# CONTRASTIVE TRAJECTORY-GUIDED VISUAL EVIDENCE REINFORCEMENT FOR HALLUCINATION MITIGATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Visual hallucination remains a major obstacle to the reliable deployment of large vision-language models (LVLMs), as fluent responses may contain objects, attributes, or relations that are not adequately grounded in the input image. Existing inference-time methods commonly rely on predictive uncertainty or direct attention to visual tokens to determine when and where to intervene. However, these signals do not isolate the effect of visual access on the current token after image information has propagated through the decoder. We propose VisDelta, a training-free method that uses controlled visual-access contrast to identify where visual influence begins to diminish and selectively reinforces query-relevant image evidence. Specifically, Dual-branch Contrastive Trajectory (DCT) compares a standard decoding branch with a matched reference branch in which non-visual queries are prevented from attending to visual keys, and selects an intervention layer based on a sustained decline in their normalized current-token attention-output difference. At the selected layer, Coarse-to-Fine Evidence Reinforcement (CFER) localizes query-relevant image regions and re-encodes their original pixels with the frozen vision encoder. Contribution-Controlled Reinforcement (CCR) then injects the retrieved local features into the feed-forward computation and adaptively scales the update based on their source-region contributions. Experiments on CHAIR and POPE across multiple LVLM backbones demonstrate that VisDelta effectively reduces object hallucination and outperforms existing inference-time baselines. Results on MME and MMBench further indicate that these improvements are achieved while preserving general multimodal capabilities, and controlled ablations validate the respective contributions of layer selection, regional evidence retrieval, and contribution-controlled integration.

## 1 INTRODUCTION

Large vision-language models (LVLMs) (Liu et al., 2024b; Dai et al., 2023; Bai et al., 2025) have demonstrated strong capabilities in visual question answering, image captioning, and open-ended multimodal interaction. Despite these advances, they can generate fluent descriptions of objects, attributes, or relations that are not adequately grounded in the input image, thereby compromising the factual reliability of their responses (Rohrbach et al., 2018; Li et al., 2023; Wang et al., 2024a). Such hallucinations can hinder the safe deployment of LVLMs in high-stakes domains, including medical imaging (Nguyen et al., 2026), autonomous driving (Mahawatta Dona et al., 2025), and content moderation (Yang et al., 2026).

Existing methods mitigate hallucinations during training through robust instruction tuning, hallucination-aware data construction, and preference optimization (Liu et al., 2024a; Zhou et al., 2024; Zhao et al., 2025b). However, their reliance on additional data and parameter updates limits their direct applicability to existing LVLMs. To strengthen visual evidence during inference, some existing methods contrast predictions under different visual conditions to make better use of available visual information (Leng et al., 2024; Favero et al., 2024; Huo et al., 2025) , while others employ auxiliary perception or verification models to acquire or validate visual cues (Zhao et al., 2025a; Yin et al., 2024). However, the effectiveness of these approaches depends on the reliability



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.

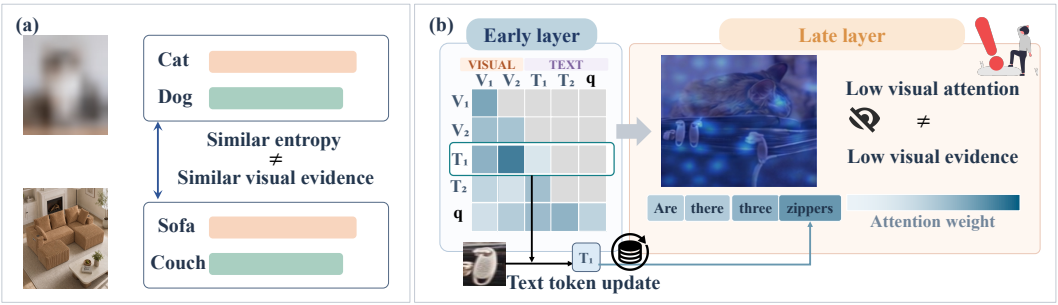


Figure 1: Two examples illustrating the limitations of entropy and visual attention as intervention signals. (a) Visual ambiguity and synonymous wording can yield similar entropy. (b) Visual information carried by text states can influence later predictions despite low direct visual attention.

of guidance signals derived either from the LVLM itself or from auxiliary models, allowing errors or biases in these signals to propagate into the final prediction.

More recent model-internal approaches avoid these dependencies by re-encoding selected image regions (Zhu et al., 2026) or amplifying existing visual representations (Liu et al., 2024c; Lei et al., 2026; He et al., 2025). Whereas, they rely heavily on attention magnitude, predictive uncertainty, or hidden states after the combination of visual and textual information which can be influenced by various factors. As illustrated in Fig. 1, predictive entropy can indicate uncertainty (Zou et al., 2025; Zhu et al., 2026), but cannot establish whether it reflects insufficient visual evidence or multiple plausible linguistic continuations (Farquhar et al., 2024). Likewise, attention to visual-token positions measures direct access to those positions (He et al., 2025; Cheng et al., 2026), not the full influence of visual information after it has propagated into text-position states (Kaduri et al., 2025).

To address these limitations, we propose VisDelta, a training-free method that reinforces visual evidence by tracing its influence across decoder layers and reinjecting features re-encoded from zoomed-in image regions. To determine where to intervene based on visual influence on the current token, VisDelta employs Dual-branch Contrastive Trajectory (DCT), which compares a standard decoding branch with a matched reference branch in which non-visual queries are prevented from attending to visual keys. It measures the normalized difference between their final-token attention outputs across decoder layers and selects an intervention layer according to a sustained decline in this trajectory. At the selected intervention layer, Coarse-to-Fine Evidence Reinforcement (CFER) locates candidate image regions using the current query, re-encodes the corresponding raw-image regions with the frozen vision encoder, and retains query-relevant local features. Its Contribution-Controlled Reinforcement (CCR) component then integrates the retrieved features into the feed-forward computation and scales the resulting update according to the concentration of the contributions associated with their source regions. Our contributions are threefold:

- We introduce Dual-branch Contrastive Trajectory, which compares matched decoding branches with and without visual-token access and uses the layer-wise trajectory of their normalized attention-output difference to select an intervention layer.
- We propose Coarse-to-Fine Evidence Reinforcement to retrieve visual evidence re-encoded from the original image using the selected-layer query, and Contribution-Controlled Reinforcement to regulate their integration based on source-region contributions.
- Experiments across multiple LVLMs and hallucination benchmarks demonstrate that VisDelta effectively reduces hallucinations and outperforms existing inference-time baselines.

## 2 Related Work

**Inference-time Hallucination Mitigation.** Inference-time hallucination mitigation guides generation by contrasting predictive distributions, perturbing visual tokens, or incorporating additional perceptual evidence. Contrastive approaches differ primarily in how they construct the reference branch. VCD and M3ID modify visual access by contrasting predictions from original and dis-

torted images or from image-conditioned and image-free inputs, respectively (Leng et al., 2024; Favero et al., 2024), whereas ICD perturbs the instruction to obtain a reference distribution (Wang et al., 2024b). Visual-token interventions offer finer control over the predictive context. SID selects visual tokens within the decoder to construct a contrastive branch (Huo et al., 2025), while Dropout Decoding masks tokens according to their estimated uncertainty and aggregates predictions across the resulting contexts (Fang et al., 2025). Complementary approaches use perceptual evidence to guide or correct generation, including auxiliary object-level cues in MARINE (Zhao et al., 2025a), classifier-based validation of candidate object tokens in TTH (Tamjidi et al., 2026), and verification and revision of generated statements in Woodpecker (Yin et al., 2024). HALC connects visual grounding with contrastive decoding through adaptive fields of view (Chen et al., 2024). These methods primarily modify output predictions or introduce additional guidance. In VisDelta, branch contrast instead serves to select an internal intervention layer, where visual evidence obtained through the target model's own vision encoder is reinforced.

**Visual enhancement in LVLMs.** Attention-based approaches amplify or calibrate visual signals using image-token weighting in PAI (Liu et al., 2024c), model confidence in CAAC (Fazli et al., 2025), and patch weighting with uncertainty gating in CAI (Lei et al., 2026). More targeted interventions enhance visual signals in middle decoder layers, as in ClearSight (Yin et al., 2025), or select attention heads according to their sensitivity to visual context, as in VHR (He et al., 2025). At the representation level, CRG estimates the effects of visual and textual routes (Cheng et al., 2026), while HulluEdit and HIRE introduce a context-dependent visual-subspace edit (Lin et al., 2026) and a learned editor with a router (Suo et al., 2026), respectively. ONLY targets a single layer using a text-to-visual entropy ratio (Wan et al., 2025), whereas MemVR reinjects existing visual tokens under predictive uncertainty (Zou et al., 2025) and AIR selects encoded image patches for reinforcement (Zhu et al., 2026). These approaches motivate two distinct considerations for intervention. Attention and uncertainty remain indirect indicators of visual influence after visual information has propagated into hidden states at text positions (Kaduri et al., 2025), while selecting patches does not itself quantify their contributions to the resulting update. VisDelta distinguishes intervention selection from evidence integration by using DCT to compare matched branches with different visual access, rather than relying solely on attention or uncertainty, CFR to retrieve query-relevant details from re-encoded image regions, and CCR to scale the resulting update according to regional contributions rather than patch selection alone.

