↑4.03	82.87 ↑7.22
OPERA	86.99 ↑6.02	85.00 ↑8.07	81.85 ↑2.32	82.06 ↑6.14	80.53 ↑3.10	81.35 ↑7.24	83.12 ↑3.81	82.80 ↑7.15
MemVR	88.07 ↑7.10	86.90 ↑9.97	82.43 ↑2.90	81.83 ↑5.91	80.93 ↑3.50	80.41 ↑6.30	83.81 ↑4.50	83.05 ↑7.40
LEAD	88.53 ↑7.56	87.64 ↑10.71	84.67 ↑5.14	82.88 ↑6.96	80.57 ↑3.14	80.41 ↑6.30	84.59 ↑5.28	83.64 ↑7.99
VSV	85.93 ↑4.96	85.75 ↑8.82	84.13 ↑4.60	84.15 ↑8.23	79.33 ↑1.90	80.33 ↑6.22	83.13 ↑3.82	83.41 ↑7.76
TVS (Ours)	**90.20** ↑9.23	**89.75** ↑12.82	**87.07** ↑7.54	**86.88** ↑10.96	**82.97** ↑5.54	**83.38** ↑9.27	**86.74** ↑7.43	**86.67** ↑11.02

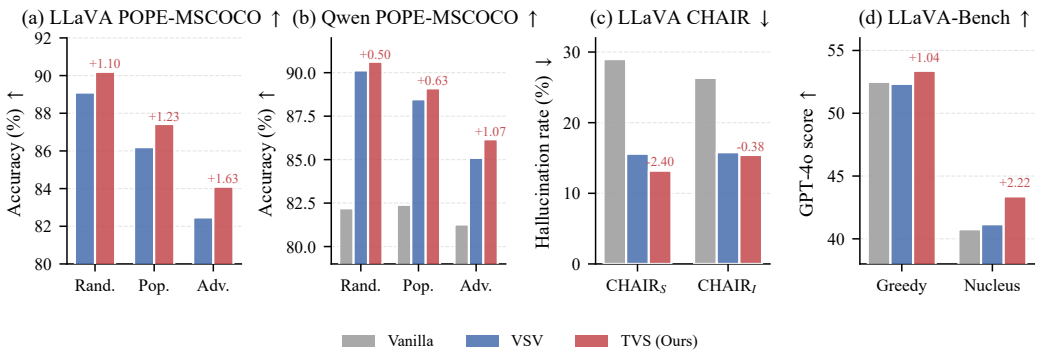

Figure 3: **Main results.** POPE, CHAIR and LLaVA-Bench comparisons; changes are relative to the visual-steering baseline.

contrastive prefills, injection rule and gain, isolating the contribution of text-conditioned spatial refinement. Detailed settings are given in Appendix A.

Benchmarks. We evaluate POPE Li et al. (2023b) on MSCOCO Lin et al. (2014), A-OKVQA Schwenk et al. (2022) and GQA Hudson & Manning (2019), each with random, popular and adversarial splits of 3000 questions. Both backbones are complete on all three datasets. On LLaVA, we additionally evaluate CHAIR Rohrbach et al. (2018), MME Fu et al. (2023), MM-Vet Yu et al. (2024) and LLaVA-Bench Liu et al. (2023).

RQ1: Does text-conditioned routing improve hallucination mitigation? POPE across datasets. Table 1 shows that TVS improves over the visual-steering baseline on all three Qwen datasets and on both backbones for MSCOCO. The strongest and most consistent gains occur on Qwen, whereas LLaVA gains on A-OKVQA and GQA are smaller on average and mixed across individual splits. The pattern supports text-conditioned routing and also exposes its architecture dependence; complete results are provided in the supplementary material.

Free-form captioning. POPE evaluates binary object existence, whereas CHAIR measures unsupported object mentions in generated captions. On LLaVA, **TVS** lowers CHAIR_S from 15.6 to 13.2 and CHAIR_I from 15.81 to 15.43 (Tab. 2 and Fig. 3(c)).

Table 2: **CHAIR and caption coverage** with LLaVA-1.5-7B under greedy decoding. Length, Obj. and Recall measure generated tokens, mentioned objects and object recall.

Method	$\text{CHAIR}_S \downarrow$	$\text{CHAIR}_I \downarrow$	Length↑	Obj.↑	Recall↑
Vanilla	29.0	26.34	93.5	1.45	32.77
VSV Li et al. (2025)	15.6	15.81	85.0	1.09	29.30
TVS (Ours)	**13.2**	**15.43**	**87.5**	**1.27**	**30.23**

This change is accompanied by better coverage than the global-direction baseline: TVS generates 2.9% more tokens, mentions 16.5% more objects and improves object recall by 3.2%. CHAIR_S decreases by 15.4%, while the per-mention CHAIR_I changes by only 2.4%. The result therefore improves hallucination metrics without the earlier coverage loss.

RQ2: Does hallucination mitigation preserve general multimodal ability? Capability controls. Tables 3 and 4 test general multimodal ability under matched decoding. Under greedy decoding, TVS remains close to the visual-steering baseline on MME, MM-Vet and LLaVA-Bench.