## 3 METHOD

As illustrated in Fig. 2, VisDelta modifies autoregressive decoding through Dual-branch Contrastive Trajectory (DCT) and Coarse-to-Fine Evidence Reinforcement (CFER), while keeping the LVLM parameters frozen. Given an image $I$ and a textual prompt $T$, an LVLM with parameters $\theta$ generates token $y_t$ according to $p_\theta(y_t \mid I, T, y_{<t})$, where $y_{<t}$ denotes the generation prefix. At decoder layer $\ell$, let $\mathbf{O}_\ell \in \mathbb{R}^{N\times d}$ be the attention output before residual addition and $\mathbf{X}_\ell \in \mathbb{R}^{N\times d}$ the input to the feed-forward network (FFN), where $N$ is the context length and $d$ is the hidden dimension. DCT uses the final-token attention output $\mathbf{o}_\ell = \mathbf{O}_\ell[N,:]^\top$ to select the layer, while CFER uses the corresponding query $\mathbf{q}_\ell = \mathbf{X}_\ell[N,:]^\top$ to retrieve and integrate regional visual evidence.

### 3.1 DUAL-BRANCH CONTRASTIVE TRAJECTORY (DCT)

To determine where to reinforce visual evidence, DCT measures how the attention output changes when direct access to visual tokens is blocked. Specifically, the two branches use the same model parameters, initial input sequence, and token positions. The standard branch $B^{\text{full}}$ retains the original attention mask, whereas the reference branch $B^{\text{blk}}$ additionally prevents non-visual queries from attending to visual keys at every decoder layer. To quantify the resulting computational difference, we extract the attention output and its final-token representation from each branch $b \in \{\text{full}, \text{blk}\}$:

$$\mathbf{O}_\ell^b = \mathrm{MHA}_\ell\big(\mathbf{Q}_\ell^b, \mathbf{K}_\ell^b, \mathbf{V}_\ell^b, \mathbf{M}_\ell^b\big), \qquad \mathbf{o}_\ell^b = \mathbf{O}_\ell^b[N,:]^\top, \tag{1}$$

where $\mathbf{Q}_\ell^b$, $\mathbf{K}_\ell^b$, and $\mathbf{V}_\ell^b$ are the branch-specific projected queries, keys, and values, and $\mathbf{M}_\ell^b$ is the additive attention mask. Specifically, $\mathbf{M}_\ell^{\text{full}}$ retains the original causal and padding constraints, while $\mathbf{M}_\ell^{\text{blk}}$ additionally sets entries from non-visual queries to visual keys to $-\infty$, leaving all other entries unchanged. Because attention-output magnitudes vary across layers, we normalize the difference

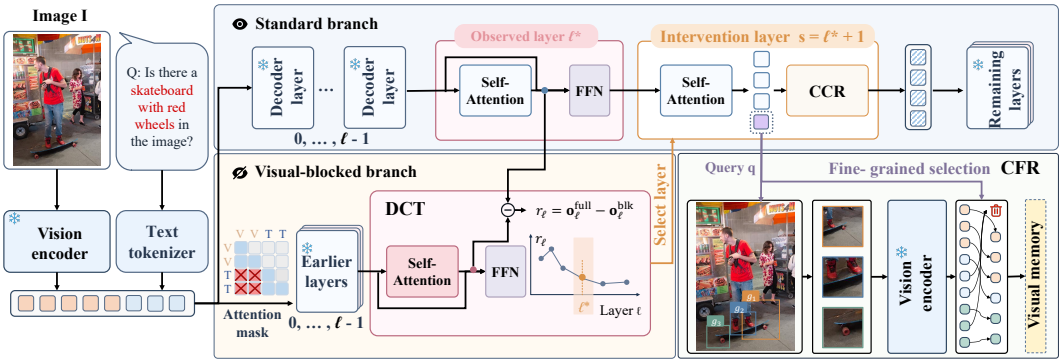


Figure 2: Overview of the VisDelta architecture. VisDelta selects the intervention layer, re-encodes relevant image regions, and controls the strength of visual reinjection.

between the two branches by the sum of their output norms:

$$r_\ell = \frac{\|\mathbf{o}_\ell^{\text{full}} - \mathbf{o}_\ell^{\text{blk}}\|_2}{s_\ell}, \qquad s_\ell = \|\mathbf{o}_\ell^{\text{full}}\|_2 + \|\mathbf{o}_\ell^{\text{blk}}\|_2 + \epsilon, \tag{2}$$

where $\epsilon > 0$ stabilizes the denominator. The resulting $r_\ell$ measures the relative change in the final-token attention output caused by blocking direct access to visual tokens.

Across decoder layers, the values of $r_\ell$ form a trajectory that DCT uses to identify a sustained decline in the effect of direct visual-token access. Because an isolated decrease may reflect a local fluctuation, we require two consecutive decreases. Each decrease must also exceed the combined reconstruction discrepancies at its adjacent layers, reducing the influence of small computational differences on layer selection. To estimate these discrepancies, we reconstruct the final-query attention output from the projected query, key, and value tensors and the attention mask of each branch:

$$\widehat{\mathbf{o}}_\ell^b = \mathrm{MHA}_\ell\big(\mathbf{Q}_\ell^b[N,:], \mathbf{K}_\ell^b, \mathbf{V}_\ell^b, \mathbf{M}_\ell^b[N,:]\big)^\top, \qquad \eta_\ell = \frac{\sum_b \|\widehat{\mathbf{o}}_\ell^b - \mathbf{o}_\ell^b\|_2}{s_\ell}. \tag{3}$$

Here, $[N,:]$ selects the final query and its mask row, retaining the singleton query dimension and all attention heads, while $\mathbf{K}_\ell^b$ and $\mathbf{V}_\ell^b$ retain the full context. The reconstruction includes the output projection and bias. Normalization by $s_\ell$ places the resulting tolerance $\eta_\ell$ on the same scale as $r_\ell$.

We use the standard branch's intermediate prediction entropy to define the eligible layers $\mathcal{E} = \{\ell \in \{1, \dots, L-1\} : H_\ell \geq 0.75\}$. Here, $H_\ell$ is the entropy of the softmax distribution over the ten largest intermediate next-token logits, normalized by $\log 10$. Among eligible layers, we retain those satisfying both trajectory comparisons:

$$\mathcal{A} = \{\ell \in \mathcal{E} : \ell \geq 3, \quad r_{j-1} - r_j > \eta_{j-1} + \eta_j \text{ for } j \in \{\ell - 1, \ell\}\}. \tag{4}$$

The following intervention is applied one layer after the earliest layer satisfying the trajectory criterion and no intervention is performed when $\mathcal{A}$ is empty.

At the selected intervention layer $\ell_{\text{int}}$, the current query $\mathbf{q}_{\ell_{\text{int}}}$ guides the retrieval and integration of regional visual evidence to mitigate the potential for hallucination.

### 3.2 Coarse-to-Fine Evidence Reinforcement

At the layer selected by DCT, Coarse-to-Fine Evidence Reinforcement (CFER) performs two operations. Coarse-to-Fine Retrieval (CFR) locates relevant regions in the original image and selects visual tokens from their re-encoded crops. Contribution-Controlled Reinforcement (CCR) then integrates these tokens into the decoder computation.

**Coarse-to-Fine Retrieval (CFR).** At the coarse stage, CFR scores the original projected visual tokens $\mathbf{V}^{(0)} \in \mathbb{R}^{N_v \times d}$ against the current query $\mathbf{q} = \mathbf{q}_{\ell_{\text{int}}}$ to identify regions for closer inspection:

$$\mathbf{s}^{\text{loc}} = \left|\mathbf{V}^{(0)}\mathbf{q}\right|, \qquad \mathcal{S}_{\text{loc}} = \mathrm{TopK}\big(\mathbf{s}^{\text{loc}}, \min(K_{\text{loc}}, N_v)\big), \tag{5}$$

where $\mathbf{s}^{\text{loc}} \in \mathbb{R}^{N_v}$ is the score vector and the absolute value is applied elementwise. The operator $\text{TopK}(\mathbf{s}, k)$ returns the set of indices of the $k$ largest entries of $\mathbf{s}$, and $K_{\text{loc}}$ specifies the maximum number of global visual tokens selected for regional localization. The absolute inner product retains strong interactions of either sign, consistent with the signed memory operation used below. To convert the selected token indices into image regions, we group the corresponding patches by spatial connectivity, yielding $G$ candidate groups. For each group $g$ in $G$, we map its bounding box to the original image and expand it to include surrounding context, defining the candidate region $\mathcal{R}_g$. Details of the grouping and expansion are provided in Appendix C.