Table 3: **MME under greedy and nucleus decoding** with LLaVA-1.5-7B and λ=0.01. Differences are measured from Vanilla.

Method	Greedy↑	Nucleus↑
Vanilla	1755.9	1479.2
VSV Li et al. (2025)	1752.5 ↓3.4	1524.8 ↑45.6
TVS (Ours)	**1762.0** ↑6.1	**1566.9** ↑87.7

Table 4: **General multimodal capability** on LLaVA-1.5-7B with λ=0.01. Results cover MM-Vet and LLaVA-Bench under greedy and nucleus decoding.

Method	MM-Vet		LLaVA-Bench	
	Greedy↑	Nucleus↑	Greedy↑	Nucleus↑
Vanilla	27.43	24.02	52.50	40.76
VSV Li et al. (2025)	26.17	22.10	52.34	41.16
TVS (Ours)	**27.69**	**24.70**	**53.38**	**43.38**

Nucleus decoding is more benchmark-dependent, while the results remain consistent with capability retention under matched decoding. The method is designed for hallucination mitigation, so these controls establish preservation of general ability rather than claim a separate capability improvement.

RQ3: Which direction components explain the observed gains? The component ablations use LLaVA-1.5-7B and the full POPE-MSCOCO protocol (3 splits $\times$ 3000 samples), with $\lambda = 0.17$, batch size 8 and seed 42. Table 5 reports the matched direction-component comparison. Direction components are compared in a last-token shared-evaluator rerun; only within this matched suite are the component differences interpreted.

Table 5: **Direction-component ablation on LLaVA POPE-MSCOCO.** All variants share the evaluator and configuration; Uncentered (NM) omits centering and is norm-matched to TVS, while Shuffled averages three permutations.

Direction	Random	Popular	Adversarial	Average
Vanilla	89.07	85.63	79.20	84.63
VSV	89.80	87.20	82.47	86.49
Spatial only	88.83	86.97	82.27	86.02
Uncentered (NM)	**89.93**	87.30	82.97	86.73
Shuffled	89.18	86.20	81.71	85.70
Full TVS	89.77	**87.33**	**83.27**	**86.79**

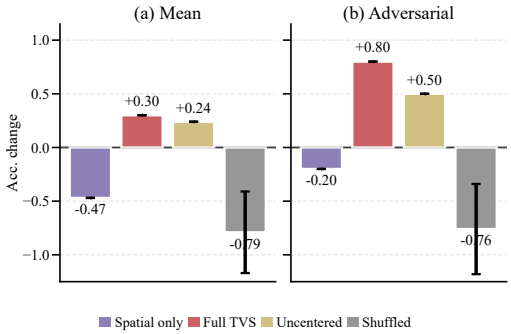

Figure 4: **Direction components.** Accuracy changes for mean and adversarial splits; shuffled routing averages three seeds.

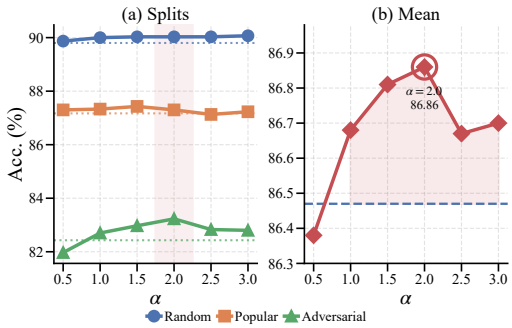

Figure 5: **Sensitivity to α.** POPE-MSCOCO accuracy; shading marks the default neighborhood.

Direction components. The full direction attains the highest mean Accuracy, improving the baseline by 0.30 points and the spatial-only variant by 0.77 points. Text-conditioned spatial correction complements the global image–text contrast. Shuffling the visual-token correspondence reduces mean Accuracy by 1.09 points relative to Full TVS and weakens the adversarial split by 1.56 points (Fig. 4), showing that the structured token assignment matters.

The uncentered, norm-matched (NM) control pools directly with g and rescales the correction to match the TVS norm. It comes within 0.06 points of Full TVS on average, while Full TVS is 0.30 points higher on adversarial questions. We retain zero-sum centering for its exact common-component cancellation property; the clearest empirical gain comes from combining global steering with aligned token-level correction.

Configuration and validation analyses. We next group the tunable coefficients, attention readout choices and coverage-constrained validation in one configuration analysis. These studies are not additional benchmark claims: all comparisons use matched LLaVA POPE-MSCOCO settings, and the selected values are fixed before the corresponding backbone-level evaluations.

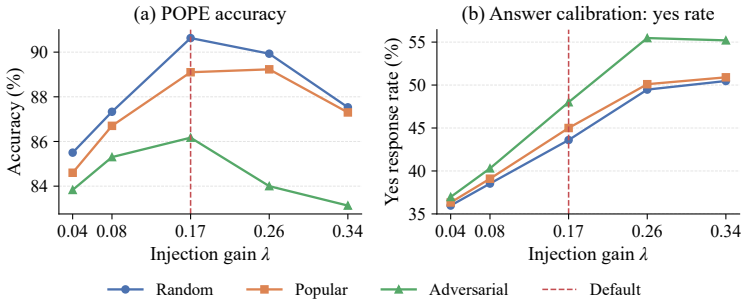

Figure 6: **Qwen injection-gain sensitivity.** Accuracy and yes-response rate on Qwen2.5-VL-7B-Instruct POPE-MSCOCO versus λ; the dashed line marks the default and yes rate summarizes answer calibration.

Injection gain λ. Figure 6 shows the Qwen2.5-VL-7B-Instruct POPE-MSCOCO sweep. With final-layer, last-token guidance, $\lambda = 0.17$ maximizes mean Accuracy; smaller gains over-favor negative answers, while larger gains hurt adversarial accuracy. The broad plateau around this value shows that the method is robust to moderate coefficient changes.

Sharpening coefficient α. Together with the gain sweep, this study tests how strongly the attention profile is used for token routing.

The coefficient α controls how far the signed coefficient vector moves from uniform weighting toward the normalized attention distribution. Figure 5 sweeps $\alpha \in \{0.5, 1.0, 1.5, 2.0, 2.5, 3.0\}$; the supplementary material reports all split-wise values.

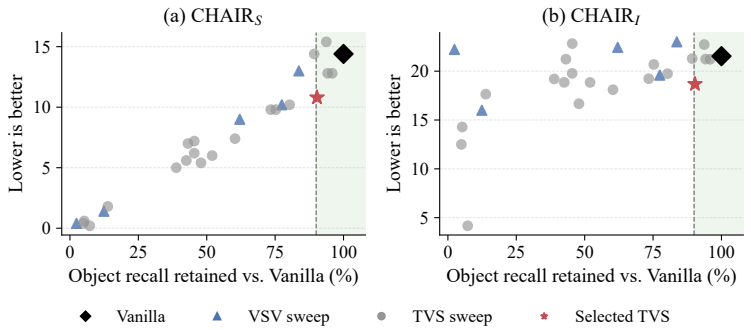

Figure 7: **Qwen CHAIR–coverage trade-off.** Qwen2.5-VL-7B-Instruct results on a fixed 500-image validation set; each point is one (λ, α) configuration, and the shaded region retains at least 90% of Vanilla object recall.

Mean accuracy rises from 86.38 at $\alpha = 0.5$ to 86.86 at $\alpha = 2.0$. Every setting from 1.0 to 3.0 remains above the baseline of 86.47, showing that the benefit spans a broad range around the default. The curve is relatively flat near its maximum: $\alpha = 1.5$, 2.0 and 2.5 score 86.81, 86.86 and 86.67.

Figure 7 shows the Qwen2.5-VL-7B-Instruct validation sweep on a fixed 500-image CHAIR set. The selected TVS setting reduces CHAIR while preserving caption length and object recall. Configurations in the shaded region retain at least 90% of Vanilla object recall, so the selected point reflects a practical accuracy–hallucination trade-off.

The three splits respond differently to sharpening. Random accuracy changes by only 0.20 points over the sweep, and popular changes by 0.30. Adversarial accuracy increases from 81.97 at $\alpha = 0.5$ to 83.23 at $\alpha = 2.0$, accounting for most of the change in the mean. This mirrors the POPE-MSCOCO comparison, where adversarial questions show the largest gain on both backbones.

The supplementary material reports the complete source-layer and query-position comparisons. Final-layer guidance is stronger than the tested middle-layer band, while last-token and object-token queries differ by at most 0.17 points on LLaVA MSCOCO. The frozen POPE pipelines use the object token for LLaVA and the last text token for Qwen.

RQ4: Does the cached implementation reduce decoding overhead? VCD contrasts logits from the original and diffusion-noised image branches at every decoding step, requiring two model forwards per generated token. Cached global steering instead extracts one direction from two one-time prefills and reuses it during decoding; TVS preserves this cached-direction design while adding its text-conditioned visual-token correction during prefill. As shown in Fig. 8, TVS also uses one decoding forward per token. Table 16 reports the timing and memory protocol in the supplement; the fused implementation stays within 0.08 mean Accuracy points of the two-forward reference.

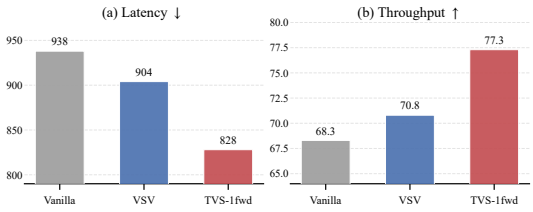

Figure 8: **Efficiency.** Decode latency and throughput on LLaVA-1.5-7B.

5 Conclusion

TVS is a training-free refinement of global visual steering that adds text-conditioned token structure. Across POPE and CHAIR, a reusable global direction benefits from a lightweight correction that preserves text-selected visual organization. Experiments across these datasets and on both Qwen- and LLaVA-based architectures consistently validate the effectiveness of our method. The ablations support complementary global and aligned token-level corrections, and shuffled correspondence weakens the result. The sensitivity studies identify a usable range for steering strength and attention sharpening. TVS requires no training, computes its correction during existing prefills, caches it with the global direction, and uses one ordinary model forward per generated token, making text-dependent visual refinement practical for frozen MLLMs.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations (ICLR)*, 2024.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Xuan Gong, Tianshi Ming, Xinpeng Wang, and Zhihua Wei. Damro: Dive into the attention mechanism of lvlm to reduce object hallucination. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7696–7712, 2024.