At the fine-grained stage, we crop the localized regions from the original image and re-encode each crop at the visual encoder's native input resolution and settings, allocating greater spatial resolution to the selected image content. The same query then ranks the newly encoded visual tokens to select relevant details within these regions:

$$\mathbf{V}_g = P_{\text{vis}}\big(E_{\text{vis}}(\text{Crop}(I, \mathcal{R}_g))\big), \qquad s^{\text{ret}}_{g,j} = |\mathbf{q}^\top \mathbf{v}_{g,j}|, \tag{6}$$

where $E_{\text{vis}}$ and $P_{\text{vis}}$ denote the frozen visual encoder with standard image preprocessing and multimodal projector, respectively, and $\mathbf{V}_g \in \mathbb{R}^{n_g \times d}$ contains the $n_g$ visual tokens encoded from region $\mathcal{R}_g$, with $\mathbf{v}_{g,j}$ denoting its $j$-th visual token. Given a visual-token budget $K_{\text{ret}}$, we retain the $N_{\text{ret}} = \min(K_{\text{ret}}, \sum_{g=1}^{G} n_g)$ highest-scoring visual tokens across all regions and restore their source order to form the final visual memory $\mathbf{V}_{\text{ret}} \in \mathbb{R}^{N_{\text{ret}} \times d}$. Each selected token retains its source-region identifier, enabling the subsequent Contribution-Controlled Reinforcement (CCR) module to quantify regional contributions during evidence aggregation.

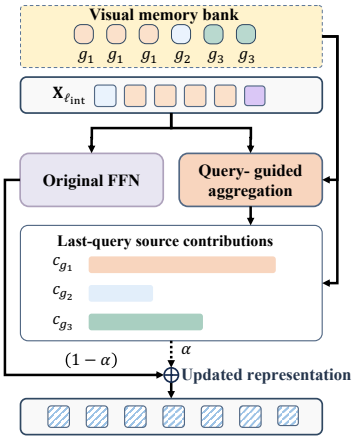


Figure 3: Structure of the Contribution-Controlled Reinforcement (CCR).

**Contribution-Controlled Reinforcement (CCR).** As illustrated in Fig. 3, CCR aggregates the retrieved visual memory into evidence for the current decoder computation and regulates the overall injection strength according to regional contributions. To obtain this evidence, we use the FFN input $\mathbf{X}_{\ell_{\text{int}}}$ as queries and $\mathbf{V}_{\text{ret}}$ as both keys and values. The query-key inner products weight the visual values, yielding an evidence representation for each context position:

$$\widetilde{\mathbf{M}} = \gamma_{\ell_{\text{int}}} \mathbf{X}_{\ell_{\text{int}}} \mathbf{V}_{\text{ret}}^\top \mathbf{V}_{\text{ret}}, \qquad \gamma_{\ell_{\text{int}}} = \frac{\mu(\mathbf{W}^{\text{up}}_{\ell_{\text{int}}})\mu(\mathbf{W}^{\text{down}}_{\ell_{\text{int}}})}{\mu(\mathbf{V}_{\text{ret}})^2}, \tag{7}$$

where $\mu(\cdot)$ denotes the mean absolute matrix entry, and $\mathbf{W}^{\text{up}}_{\ell_{\text{int}}}$ and $\mathbf{W}^{\text{down}}_{\ell_{\text{int}}}$ are the frozen FFN up- and down-projection matrices, respectively. $\gamma_{\ell_{\text{int}}}$ calibrates the scale of visual aggregation using the mean absolute magnitudes of the FFN weights.

The query-guided aggregation weights the retrieved visual evidence according to the current query, potentially concentrating regional contributions in a small subset of sources. To characterize this concentration, we decompose the aggregated evidence by source region. Let $\mathbf{V}^{(g)}_{\text{ret}}$ denote the retained visual tokens originating from region $g$, whose contribution is given by

$$\mathbf{c}_g = \gamma_{\ell_{\text{int}}} (\mathbf{V}^{(g)}_{\text{ret}})^\top \mathbf{V}^{(g)}_{\text{ret}} \mathbf{q}. \tag{8}$$

To quantify the dominance of a single region, we define the contribution concentration $D$ as the largest regional contribution norm relative to the sum of all regional contribution norms:

$$D = \frac{\max_{1 \le g \le G} \|\mathbf{c}_g\|_2}{\sum_{g=1}^{G} \|\mathbf{c}_g\|_2 + \epsilon}. \tag{9}$$

Here, $\epsilon$ stabilizes the denominator. To limit the overall influence of evidence dominated by one source when several regions contribute, we apply a conservative attenuation rule. The interpolation coefficient is $\alpha = \alpha_1(1 - D) + \alpha_2$, where $\alpha_1$ scales the concentration-dependent term and $\alpha_2$ provides an additive baseline. A region is considered active when $\|\mathbf{c}_g\|_2 > \epsilon$.

Table 1: CHAIR evaluation results on MS COCO val2014 with LLaVA-1.5-7B and InstructBLIP-7B. We report mean $\pm$std across three 500-image subsets at 64- and 512-token generation limits.

| Model | Method | Max new tokens set to 64 | | | | Max new tokens set to 512 | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | $CHAIR_S \downarrow$ | $CHAIR_I \downarrow$ | Recall↑ | Length | $CHAIR_S \downarrow$ | $CHAIR_I \downarrow$ | Recall↑ | Length |
| LLaVA-1.5 | Greedy | 21.80 $\pm$1.25 | 7.16 $\pm$0.45 | 63.82 $\pm$0.51 | 54.45 $\pm$0.20 | 50.33 $\pm$4.03 | 15.08 $\pm$1.11 | 78.57 $\pm$0.93 | 92.35 $\pm$0.92 |
| | OPERA | 22.80 $\pm$3.99 | 7.21 $\pm$1.24 | 62.32 $\pm$1.16 | 52.64 $\pm$0.09 | 45.93 $\pm$3.00 | 14.36 $\pm$1.02 | 77.11 $\pm$0.76 | 89.61 $\pm$0.88 |
| | ICD | 23.27 $\pm$0.64 | 7.69 $\pm$0.87 | 62.83 $\pm$1.08 | 54.43 $\pm$0.19 | 41.93 $\pm$3.03 | 12.95 $\pm$1.54 | 75.26 $\pm$0.16 | 84.51 $\pm$1.18 |
| | VCD | 25.93 $\pm$0.42 | 8.94 $\pm$0.45 | 61.41 $\pm$1.42 | 53.87 $\pm$0.17 | 52.20 $\pm$1.74 | 16.37 $\pm$1.05 | 73.86 $\pm$1.62 | 91.47 $\pm$0.33 |
| | DD | 23.93 $\pm$1.03 | 8.37 $\pm$0.46 | 63.01 $\pm$0.43 | 54.55 $\pm$0.14 | 51.73 $\pm$1.21 | 13.82 $\pm$0.25 | 76.07 $\pm$1.21 | 98.94 $\pm$1.05 |
| | MemVR | 16.60 $\pm$1.04 | 5.51 $\pm$0.22 | 50.92 $\pm$1.10 | 51.11 $\pm$0.21 | 43.53 $\pm$2.93 | 12.18 $\pm$0.99 | 79.33 $\pm$0.71 | 97.72 $\pm$0.33 |
| | VAF | 19.73 $\pm$1.47 | 6.60 $\pm$0.10 | 61.79 $\pm$1.44 | 52.56 $\pm$0.11 | 45.07 $\pm$1.10 | 12.78 $\pm$0.27 | 78.14 $\pm$1.65 | 88.62 $\pm$1.40 |
| | AIR | 20.00 $\pm$0.53 | 7.09 $\pm$0.39 | 62.26 $\pm$0.98 | 54.74 $\pm$0.13 | 49.13 $\pm$0.46 | 14.22 $\pm$0.36 | 78.36 $\pm$0.92 | 99.04 $\pm$1.19 |
| | HulluEdit | 17.93 $\pm$0.81 | 5.92 $\pm$0.19 | 62.68 $\pm$0.67 | 53.49 $\pm$0.34 | 40.67 $\pm$2.95 | 12.49 $\pm$1.52 | 75.03 $\pm$0.16 | 83.92 $\pm$1.17 |
| | VisDelta | **16.20** $\pm$1.31 | **5.31** $\pm$0.22 | **64.17** $\pm$0.42 | 53.05 $\pm$0.24 | **37.87** $\pm$1.45 | **7.98** $\pm$0.30 | **82.18** $\pm$2.24 | 99.58 $\pm$3.35 |
| InstructBLIP | Greedy | 24.73 $\pm$1.29 | 8.42 $\pm$0.28 | 63.66 $\pm$0.68 | 56.58 $\pm$0.13 | 48.07 $\pm$1.67 | 14.00 $\pm$0.73 | 73.09 $\pm$1.19 | 104.23 $\pm$0.46 |
| | OPERA | 22.87 $\pm$3.89 | 7.16 $\pm$2.04 | 62.22 $\pm$0.76 | 52.88 $\pm$0.14 | 46.40 $\pm$3.03 | 14.37 $\pm$1.04 | 76.64 $\pm$0.5 | 88.89 $\pm$0.90 |
| | ICD | 21.67 $\pm$1.70 | 7.61 $\pm$0.33 | 55.89 $\pm$0.82 | 55.28 $\pm$0.12 | 45.40 $\pm$2.25 | 13.76 $\pm$0.69 | 68.18 $\pm$0.36 | 106.01 $\pm$2.90 |
| | VCD | 27.73 $\pm$0.64 | 9.69 $\pm$0.51 | 61.10 $\pm$0.48 | 55.44 $\pm$0.10 | 55.20 $\pm$2.62 | 16.19 $\pm$0.14 | 72.56 $\pm$0.60 | 109.51 $\pm$1.02 |
| | DD | 14.80 $\pm$0.80 | 8.64 $\pm$0.64 | 64.80 $\pm$0.48 | 56.60 $\pm$0.15 | 48.27 $\pm$1.22 | 14.22 $\pm$0.70 | 72.91 $\pm$0.75 | 102.49 $\pm$0.93 |
| | MemVR | 11.80 $\pm$1.00 | 5.52 $\pm$0.37 | 64.50 $\pm$0.75 | 57.80 $\pm$0.14 | 41.33 $\pm$1.51 | 8.70 $\pm$0.82 | 77.02 $\pm$2.85 | 103.28 $\pm$1.88 |
| | VAF | 12.60 $\pm$0.92 | 6.21 $\pm$0.39 | 64.64 $\pm$0.84 | 58.04 $\pm$0.13 | 38.87 $\pm$1.42 | 8.74 $\pm$0.78 | 73.71 $\pm$2.99 | 106.58 $\pm$1.87 |
| | AIR | 11.13 $\pm$0.95 | 5.95 $\pm$0.37 | 66.30 $\pm$0.77 | 57.92 $\pm$0.14 | 39.93 $\pm$1.33 | 8.48 $\pm$0.78 | 80.40 $\pm$2.99 | 106.57 $\pm$1.86 |
| | HulluEdit | 11.87 $\pm$0.95 | 6.41 $\pm$0.36 | 65.11 $\pm$0.74 | 58.98 $\pm$0.13 | 39.13 $\pm$1.33 | 8.24 $\pm$0.82 | 77.16 $\pm$2.81 | 107.12 $\pm$1.83 |
| | VisDelta | **10.53** $\pm$0.81 | **5.21** $\pm$0.33 | **66.77** $\pm$0.67 | 60.73 $\pm$0.13 | **36.93** $\pm$1.21 | **8.08** $\pm$0.75 | **81.91** $\pm$2.68 | 105.28 $\pm$2.43 |

To align the magnitude of the aggregated evidence with that of the FFN output, we normalize the evidence before interpolation:

$$\mathbf{M} = \frac{\mu(F_{\ell_{\text{int}}}(\mathbf{X}_{\ell_{\text{int}}}))}{\mu(\widetilde{\mathbf{M}})}\widetilde{\mathbf{M}}, \qquad \mathbf{Y}_{\ell_{\text{int}}} = (1-\alpha)F_{\ell_{\text{int}}}(\mathbf{X}_{\ell_{\text{int}}}) + \alpha\mathbf{M}. \tag{10}$$

Here, $F_\ell$ denotes the frozen FFN at decoder layer $\ell$. The updated representation $\mathbf{Y}_{\ell_{\text{int}}}$ replaces the original FFN output, incorporating the retrieved visual evidence into subsequent decoder computation with an overall strength determined by regional contributions.

# 4 Experiments

## 4.1 Experimental Setup

**Base models and baselines.** To assess the applicability of VisDelta across different model families, we use LLaVA-1.5-7B (Liu et al., 2024b), InstructBLIP (Dai et al., 2023), Qwen-VL (Bai et al., 2023), and GLM-4V-9B (Team GLM, 2024) as base models. To evaluate its effectiveness in mitigating object hallucination, we compare it with greedy decoding, OPERA (Huang et al., 2024), VCD (Leng et al., 2024), ICD (Wang et al., 2024b), AGLA An et al. (2025), MemVR (Zou et al., 2025), Dropout Decoding (DD) (Fang et al., 2025), VAF (Yin et al., 2025), AIR (Zhu et al., 2026), CRG (Cheng et al., 2026), and HulluEdit (Lin et al., 2026).

**Benchmarks and metrics.** We assess hallucination using CHAIR (Rohrbach et al., 2018) on MS COCO (Lin et al., 2014) and POPE (Li et al., 2023) on MS COCO, A-OKVQA (Schwenk et al., 2022), and GQA (Hudson & Manning, 2019), and evaluate general capabilities using MME (Fu et al., 2023) and MMBench (Liu et al., 2024d). We report $\text{CHAIR}_S$, $\text{CHAIR}_I$, recall, and mean description length for CHAIR, accuracy and F1-score for POPE, hallucination and cognition scores for MME, and accuracy for MMBench. More benchmark descriptions and detailed settings are provided in Appendices A and B.

## 4.2 Comparison with State-of-the-Art Methods

**CHAIR.** CHAIR measures object hallucination in free-form captions through sentence- and instance-level error rates (Rohrbach et al., 2018). Table 1, VisDelta achieves the lowest mean

Table 2: POPE evaluation results on MS COCO, A-OKVQA, and GQA using LLaVA-1.5-7B. We report accuracy and F1 for Random, Popular, and Adversarial settings and their averages.