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 (CVPR)*, pp. 13418–13427, 2024.

Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6700–6709, 2019.

Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-introspective decoding: Alleviating hallucinations for large vision-language models. *arXiv preprint arXiv:2408.02032*, 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 (CVPR)*, pp. 13872–13882, 2024.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems (NeurIPS)*, 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 (EMNLP)*, pp. 292–305, 2023b.

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N. Metaxas. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. In *International Conference on Machine Learning (ICML)*, 2025.

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 (ECCV)*, pp. 740–755, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26296–26306, 2024a.

Sheng Liu, Haotian Ye, Lei Xing, and James Zou. In-context vectors: Making in context learning more effective and controllable through latent space steering. In *International Conference on Machine Learning (ICML)*, 2024b.

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 (ECCV)*, pp. 125–140, 2024c.

Jiayun Luo, Wan-Cyuan Fan, Lyuyang Wang, Xiangteng He, Tanzila Rahman, Purang Abolmaesumi, and Leonid Sigal. To sink or not to sink: Visual information pathways in large vision-language models. *arXiv preprint arXiv:2510.08510*, 2025.

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

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 (ECCV)*, pp. 146–162, 2022.

Feilong Tang, Chengzhi Liu, Zhongxing Xu, Ming Hu, Zile Huang, Haochen Xue, Ziyang Chen, Zelin Peng, Zhiwei Yang, Sijin Zhou, et al. Seeing far and clearly: Mitigating hallucinations in mllms with attention causal decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26147–26159, 2025.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. *arXiv preprint arXiv:2403.18715*, 2024.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *International Conference on Learning Representations (ICLR)*, 2024.

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

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *International Conference on Machine Learning (ICML)*, 2024.

Xiaofeng Zhang, Yihao Quan, Chaochen Gu, Chen Shen, Xiaosong Yuan, Shaotian Yan, Hao Cheng, Kaijie Wu, and Jieping Ye. Seeing clearly by layer two: Enhancing attention heads to alleviate hallucination in lvlms. *arXiv preprint arXiv:2411.09968*, 2024.

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 (ICLR)*, 2024.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

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. *arXiv preprint arXiv:2410.03577*, 2024.

SUPPLEMENTARY MATERIAL

A CONFIGURATION

TVS augments a cached global visual direction with query-conditioned token routing while retaining the same image-conditioned and text-only prefills, decoding-time injection rule and gain. In our experiments, the cached direction is instantiated with the VSV extractor from VISTA. For LLaVA-1.5-7B, the visual span contains a fixed grid of 576 tokens; Qwen2.5-VL-7B-Instruct uses a sample-dependent multimodal layout. All POPE runs use $\alpha = 2.0$, $\lambda = 0.17$ and the final decoder layer as the attention source; LLaVA queries the object token, whereas Qwen queries the last text token. Within every benchmark and backbone, paired global-direction-baseline and TVS runs share prompts, decoding parameters and evaluation settings.

In object-query ablations, tokenizer offsets identify the queried noun phrase and the source attention weights are averaged over its subword positions. TVS records and reduces visual-span attention and token states during the image-conditioned prefill, caches the combined direction and uses one model forward per generated token. The POS/NEG prefills remain unchanged; "single-forward" refers to decoding-time branches. Algorithm 1 in the main paper summarizes direction extraction.

B COMPLETE QUANTITATIVE RESULTS

Table 6 summarizes the hallucination and capability results. For LLaVA CHAIR, average caption length/object mentions/object recall are $93.5/1.45/32.77$ for Vanilla, $85.0/1.09/29.30$ for the global-direction baseline and $87.5/1.27/30.23$ for TVS. The lower CHAIR scores are therefore accompanied by improved caption coverage relative to the global-direction baseline. Table 7 reports the complete LLaVA three-dataset POPE results; Table 8 gives its cross-dataset macro-average. Table 9 reports the Qwen results across all three datasets.

Table 6: **Comparison with VSV.** POPE rows distinguish MSCOCO from macro-averages over completed datasets. Best-value highlighting is computed only between VSV and TVS; Vanilla is a non-competing reference.