| Dataset | Method | Random | | Popular | | Adversarial | | Average | |
|---|---|---|---|---|---|---|---|---|---|
| | | Acc. ↑ | F1 ↑ | Acc. ↑ | F1 ↑ | Acc. ↑ | F1 ↑ | Acc. ↑ | F1 ↑ |
| MSCOCO | LLaVA-1.5 | 83.49(0.0) | 82.28(0.0) | 79.98(0.0) | 79.34(0.0) | 76.03(0.0) | 76.26(0.0) | 79.83(0.0) | 79.29(0.0) |
| | OPERA | 84.87(+1.4) | 83.22(+0.9) | 83.00(+3.0) | 81.08(+1.7) | 80.28(+4.3) | 80.09(+3.8) | 82.72(+2.9) | 81.46(+2.2) |
| | ICD | 84.59(+1.1) | 82.17(-0.1) | 82.65(+2.7) | 80.48(+1.1) | 80.77(+4.7) | 79.36(+3.1) | 82.67(+2.8) | 80.67(+1.4) |
| | VCD | 85.84(+2.4) | 84.82(+2.5) | 81.31(+1.3) | 82.03(+2.7) | 77.11(+1.1) | 78.68(+2.4) | 81.42(+1.6) | 81.84(+2.6) |
| | DD | 84.87(+1.4) | 84.87(+2.6) | 85.03(+5.1) | 83.61(+4.3) | 78.47(+2.4) | 81.03(+4.8) | 82.79(+3.0) | 83.17(+3.9) |
| | MemVR | 85.17(+1.7) | 85.84(+3.6) | 83.59(+3.6) | 82.50(+3.2) | 84.73(+8.7) | 83.08(+6.8) | 84.50(+4.7) | 83.81(+4.5) |
| | VAF | 88.15(+4.7) | **88.14**(+5.9) | 83.71(+3.7) | 85.51(+6.2) | 82.87(+6.8) | 82.30(+6.0) | 84.91(+5.1) | 85.32(+6.0) |
| | AIR | 86.15(+2.7) | 87.10(+4.8) | 84.35(+4.4) | 84.20(+4.9) | 83.20(+7.2) | 83.10(+6.8) | 84.57(+4.7) | 84.80(+5.5) |
| | HulluEdit | 88.27(+4.8) | 87.59(+5.3) | 83.40(+3.4) | 85.18(+5.8) | 79.80(+3.8) | 81.69(+5.4) | 83.82(+4.0) | 84.82(+5.5) |
| | CRG | 86.84(+3.4) | 85.42(+3.1) | 86.06(+6.1) | 85.86(+6.5) | 83.77(+7.7) | 83.42(+7.2) | 85.56(+5.7) | 84.90(+5.6) |
| | VisDelta | **88.97**(+5.5) | 88.06(+5.8) | **87.53**(+7.6) | **86.72**(+7.4) | **84.97**(+8.9) | **84.41**(+8.1) | **87.16**(+7.3) | **86.40**(+7.1) |
| A-OKVQA | LLaVA-1.5 | 83.45(0.0) | 82.56(0.0) | 79.90(0.0) | 79.59(0.0) | 74.04(0.0) | 75.15(0.0) | 79.13(0.0) | 79.10(0.0) |
| | OPERA | 87.12(+3.7) | 87.10(+4.5) | 85.60(+5.7) | 84.40(+4.8) | 76.85(+2.8) | 72.77(-2.4) | 83.19(+4.1) | 81.42(+2.3) |
| | ICD | 84.89(+1.4) | 85.83(+3.3) | 81.27(+1.4) | 82.36(+2.8) | 77.35(+3.3) | 79.15(+4.0) | 81.17(+2.0) | 82.44(+3.3) |
| | VCD | 85.46(+2.0) | 85.48(+2.9) | 81.03(+1.1) | 82.73(+3.1) | 74.90(+0.9) | 77.65(+2.5) | 80.46(+1.3) | 81.95(+2.9) |
| | DD | 86.17(+2.7) | 87.33(+4.8) | 78.86(-1.0) | 81.29(+1.7) | 67.03(-7.0) | 74.47(-0.7) | 77.35(-1.8) | 81.03(+1.9) |
| | MemVR | 90.01(+6.6) | 90.10(+7.5) | 84.64(+4.7) | 86.04(+6.4) | 77.81(+3.8) | 79.15(+4.0) | 84.15(+5.0) | 85.09(+6.0) |
| | VAF | 89.13(+5.7) | 88.39(+5.8) | 82.85(+3.0) | 84.26(+4.7) | 77.97(+3.9) | 79.75(+4.6) | 83.32(+4.2) | 84.13(+5.0) |
| | AIR | 88.93(+5.5) | 88.46(+5.9) | 83.56(+3.7) | 85.89(+6.3) | 78.23(+4.2) | 78.98(+3.8) | 83.57(+4.4) | 84.44(+5.3) |
| | HulluEdit | 86.76(+3.3) | 88.10(+5.5) | 80.73(+0.8) | 82.76(+3.2) | 70.57(-3.5) | 76.55(+1.4) | 79.35(+0.2) | 82.47(+3.4) |
| | CRG | 88.95(+5.5) | 88.75(+6.2) | 84.82(+4.9) | 84.95(+5.4) | 76.81(+2.8) | 78.28(+3.1) | 83.53(+4.4) | 83.99(+4.9) |
| | VisDelta | **90.87**(+7.4) | **90.73**(+8.2) | **85.87**(+6.0) | **86.35**(+6.8) | **78.67**(+4.6) | **80.73**(+5.6) | **85.14**(+6.0) | **85.94**(+6.8) |
| GQA | LLaVA-1.5 | 83.73(0.0) | 82.95(0.0) | 78.17(0.0) | 78.37(0.0) | 75.08(0.0) | 76.06(0.0) | 78.99(0.0) | 79.13(0.0) |
| | OPERA | 82.75(-1.0) | 83.21(+0.3) | 78.20(+0.0) | 77.98(-0.4) | 74.07(-1.0) | 77.62(+1.6) | 78.34(-0.7) | 79.60(+0.5) |
| | ICD | 84.55(+0.8) | 84.64(+1.7) | 78.37(+0.2) | 79.20(+0.8) | 74.53(-0.5) | 77.16(+1.1) | 79.15(+0.2) | 80.33(+1.2) |
| | VCD | 86.28(+2.5) | 86.47(+3.5) | 80.73(+2.6) | 80.84(+2.5) | 74.95(-0.1) | 77.91(+1.8) | 80.65(+1.7) | 81.74(+2.6) |
| | DD | 85.30(+1.6) | 86.89(+3.9) | 81.80(+3.6) | 81.98(+3.6) | 66.96(-8.1) | 74.36(-1.7) | 78.02(-1.0) | 81.08(+2.0) |
| | MemVR | 89.24(+5.5) | 89.14(+6.2) | 82.20(+4.0) | 82.79(+4.4) | 78.87(+3.8) | 79.62(+3.6) | 83.44(+4.4) | 83.85(+4.7) |
| | VAF | 87.23(+3.5) | 87.62(+4.7) | 79.69(+1.5) | 81.33(+3.0) | 78.19(+3.1) | 79.50(+3.4) | 81.70(+2.7) | 82.82(+3.7) |
| | AIR | 89.30(+5.6) | 88.98(+6.0) | 81.43(+3.3) | 82.87(+4.5) | 78.56(+3.5) | 79.82(+3.8) | 83.10(+4.1) | 83.89(+4.8) |
| | HulluEdit | 86.54(+2.8) | 87.15(+4.2) | 79.31(+1.1) | 81.89(+3.5) | 78.89(+3.8) | 78.16(+2.1) | 81.58(+2.6) | 82.40(+3.3) |
| | CRG | 86.29(+2.6) | 86.98(+4.0) | 79.53(+1.4) | 81.28(+2.9) | 78.00(+2.9) | 78.94(+2.9) | 81.27(+2.3) | 82.40(+3.3) |
| | VisDelta | **89.60**(+5.9) | **89.50**(+6.6) | **82.73**(+4.6) | **83.70**(+5.3) | **80.23**(+5.2) | **81.77**(+5.7) | **84.19**(+5.2) | **84.99**(+5.9) |

$\text{CHAIR}_S$ and $\text{CHAIR}_I$ and the highest mean object recall among the evaluated methods for both backbones under both generation limits. At the 64-token limit, VisDelta reduces InstructBLIP's $\text{CHAIR}_S$ from 24.73% to 10.53% while increasing object recall from 63.66% to 66.77% relative to Greedy. At the 512-token limit, the lower hallucination rates and higher object recall are accompanied by longer captions on average than Greedy for both backbones.

**POPE.** POPE tests object hallucination through balanced yes/no questions with random, popular, and adversarial negatives. As shown in Table 2, VisDelta achieves the highest average accuracy and F1-score on MS COCO, A-OKVQA, and GQA using LLaVA-1.5-7B. Average F1-score increases from 79.29% to 86.40% on MS COCO, from 79.10% to 85.94% on A-OKVQA, and from 79.13% to 84.99% on GQA relative to the baseline, although VAF retains a slight advantage on MS COCO Random F1-score at 88.14% compared with 88.06% for VisDelta. Our method also achieves the highest accuracy and F1-score on the adversarial split of each dataset. These gains over contrastive decoding methods, visual reinforcement methods and other methods based on routing or representations are consistent with the benefit of reinforcing query-relevant visual evidence. Values in red denote differences in percentage points relative to the LLaVA-1.5 baseline for each dataset.

**MME and MMBench.** MME-Hall evaluates object and attribute hallucinations, while MME cognition and MMBench assess broader multimodal capabilities. As shown in Table 3, VisDelta achieves the highest MME-Hall total on all three backbones, with 10-point gains over Vanilla on LLaVA-1.5 and Qwen-VL and a smaller increase on GLM-4V. In comparison, VCD lowers MME-Hall scores across all three backbones, whereas MemVR, VAF, and AIR yield smaller gains or

Table 3: Comparison on MME hallucination and cognition scores and MMBench accuracy. Bold marks a unique best score within each model block.

| Model | Method | MME-Hall | Object-Level | | Attribute-Level | | Cognition | MMBench |
|---|---|---|---|---|---|---|---|---|
| | | Total ↑ | Existence ↑ | Count ↑ | Position ↑ | Color ↑ | Score ↑ | Accuracy ↑ |
| LLaVA-1.5 | Vanilla | 643.3 (0.0) | 190.0 (0.0) | 155.0 (0.0) | 128.3 (0.0) | 170.0 (0.0) | 357.8 (0.0) | 64.6 (0.0) |
| | VCD | 613.3 (-30.0) | 190.0 (0.0) | 140.0 (-15.0) | 118.3 (-10.0) | 165.0 (-5.0) | 337.1 (-20.7) | 61.6 (-3.0) |
| | MemVR | 648.3 (+5.0) | 190.0 (0.0) | 155.0 (0.0) | 133.3 (+5.0) | 170.0 (0.0) | 378.6 (+20.8) | 64.6 (0.0) |
| | VAF | 603.3 (-40.0) | 190.0 (0.0) | 135.0 (-20.0) | 108.3 (-20.0) | 170.0 (0.0) | 322.8 (-35.0) | 14.8 (-49.8) |
| | AIR | 638.3 (-5.0) | 190.0 (0.0) | 155.0 (0.0) | 123.3 (-5.0) | 170.0 (0.0) | 372.5 (+14.7) | 64.7 (+0.1) |
| | VisDelta | **653.3** (+10.0) | 190.0 (0.0) | **160.0** (+5.0) | 133.3 (+5.0) | 170.0 (0.0) | **382.1** (+24.3) | 64.7 (+0.1) |
| Qwen-VL | Vanilla | 631.7 (0.0) | 185.0 (0.0) | 145.0 (0.0) | 126.7 (0.0) | 175.0 (0.0) | 342.8 (0.0) | 59.9 (0.0) |
| | VCD | 626.7 (-5.0) | 180.0 (-5.0) | 145.0 (0.0) | 126.7 (0.0) | 175.0 (0.0) | 348.9 (+6.1) | 60.3 (+0.4) |
| | MemVR | 636.7 (+5.0) | 185.0 (0.0) | 145.0 (0.0) | **131.7** (+5.0) | 175.0 (0.0) | 337.5 (-5.3) | 60.0 (+0.1) |
| | VAF | 631.7 (0.0) | 185.0 (0.0) | 145.0 (0.0) | 126.7 (0.0) | 175.0 (0.0) | 329.3 (-13.5) | 60.0 (+0.1) |
| | AIR | 636.7 (+5.0) | 185.0 (0.0) | 145.0 (0.0) | 126.7 (0.0) | 180.0 (+5.0) | **352.1** (+9.3) | 60.0 (+0.1) |
| | VisDelta | **641.7** (+10.0) | 185.0 (0.0) | **150.0** (+5.0) | 126.7 (0.0) | 180.0 (+5.0) | 351.8 (+9.0) | **63.8** (+3.9) |
| GLM-4V | Vanilla | 703.3 (0.0) | 200.0 (0.0) | 168.3 (0.0) | 156.7 (0.0) | 178.3 (0.0) | 479.6 (0.0) | 81.3 (0.0) |
| | VCD | 696.7 (-6.6) | 200.0 (0.0) | 170.0 (+1.7) | 151.7 (-5.0) | 175.0 (-3.3) | **485.0** (+5.4) | 80.2 (-1.1) |
| | MemVR | 703.3 (0.0) | 200.0 (0.0) | 168.3 (0.0) | 156.7 (0.0) | 178.3 (0.0) | 479.6 (0.0) | 81.3 (0.0) |
| | VAF | 703.3 (0.0) | 200.0 (0.0) | 173.3 (+5.0) | 151.7 (-5.0) | 178.3 (0.0) | 479.6 (0.0) | 81.3 (0.0) |
| | AIR | 703.3 (0.0) | 200.0 (0.0) | 173.3 (+5.0) | 151.7 (-5.0) | 178.3 (0.0) | 479.6 (0.0) | 81.3 (0.0) |
| | VisDelta | **705.0** (+1.7) | 200.0 (0.0) | **175.0** (+6.7) | 151.7 (-5.0) | 178.3 (0.0) | 479.6 (0.0) | 81.3 (0.0) |