Backbone / metric	Vanilla	VSV	TVS
Hallucination evaluation			
LLaVA: POPE-MSCOCO avg. Acc ↑	79.83	85.92	**87.24**
LLaVA: TVS-1fwd audit avg. Acc ↑	81.23	86.47	**86.83**
LLaVA: POPE 3-dataset macro Acc ↑	79.32	83.82	**86.31**
Qwen: POPE-MSCOCO avg. Acc ↑	81.96	87.90	**88.63**
Qwen: POPE 3-dataset macro Acc ↑	81.89	85.83	**87.73**
LLaVA: CHAIR_S ↓	29.0	15.6	**13.2**
LLaVA: CHAIR_I ↓	26.34	15.81	**15.43**
LLaVA: caption length ↑	93.5	85.0	**87.5**
LLaVA: object recall ↑	32.77	29.30	**30.23**
General-capability evaluation on LLaVA			
MME Greedy ↑	1755.9	1752.5	**1762.0**
MME Nucleus ↑	1479.2	1524.8	**1566.9**
MM-Vet Greedy ↑	27.43	26.17	**27.69**
MM-Vet Nucleus ↑	24.02	22.10	**24.70**
LLaVA-Bench Greedy ↑	52.50	52.34	**53.38**
LLaVA-Bench Nucleus ↑	40.76	41.16	**43.38**

Table 8: **LLaVA macro-average over three POPE datasets and nine splits.** Values are arithmetic means of the dataset-level averages in Tab. 7.

Method	Mean Acc.	Mean F1
Vanilla	79.32	79.17
VSV Li et al. (2025)	83.82	84.83
TVS (Ours)	**86.31**	**86.06**

Table 7: **POPE results on LLaVA-1.5-7B** across MSCOCO, A-OKVQA and GQA (3000 questions per split). Each cell reports Accuracy/F1 followed by its absolute point change from Vanilla; Average is the arithmetic mean over three splits.

Method	Random		Popular		Adversarial		Average	
	Acc.↑	F1↑	Acc.↑	F1↑	Acc.↑	F1↑	Acc.↑	F1↑
MSCOCO								
Vanilla	83.49 ↑0.00	82.28 ↑0.00	79.98 ↑0.00	79.34 ↑0.00	76.03 ↑0.00	76.26 ↑0.00	79.83 ↑0.00	79.29 ↑0.00
OPERA	87.53 ↑4.04	86.45 ↑4.17	84.21 ↑4.23	83.50 ↑4.16	80.88 ↑4.85	80.69 ↑4.43	84.21 ↑4.38	83.55 ↑4.26
ICD	84.87 ↑1.38	83.27 ↑0.99	82.93 ↑2.95	81.45 ↑2.11	81.07 ↑5.04	79.96 ↑3.70	82.96 ↑3.13	81.56 ↑2.27
VCD	86.84 ↑3.35	86.83 ↑4.55	82.65 ↑2.67	83.37 ↑4.03	77.31 ↑1.28	79.28 ↑3.02	82.27 ↑2.44	83.16 ↑3.87
MemVR	88.50 ↑5.01	87.34 ↑5.06	87.10 ↑7.12	86.01 ↑6.67	85.20 ↑9.17	84.28 ↑8.02	86.93 ↑7.10	85.88 ↑6.59
VAF	87.60 ↑4.11	86.20 ↑3.92	86.20 ↑6.22	85.00 ↑5.66	83.90 ↑7.87	82.80 ↑6.54	85.90 ↑6.07	84.67 ↑5.38
AIR	89.00 ↑5.51	88.20 ↑5.92	87.10 ↑7.12	86.40 ↑7.06	83.90 ↑7.87	83.60 ↑7.34	86.67 ↑6.84	86.07 ↑6.78
VSV	89.10 ↑5.61	89.33 ↑7.05	86.20 ↑6.22	86.97 ↑7.63	82.47 ↑6.44	82.97 ↑6.71	85.92 ↑6.09	86.42 ↑7.13
TVS (Ours)	**90.20** ↑6.71	**89.69** ↑7.41	**87.43** ↑7.45	**87.16** ↑7.82	**84.10** ↑8.07	**83.46** ↑7.20	**87.24** ↑7.41	**86.77** ↑7.48
A-OKVQA								
Vanilla	83.45 ↑0.00	82.56 ↑0.00	79.90 ↑0.00	79.59 ↑0.00	74.04 ↑0.00	75.15 ↑0.00	79.13 ↑0.00	79.10 ↑0.00
OPERA	88.27 ↑4.82	87.54 ↑4.98	85.17 ↑5.27	83.74 ↑4.15	79.37 ↑5.33	79.97 ↑4.82	84.27 ↑5.14	83.75 ↑4.65
ICD	85.57 ↑2.12	85.06 ↑2.50	81.93 ↑2.03	81.95 ↑2.36	77.43 ↑3.39	78.99 ↑3.84	81.64 ↑2.51	82.00 ↑2.90
VCD	86.15 ↑2.70	86.34 ↑3.78	81.85 ↑1.95	82.82 ↑3.23	74.97 ↑0.93	77.73 ↑2.58	80.99 ↑1.86	82.30 ↑3.20
MemVR	91.10 ↑7.65	**90.83** ↑8.27	87.33 ↑7.43	87.43 ↑7.84	80.20 ↑6.16	**81.66** ↑6.51	86.21 ↑7.08	**86.64** ↑7.54
VAF	88.10 ↑4.65	88.30 ↑5.74	82.10 ↑2.20	84.30 ↑4.71	73.90 ↓0.14	78.70 ↑3.55	81.37 ↑2.24	83.77 ↑4.67
AIR	89.20 ↑5.75	88.90 ↑6.34	84.40 ↑4.50	84.70 ↑5.11	78.00 ↑3.96	79.70 ↑4.55	83.87 ↑4.74	84.43 ↑5.33
VSV	87.67 ↑4.22	89.02 ↑6.46	84.13 ↑4.23	85.27 ↑5.68	74.83 ↑0.79	78.50 ↑3.35	82.21 ↑3.08	84.26 ↑5.16
TVS (Ours)	**91.50** ↑8.05	**91.70** ↑9.14	**90.83** ↑10.93	**88.81** ↑9.22	**81.40** ↑7.36	80.91 ↑5.76	**87.91** ↑8.78	**87.14** ↑8.04
GQA								
Vanilla	83.73 ↑0.00	82.95 ↑0.00	78.17 ↑0.00	78.37 ↑0.00	75.08 ↑0.00	76.06 ↑0.00	78.99 ↑0.00	79.13 ↑0.00
OPERA	88.27 ↑4.54	87.52 ↑4.57	**83.07** ↑4.90	82.93 ↑4.56	80.77 ↑5.69	81.05 ↑4.99	84.04 ↑5.05	83.83 ↑4.70
ICD	84.90 ↑1.17	84.22 ↑1.27	78.37 ↑0.20	78.81 ↑0.44	75.97 ↑0.89	76.93 ↑0.87	79.75 ↑0.76	79.99 ↑0.86
VCD	86.65 ↑2.92	86.99 ↑4.04	80.73 ↑2.56	82.24 ↑3.87	76.09 ↑1.01	78.78 ↑2.72	81.16 ↑2.17	82.67 ↑3.54
MemVR	89.60 ↑5.87	89.32 ↑6.37	84.63 ↑6.46	84.98 ↑6.61	81.53 ↑6.45	**82.48** ↑6.42	85.25 ↑6.26	85.59 ↑6.46
VAF	88.10 ↑4.37	87.70 ↑4.75	79.40 ↑1.23	80.60 ↑2.23	78.20 ↑3.12	79.70 ↑3.64	81.90 ↑2.91	82.67 ↑3.54
AIR	89.70 ↑5.97	88.50 ↑5.55	83.00 ↑4.83	82.80 ↑4.43	80.40 ↑5.32	81.70 ↑5.64	**84.37** ↑5.38	**84.33** ↑5.20
VSV	88.67 ↑4.94	88.09 ↑5.14	82.12 ↑3.95	83.12 ↑4.75	79.20 ↑4.12	80.23 ↑4.17	83.33 ↑4.34	83.81 ↑4.68
TVS (Ours)	**89.93** ↑6.20	**89.95** ↑7.00	79.23 ↑1.06	**83.51** ↑5.14	**82.17** ↑7.09	79.34 ↑3.28	83.78 ↑4.79	84.27 ↑5.14