mixed results. At the subtask level, counting scores improve consistently across backbones, while object-existence scores remain unchanged. These gains are accompanied by preserved or improved performance on MME cognition. Specifically, VisDelta achieves the highest reported cognition score on LLaVA-1.5, outperforms Vanilla and nearly matches AIR on Qwen-VL, and maintains the Vanilla score on GLM-4V. Together, these results suggest that the aggregate gains in hallucination-related perception are achieved without a broad decline in the capabilities evaluated here.

### 4.3 Ablation Study

**Impact of Main Components.** To assess the contribution of each design choice, we evaluate three component replacements on LLaVA-1.5-7B and the results are shown in Table 4. We replace DCT layer selection with a fixed layer, CFR's re-encoded crop features with the original visual tokens, and CCR's contribution-dependent interpolation coefficient with a fixed value. Each replacement worsens all five reported metrics. Using the original visual tokens produces the largest degradation, reducing POPE average F1 from 85.77% to 83.14% and increasing $\mathrm{CHAIR}_S$ from 16.20% to 19.80%. The fixed-layer and fixed-coefficient variants reduce POPE average F1 by 1.74 and 0.99 percentage points, respectively. These comparisons support the benefit of each design choice relative to its tested replacement on this backbone.

**Parameter Sensitivity.** To examine how reduced budgets and reinforcement strength affect performance, we halve the localization budget, retrieval budget, and maximum reinforcement strength one at a time on LLaVA-1.5-7B (Table 4). Relative to the common configuration, all three reductions lower POPE average accuracy and F1 and increase both CHAIR hallucination rates, with the largest F1 drop occurring when the localization budget is halved. Additional results for the localization budget are provided in Appendix D.

Table 4: Abalation study on LLaVA-1.5-7B.

| Variant | MME-Hall Total ↑ | POPE Avg. Acc./F1 ↑ | CHAIR $I/S$ ↓ |
|---|---|---|---|
| Base | 643.3 | 79.32/79.17 | 7.16/21.80 |
| DCT → fixed | 648.3 | 83.92/84.03 | 5.92/19.73 |
| CFR → original | 645.0 | 82.88/83.14 | 6.41/19.80 |
| CCR → fixed $\alpha$ | 650.0 | 84.61/84.78 | 5.67/16.60 |
| VisDelta | **653.3** | **85.49/85.77** | **5.31/16.20** |
| $K_{\mathrm{loc}}$ (64 → 32) | | 84.18/84.45 | 5.47/16.67 |
| $K_{\mathrm{ret}}$ (576 → 288) | | 84.95/85.23 | 5.47/16.67 |
| $\alpha_1$ (0.18 → 0.09) | | 85.21/85.49 | 5.42/16.47 |

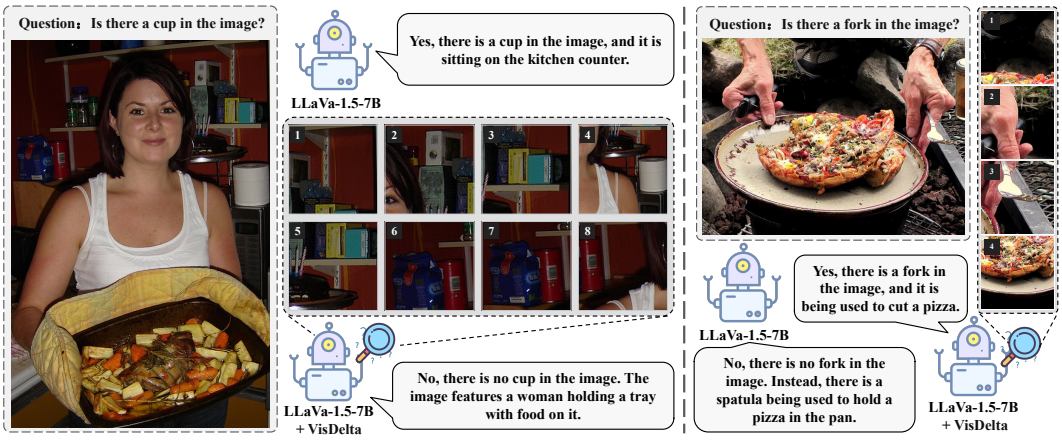


Figure 4: Qualitative examples of object-existence reasoning with retrieved image regions.

### 4.4 Qualitative Analysis

**Object-Presence Cases.** To illustrate how regional evidence informs object-existence judgments, Figure 4 presents two questions that require local visual inspection. In the cup example, the kitchen context makes an affirmative answer plausible despite the absence of the queried object. In the fork example, distinguishing similar utensils requires close inspection of a cluttered scene.

**DCT Layer Selection.** To examine how the entropy condition and trajectory criterion contribute to DCT layer selection, Figure 5 reports gate status across intervention layer groups. DCT selects the earliest qualifying observation layer and intervenes at the following layer. The plotted proportions therefore characterize gate status rather than the distribution of selected intervention layers. Entropy-only proportions use the left axis, whereas trajectory-only and joint proportions use the right axis. Entropy-only cases peak at layers 12–13 and decline at layers 14–16, whereas trajectory-only cases increase across the displayed groups and are most frequent at layers 14–16. Joint satisfaction remains low across all displayed groups. The prevalence of entropy-only cases at layers 10–13 shows that the trajectory criterion excludes many candidates admitted by the entropy condition alone. Since DCT requires both conditions, the increasing prevalence of the trajectory criterion alone does not imply more interventions at later layers. These patterns show that the two conditions impose distinct, complementary constraints on intervention eligibility.

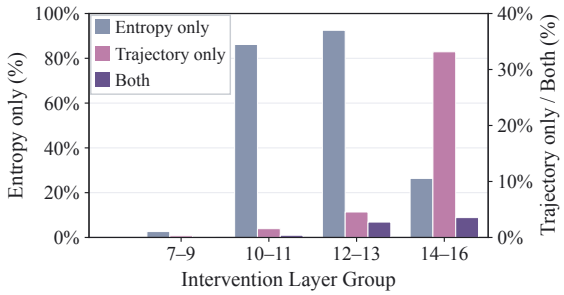


Figure 5: DCT entropy and trajectory gate satisfaction across candidate intervention layer groups.

## 5 Conclusion

We presented VisDelta, a training-free method that uses controlled visual-access contrast to identify where visual influence begins to diminish and selectively reinforces query-relevant image evidence. DCT uses the contrast between standard and visually blocked branches to select an intervention layer, while CFER re-encodes query-relevant image regions and controls the resulting update through source-region contributions. On CHAIR and POPE, VisDelta reduces object hallucination relative to the evaluated inference-time baselines. Results on MME and MMBench show no broad decline in the capabilities evaluated here, while component-replacement ablations on LLaVA-1.5-7B support the roles of layer selection, regional retrieval, and contribution control.

### AI USE STATEMENT

In this work, we used generative AI tools to assist with manuscript drafting, language editing, and consistency checks, including feedback on the presentation of experimental procedures and the interpretation of results. All AI-assisted text was reviewed and verified by the authors against the underlying experimental results and implementation details. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the assistance of generative AI.

### ETHICS STATEMENT

This work aims to improve the reliability of large vision-language models by reducing object hallucination through selective reinforcement of query-relevant visual evidence. Although the reported results show improved visual grounding, the method does not guarantee factual correctness or eliminate biases and potentially harmful outputs inherited from the underlying models. Its effectiveness beyond the evaluated settings requires further validation. Practical deployment should therefore include application-specific evaluation and appropriate human oversight.

### REPRODUCIBILITY STATEMENT

The main text and appendix (Appendix A and B) describe the proposed method, evaluation benchmarks, and inference procedures. The experimental appendix specifies the evaluation protocols, baseline configurations, decoding parameters, and sources of comparison results. Appendix C further details regional grouping, image-space mapping, context expansion, and crop encoding. These descriptions provide the implementation and evaluation details needed to reproduce our experiments.

## REFERENCES

Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Ping Chen, Xiaoqin Zhang, and Shijian Lu. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29915–29926. IEEE, 2025.

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*, 2023. doi: 10.48550/arXiv.2308.12966. URL https://arxiv.org/abs/2308.12966.

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. doi: 10.48550/arXiv.2502.13923. URL https://arxiv.org/abs/2502.13923.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. HALC: Object hallucination reduction via adaptive focal-contrast decoding. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 7824–7846. PMLR, 2024. URL https://proceedings.mlr.press/v235/chen24bi.html.

Zhe Cheng, Wenyu Chen, Fode Zhang, and Dehuan Shen. Mitigating hallucinations in large vision-language models via causal route gating. arXiv preprint arXiv:2605.24024, 2026. URL https://arxiv.org/abs/2605.24024.

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*, volume 36, 2023. URL https://arxiv.org/abs/2305.06500.

Yixiong Fang, Ziran Yang, Zhaorun Chen, Zhuokai Zhao, and Jiawei Zhou. Enhancing vision-language model reliability with uncertainty-guided dropout decoding. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4991. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/db48d94a42706019262ed8304fa658c5-Abstract-Conference.html.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024. doi: 10.1038/s41586-024-07421-0. URL https://www.nature.com/articles/s41586-024-07421-0.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multimodal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14303–14312, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Favero_Multi-Modal_Hallucination_Control_by_Visual_Information_Grounding_CVPR_2024_paper.html.

Mehrdad Fazli, Bowen Wei, and Ziwei Zhu. Mitigating hallucination in large vision-language models via adaptive attention calibration, 2025. URL https://arxiv.org/abs/2505.21472.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. doi: 10.48550/arXiv.2306.13394. URL https://arxiv.org/abs/2306.13394.

Jinghan He, Kuan Zhu, Haiyun Guo, Junfeng Fang, Zhenglin Hua, Yuheng Jia, Ming Tang, Tat-Seng Chua, and Jinqiao Wang. Cracking the code of hallucination in LVLMs with vision-aware head divergence. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3488–3501. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.175. URL https://aclanthology.org/2025.acl-long.175/.

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, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_OPERA_Alleviating_Hallucination_in_Multi-Modal_Large_Language_Models_via_Over-Trust_CVPR_2024_paper.html.

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*, pp. 6700–6709, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Hudson_GQA_A_New_Dataset_for_Real-World_Visual_Reasoning_and_Compositional_CVPR_2019_paper.html.

Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-introspective decoding: Alleviating hallucinations for large vision-language models. In *International Conference on Learning Representations*, pp. 24272–24295, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/3cc87f2bd3e3b4df8f9217326761c322-Paper-Conference.pdf.

Omri Kaduri, Shai Bagon, and Tali Dekel. What's in the image? a deep-dive into the vision of vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14549–14558, June 2025.

Yuqing Lei, Wenbo Lyu, Yingjun Du, Xiantong Zhen, Cees G. M. Snoek, and Ling Shao. See only when needed: Context-aware attention intervention for mitigating hallucinations in LVLMs. arXiv preprint arXiv:2606.29847, 2026. URL https://arxiv.org/abs/2606.29847.

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. URL https://openaccess.thecvf.com/content/CVPR2024/html/Leng_Mitigating_Object_Hallucinations_in_Large_Vision-Language_Models_through_Visual_Contrastive_CVPR_2024_paper.html.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

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 *Computer Vision – ECCV 2014*, pp. 740–755. Springer, 2014. doi: 10.1007/978-3-319-10602-1_48. URL https://doi.org/10.1007/978-3-319-10602-1_48.

Yangguang Lin, Quan Fang, Yufei Li, Jiachen Sun, Junyu Gao, and Jitao Sang. Hulluedit: Single-pass evidence-consistent subspace editing for mitigating hallucinations in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11086–11095, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Lin_HulluEdit_Single-Pass_Evidence-Consistent_Subspace_Editing_for_Mitigating_Hallucinations_in_Large_CVPR_2026_paper.html.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 57689–57733, 2024a. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/fed9263d7f1086e735904ff690527a53-Paper-Conference.pdf.

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*, 2024b. URL https://arxiv.org/abs/2310.03744.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *Computer Vision – ECCV 2024*, pp. 125–140. Springer Nature Switzerland, 2024c. doi: 10.1007/978-3-031-73010-8_8. URL https://doi.org/10.1007/978-3-031-73010-8_8.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? In *Computer Vision – ECCV 2024*, pp. 216–233, 2024d. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/959_ECCV_2024_paper.php.

Malsha Ashani Mahawatta Dona, Beatriz Cabrero-Daniel, Yinan Yu, and Christian Berger. Betercheck: Towards safeguarding VLMs for automotive perception systems. In *Proceedings of the 28th IEEE International Conference on Intelligent Transportation Systems*, pp. 3776–3783, 2025. doi: 10.1109/ITSC60802.2025.11423129.

Minh Khoi Nguyen, Dai Lam Le, Amir Reza Jafari, Tuan Dung Nguyen, Hong Son Mai, Huy Thong Mai, Quang Huy Nguyen, Thanh Trung Nguyen, Reza Farahbakhsh, Noel Crespi, and Phi Le Nguyen. Med-stepbench: A hierarchical reasoning framework for evaluating hallucinations in

medical vision-language models. In *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence*, pp. 6841–6849, 2026. doi: 10.24963/ijcai.2026/761.

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*, pp. 4035–4045. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1437. URL https://aclanthology.org/D18-1437/.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. In *Computer Vision – ECCV 2022*, pp. 146–162, 2022. doi: 10.1007/978-3-031-20074-8_9.

Wei Suo, Hanzu Zhang, Lijun Zhang, Ji Ma, Peng Wang, and Yanning Zhang. Hallucination-aware intermediate representation edit in large vision-language models. In *International Conference on Learning Representations*, volume 2026, pp. 39318–39340, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/41fd34e25626cde2dd249e5b11fb547f-Paper-Conference.pdf.

Mehran Tamjidi, Hamidreza Dastmalchi, Ali Cheraghian, Mohammadreza Alimoradijazi, Aijun An, and Hossein Rahmani. Test-time hallucination control in large vision-language models. arXiv preprint arXiv:2608.11474, 2026. URL https://arxiv.org/abs/2608.11474.

Team GLM. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024. URL https://arxiv.org/abs/2406.12793.

Zifu Wan, Ce Zhang, Silong Yong, Martin Q. Ma, Simon Stepputtis, Louis-Philippe Morency, Deva Ramanan, Katia P. Sycara, and Yaqi Xie. ONLY: One-layer intervention sufficiently mitigates hallucinations in large vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/papers/Wan_ONLY_One-Layer_Intervention_Sufficiently_Mitigates_Hallucinations_in_Large_Vision-Language_Models_ICCV_2025_paper.pdf.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. AMBER: An LLM-free multi-dimensional benchmark for MLLMs hallucination evaluation, 2024a. URL https://arxiv.org/abs/2311.07397.

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. Association for Computational Linguistics, 2024b. doi: 10.18653/v1/2024.findings-acl.937. URL https://aclanthology.org/2024.findings-acl.937/.

Kejuan Yang, Yizhuo Zhang, Mingyuan Du, Yue Zhang, Dixin Zheng, Kaili Zhao, Yang Xiao, Hanzhong Liang, and Kenan Xiao. UNIVID: Unified vision-language model for video moderation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pp. 467–479. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-industry.32. URL https://aclanthology.org/2026.acl-industry.32/.

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*, pp. 14625–14634, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/papers/Yin_ClearSight_Visual_Signal_Enhancement_for_Object_Hallucination_Mitigation_in_Multimodal_CVPR_2025_paper.pdf.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105, 2024. doi: 10.1007/s11432-024-4251-x. URL https://link.springer.com/article/10.1007/s11432-024-4251-x.

Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. Mitigating object hallucination in large vision-language models via image-grounded guidance. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 77461–77486. PMLR, 2025a. URL https://proceedings.mlr.press/v267/zhao25j.html.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond multimodal hallucinations: Enhancing LVLMs through hallucination-aware direct preference optimization. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2025b. doi: 10.1109/ICME59968.2025.11209377. URL https://ieeexplore.ieee.org/document/11209377.

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 B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 56969–56998, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/fc625e831361cfcc82cb74224fdc66cb-Paper-Conference.pdf.

Xingyu Zhu, Kesen Zhao, Liang Yi, Shuo Wang, Zhicai Wang, Beier Zhu, Hanwang Zhang, and Xiangnan He. Look carefully: Adaptive visual reinforcements in multimodal large language models for hallucination mitigation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pp. 81235–81256, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/83749a89d03df30c66af88020e4c7892-Abstract-Conference.html.

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. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 80873–80899. PMLR, 2025. URL https://proceedings.mlr.press/v267/zou25e.html.