Table 9: **Protocol-matched POPE results on Qwen2.5-VL-7B-Instruct** (3000 questions per split). Each cell reports Accuracy/F1; Average is the arithmetic mean over three splits.

Method	Random		Popular		Adversarial		Average	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
MSCOCO								
Vanilla	82.20	78.40	82.40	78.98	81.27	77.52	81.96	78.30
VSV Li et al. (2025)	90.13	89.82	88.47	88.33	85.10	85.43	87.90	87.86
TVS (Ours)	**90.63**	**90.00**	**89.10**	**88.53**	**86.17**	**85.89**	**88.63**	**88.14**
A-OKVQA								
Vanilla	87.67	86.37	85.17	83.86	80.37	79.81	84.40	83.34
VSV Li et al. (2025)	90.17	90.51	88.83	89.32	80.40	82.67	86.47	87.50
TVS (Ours)	**92.23**	**92.21**	**89.20**	**89.47**	**82.03**	**83.62**	**87.82**	**88.44**
GQA								
Vanilla	80.97	76.93	79.53	75.92	77.43	74.11	79.31	75.65
VSV Li et al. (2025)	85.93	85.75	84.13	84.15	79.33	80.33	83.13	83.41
TVS (Ours)	**90.20**	**89.75**	**87.07**	**86.88**	**82.97**	**83.38**	**86.74**	**86.67**

C Single-Forward Implementation Audit

Table 10 compares the fused single-forward implementation with the reference two-branch implementation on LLaVA POPE-MSCOCO. Both retain the two one-time POS/NEG prefills; the distinction is the number of decoding-time model branches.

Table 10: **Single-forward implementation audit on LLaVA-1.5-7B POPE-MSCOCO** (3 splits $\times$ 3000 questions). Each split reports Accuracy/F1. TVS-1fwd preserves the full/reference implementation within 0.08 mean Accuracy points while using one decoding forward per token.

Method	Random		Popular		Adversarial		Average	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Vanilla	83.30	81.33	81.87	80.02	78.53	77.10	81.23	79.48
VSV Li et al. (2025)	89.80	–	87.17	–	82.43	–	86.47	–
TVS-2fwd (reference)	**90.20**	**89.69**	**87.43**	**87.16**	83.10	**83.46**	**86.91**	**86.77**
TVS-1fwd	89.83	89.19	87.37	86.91	**83.30**	83.39	86.83	86.50

D Mechanism Analysis: Direction Components

The component suite uses one shared evaluator with $\lambda = 0.17$, $\alpha = 2.0$, final-layer guidance and the last text token. Table 11 reports all Accuracy/F1 values. Full TVS improves mean Accuracy over the global-direction baseline by 0.30 points and over spatial-only steering by 0.77 points. Permuting the visual-token correspondence across three seeds lowers the mean by 1.09 points and the adversarial split by 1.56 points relative to Full TVS. The centered and uncentered norm-matched variants differ by only 0.06 points on average, so the current evidence supports aligned token routing more clearly than a standalone performance benefit from centering.

Table 11: **Complete direction-component results on LLaVA POPE-MSCOCO.** Each cell reports Accuracy/F1. Uncentered (NM) denotes the norm-matched uncentered control; shuffled variants differ only in permutation seed.

Direction	Random		Popular		Adversarial		Average	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
Vanilla	89.07	89.00	85.63	86.03	79.20	80.96	84.63	85.33
VSV	89.80	89.33	87.20	**86.97**	82.47	82.97	86.49	86.42
Spatial correction only	88.83	88.23	86.97	86.53	82.27	82.52	86.02	85.76
Uncentered (NM)	**89.93**	**89.39**	87.30	**86.97**	82.97	83.27	86.73	**86.55**
Shuffled seed 0	89.40	88.90	86.40	86.19	82.17	82.64	85.99	85.91
Shuffled seed 1	88.67	88.08	85.80	85.49	81.33	81.76	85.27	85.11
Shuffled seed 2	89.47	89.00	86.40	86.23	81.63	82.27	85.83	85.83
Full TVS	89.77	89.12	**87.33**	86.87	**83.27**	**83.36**	**86.79**	86.45

E Configuration and Validation Analyses

Sharpening, Source Layer and Query Position Table 12 reports the earlier sensitivity suite underlying the α, source-layer and query-position analyses in the main paper. Figure 9 visualizes the source-layer and query-position subsets. It uses LLaVA-1.5-7B and the full POPE protocol with three splits of 3000 samples each.

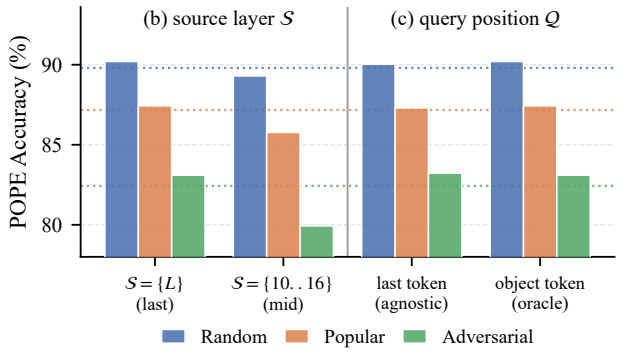

Figure 9: **Source-layer and query-position choices.** Final-layer guidance is stronger than the tested middle-layer band, while last-token and object-token queries give similar results.

Table 12: **Ablations on POPE-MSCOCO** with LLaVA-1.5-7B and λ=0.17. Changes are measured against VSV (Random 89.80, Popular 87.17, Adversarial 82.43, Average 86.47). The default is α=2.0, $\mathcal{S} = \{L\}$ and $\mathcal{Q} = \{T-1\}$.

Configuration	Random↑	Popular↑	Adversarial↑	Average↑
(a) Sharpening coefficient α ($\mathcal{S} = \{L\}$, $\mathcal{Q} = \{T-1\}$)				
$\alpha = 0.5$ (near-uniform)	89.87 ↑0.07	87.30 ↑0.13	81.97 ↓0.46	86.38 ↓0.09
$\alpha = 1.0$ ($\equiv$ raw attention weights)	90.00 ↑0.20	87.33 ↑0.16	82.70 ↑0.27	86.68 ↑0.21
$\alpha = 1.5$	90.03 ↑0.23	**87.43** ↑0.26	82.97 ↑0.54	86.81 ↑0.34
$\boldsymbol{\alpha = 2.0}$ **(default)**	90.03 ↑0.23	87.30 ↑0.13	**83.23** ↑0.80	**86.86** ↑0.39
$\alpha = 2.5$	90.03 ↑0.23	87.13 ↓0.04	82.83 ↑0.40	86.67 ↑0.20
$\alpha = 3.0$	**90.07** ↑0.27	87.23 ↑0.06	82.80 ↑0.37	86.70 ↑0.23
(b) Attention source layer $\mathcal{S}$ (α=2.0, $\mathcal{Q}$ = object)				
$\mathcal{S} = \{L\}$ (last layer)	**90.20** ↑0.40	**87.43** ↑0.26	**83.10** ↑0.67	**86.91** ↑0.44
$\mathcal{S} = \{10, \ldots, 16\}$ (middle-layer band)	89.30 ↓0.50	85.77 ↓1.40	79.93 ↓2.50	85.00 ↓1.47
(c) Attention query position $\mathcal{Q}$ (α=2.0, $\mathcal{S} = \{L\}$)				
Last token **(default, no object parser)**	90.03 ↑0.23	87.30 ↑0.13	**83.23** ↑0.80	86.86 ↑0.39
POPE object token(s)	**90.20** ↑0.40	**87.43** ↑0.26	83.10 ↑0.67	**86.91** ↑0.44
(d) $\alpha \times \mathcal{Q}$ ***interaction*** ($\mathcal{S} = \{L\}$, $\mathcal{Q}$ = object)				
$\alpha = 2.0$, object	**90.20** ↑0.40	**87.43** ↑0.26	**83.10** ↑0.67	**86.91** ↑0.44
$\alpha = 2.5$, object	90.17 ↑0.37	87.37 ↑0.20	83.03 ↑0.60	86.86 ↑0.39
$\alpha = 3.0$, object	**90.20** ↑0.40	87.37 ↑0.20	82.87 ↑0.44	86.81 ↑0.34

Benchmark-family gain Table 13 records the inherited global-steering gain sensitivity that motivates using a smaller gain on broad capability benchmarks. Each global-direction-baseline–TVS comparison nevertheless uses a matched gain.

Table 13: **Sensitivity of VSV to the injection gain** on general multimodal benchmarks with LLaVA-1.5-7B and greedy decoding. Values are rounded to one decimal where applicable; Δ denotes the score at $\lambda = 0.01$ minus that at $\lambda = 0.17$.

Benchmark	$\lambda = 0.17$	$\lambda = 0.01$	Δ
MME (overall ↑)	1617	**1752.5**	+135.5
MM-Vet (total ↑)	23.0	**26.17**	+3.17
LLaVA-Bench (all ↑)	35.3	**52.34**	+17.04

Qwen injection gain Figure 6 in the main paper visualizes the gain sensitivity and answer-calibration shift. Table 14 provides the complete Accuracy/F1/Yes-rate values.

Table 14: **TVS sensitivity to the injection gain on Qwen2.5-VL-7B-Instruct POPE-MSCOCO.** Each split reports Accuracy/F1/Yes-rate. The default $\lambda = 0.17$ provides the strongest mean Accuracy among the evaluated settings; the default row is highlighted in green.

λ	Random	Popular	Adversarial	Mean Acc.↑
0.0425	85.50 / 83.13 / 35.97	84.60 / 82.16 / 36.33	83.83 / 81.41 / 36.97	84.64
0.0850	87.33 / 85.69 / 38.53	86.70 / 85.07 / 39.10	85.30 / 83.72 / 40.30	86.44
0.1700	**90.63 / 90.00 / 43.60**	89.10 / 88.53 / 45.00	**86.17 / 85.89 / 48.00**	**88.63**
0.2550	89.93 / 89.88 / 49.47	**89.23 / 89.24 / 50.10**	84.00 / 84.83 / 55.47	87.72
0.3400	87.53 / 87.59 / 50.47	87.30 / 87.41 / 50.90	83.13 / 83.97 / 55.20	85.99

CHAIR hallucination–coverage trade-off Figure 7 in the main paper summarizes the 26 configurations on the fixed 500-image Qwen CHAIR validation set. Table 15 reports the principal coverage-preserving configurations. The selected setting is a validation result for configuration selection, with no claim about held-out test performance.

Table 15: **Length- and recall-constrained Qwen CHAIR validation.** Ratios are measured against Vanilla. Green marks the selected TVS configuration, which satisfies the pre-defined 80% length and 90% object-recall constraints.

Method	Config.	$\text{CHAIR}_S \downarrow$	$\text{CHAIR}_I \downarrow$	Length	Recall
Vanilla	–	14.40	21.54	100.0%	100.0%
VSV	$\lambda = .01$	10.20	19.59	81.1%	77.4%
TVS	$\lambda = .005, \alpha = 1$	12.80	21.22	93.3%	95.9%
TVS	$\lambda = .01, \alpha = 1$	**10.80**	**18.68**	**89.0%**	**90.3%**

High-strength steering exposes the opposite regime. Under the earlier $\lambda = 0.17$ configuration, average caption lengths were 149.6 tokens for Vanilla, 8.6 for the global-direction baseline and 0.5 for TVS. The resulting near-zero CHAIR values reflect output suppression and do not establish improved caption quality. The validation sweep therefore evaluates CHAIR jointly with length and object recall, instead of optimizing hallucination scores alone.

Efficiency timing protocol Table 16 reports prefill latency, 64-token decoding latency, throughput and peak memory for the single-forward audit. These measurements explain the decoding-overhead comparison in Fig. 8; TVS-1fwd uses one decoding forward per token while retaining the two one-time POS/NEG prefills.

Table 16: **Single-forward efficiency on LLaVA-1.5-7B** (batch size 1, n=20). Prefill and 64-token decode times are mean±std in ms; TVS-1fwd uses one decoding forward per generated token while retaining two one-time prefills.

Method	Prefill (1 tok)↓	Decode↓	Tok./s↑	Peak MB↓
Vanilla	118±31	938±203	68.3	13,941
VSV Li et al. (2025)	543±0	904±148	70.8	**13,941**
TVS-2fwd (reference)	**405±154**	1,336±214	47.9	14,529
TVS-1fwd	406±144	**828±190**	**77.3**	14,529