# APPENDIX

Contents

**A Evaluation Benchmarks and Metrics** **16**
CHAIR . . . 16
POPE . . . 16
MME . . . 16
MMBench . . . 16

**B Implementation Details** **17**
Evaluation protocol . . . 17
Baseline decoding . . . 17
VisDelta configuration . . . 17
Comparison sources . . . 17

**C Regional Grouping and Context Expansion** **17**
Valid patch locations . . . 17
Connected patch groups . . . 17
Inverse mapping to the original image . . . 18
Context expansion . . . 18
Crop encoding and source bookkeeping . . . 19

**D Parameter Sensitivity** **19**
Localization budget . . . 19
Retained visual evidence budget . . . 19
Reinforcement strength . . . 19

**E Additional Qualitative Cases** **20**
Bowl presence . . . 20
Sandwich presence . . . 20

# A Evaluation Benchmarks and Metrics

We use CHAIR and POPE to assess object hallucination, and MME and MMBench to assess broader visual understanding. Below, we define the metrics and specify which scores appear in our results.

**CHAIR.** CHAIR (Caption Hallucination Assessment with Image Relevance) evaluates object mentions in generated descriptions (Rohrbach et al., 2018). Its evaluator recognizes the 80 MS COCO object categories (Lin et al., 2014), including mapped synonyms, and treats a category as supported if it appears in the segmentation annotations or human reference captions. Let $\mathcal{H}$ be the multiset of unsupported, recognized object mentions and $\mathcal{O}$ the multiset of all recognized object mentions. Let $\mathcal{C}_{\mathcal{H}}$ be the generated descriptions containing at least one such unsupported mention and $\mathcal{C}$ all generated descriptions. Then

$$\mathrm{CHAIR}_I = \frac{|\mathcal{H}|}{|\mathcal{O}|}, \qquad \mathrm{CHAIR}_S = \frac{|\mathcal{C}_{\mathcal{H}}|}{|\mathcal{C}|}. \tag{11}$$

Thus $\mathrm{CHAIR}_I$ counts object *mentions*, whereas $\mathrm{CHAIR}_S$ counts whole descriptions, even if a description contains several sentences. Both are reported as percentages, with lower values indicating fewer hallucinations. Following the repeated-sampling procedure of Dropout Decoding (Fang et al., 2025), we evaluate 500 images from MS COCO val2014 under each of three random seeds and report the mean and standard deviation. We additionally report object recall and mean description length in words to measure coverage and verbosity.

**POPE.** POPE (Polling-based Object Probing Evaluation) asks yes/no questions about whether a named object appears in an image (Li et al., 2023). Present objects supply positive queries; absent objects supply an equal number of negative queries. The *random* policy samples absent categories, *popular* favors frequent but absent categories, and *adversarial* favors absent categories that commonly co-occur with present objects. The original MS COCO evaluation samples 500 images and asks six questions per image, yielding 3,000 questions for each policy. Our results cover MS COCO, A-OKVQA, and GQA. With "yes" as the positive class, the reported metrics are accuracy and $F_1 = 2\mathrm{TP}/(2\mathrm{TP} + \mathrm{FP} + \mathrm{FN})$; higher is better.

**MME.** MME uses paired, manually written yes/no questions to evaluate ten perception and four cognition subtasks (Fu et al., 2023). Perception covers object existence, count, position, and color; OCR; and posters, celebrities, scenes, landmarks, and artworks. Cognition covers commonsense reasoning, numerical calculation, text translation, and code reasoning. Its released evaluation contains 2,374 question–image pairs. The original benchmark instruction asks for a yes/no answer. For subtask $t$, the score is

$$S_t = \mathrm{Acc}_t + \mathrm{Acc}_t^{+}, \tag{12}$$

where $\mathrm{Acc}_t$ is question accuracy and $\mathrm{Acc}_t^{+}$ is the percentage of images for which *both* paired questions are correct. Each term is a percentage, so $S_t \le 200$. Full MME sums all 14 subtasks (maximum 2,800). The MME-Hall column in our table sums only existence, count, position, and color (maximum 800); the separately reported cognition score also has a maximum of 800.

**MMBench.** MMBench (Liu et al., 2024d) evaluates multiple-choice visual perception and reasoning. Its three-level taxonomy has perception and reasoning at Level 1, six categories at Level 2, and 20 abilities at Level 3. The official English development split contains 1,164 original questions. In CircularEval, answer choices are cyclically shifted and an item receives credit only if every shifted version is answered correctly; CircularEval accuracy is the percentage of credited original items. We use CircularEval for our MMBench runs.

## B Implementation Details

**Evaluation protocol.** Following MemVR (Zou et al., 2025), we format annotation-file questions using the chat template of each model. POPE is evaluated on MS COCO, A-OKVQA, and GQA under random, popular, and adversarial negative-object sampling, with an explicit yes/no response instruction. For MME, we use the annotated yes/no questions. For MMBench, we evaluate the English development split using CircularEval. CHAIR evaluation follows the three-seed sampling and aggregation procedure of Dropout Decoding (Fang et al., 2025). For each seed, 500 images are sampled from MS COCO val2014, and the mean and standard deviation are computed across the three runs. The CHAIR prompt is `Please describe this image in detail.` We evaluate maximum generation lengths of 64 and 512 tokens.

**Comparison sources.** All results in Tables 1 and 2 are obtained from our experiments. In Table 3, all baseline results are taken directly from AIR. We evaluate VisDelta under the same experimental settings as the corresponding AIR table and verify this alignment by reproducing the reported Vanilla baseline results. POPE averages are computed as the arithmetic mean across the three negative-object sampling policies. Values in red denote differences relative to the displayed Vanilla baseline within each dataset or model block, calculated at the reported numerical precision.

**Baseline decoding.** For our peer-method reproductions, we use the method-specific configurations documented by MemVR (Zou et al., 2025). The default is greedy decoding with `do_sample=False`, temperature 0, and one beam; for MemVR, we set its entropy threshold to 0.75. VCD uses `do_sample=True`, temperature 1, `noise_step`=500, plausibility constraint $\lambda = 0.1$, and contrastive weight $\alpha = 1$. OPERA uses `sample=True`, `beam`=5, `scale_factor`=50, `threshold`=15, `num_attn_candidates`=5, and `penalty_weights`=1. We use the respective native configurations for DD and HulluEdit.

**VisDelta configuration.** The base MLLM and visual encoder remain frozen throughout evaluation. The top-10 normalized-entropy gate activates when the normalized entropy exceeds 0.75. We fix the localization budget at $K_{\text{loc}} = 64$ and the retrieval budget at $K_{\text{ret}} = 576$. Appendix C describes the mapping to image space.

## C Regional Grouping and Context Expansion

This appendix details CFR's conversion from patch indices to regions in the original image, including valid-patch filtering, connected-component grouping, inverse coordinate mapping, and context expansion. These geometric operations are deterministic, are performed independently within each input request, and require no learned parameters.

**Valid patch locations.** The mapping below applies to the LLaVA-1.5 visual path. Its projected visual representation contains $N_v = 576$ patch tokens arranged in a square grid, without a class token. The grid side length is $m = \sqrt{N_v} = 24$. A row-major token index $i \in \{0, \dots, N_v - 1\}$ corresponds to

$$(r_i, c_i) = \left( \left\lfloor \frac{i}{m} \right\rfloor, i \bmod m \right). \tag{13}$$

When an image is padded to a square before visual encoding, some grid cells may lie entirely within the padding. Using the inverse mapping and validity test defined below, we construct the eligible index set $\mathcal{V}_{\text{valid}}$ from cells whose mapped rectangles overlap the original image with positive area. We rank the localization scores in Eq. 5 only over these indices and retain the highest-scoring $\min(K_{\text{loc}}, |\mathcal{V}_{\text{valid}}|)$ entries. A stable descending sort resolves ties by the original row-major token order.

**Connected patch groups.** Let $\mathcal{S}_{\text{loc}}$ denote the selected indices. We form an undirected graph over the selected patches, connecting distinct patches $i$ and $j$ when

$$(r_i - r_j)^2 + (c_i - c_j)^2 < \rho^2, \qquad \rho = 1.42. \tag{14}$$

On the integer grid, this threshold connects horizontal, vertical, and diagonal neighbors, but excludes pairs at a Euclidean distance of two or greater. The patch groups are therefore the eight-connected components of the selected grid cells. We enumerate components by starting each flood fill from the smallest remaining row-major index and sort the indices within each component. The configuration `max_crops=0` imposes no limit on the number of components passed to the subsequent mapping and deduplication steps.

**Inverse mapping to the original image.** We map each component's axis-aligned bounding rectangle from the processed patch grid back to the original image, accounting for padding, resizing, and center cropping. Let $(W, H)$ be the original image dimensions, $(C_w, C_h)$ the center-crop dimensions, and $S$ the target shortest-edge length. For square-padded preprocessing, the padded dimensions and offsets are

$$W_b = H_b = \max(W, H), \qquad p_x = \left\lfloor \frac{W_b - W}{2} \right\rfloor, \qquad p_y = \left\lfloor \frac{H_b - H}{2} \right\rfloor. \tag{15}$$

Without square padding, $(W_b, H_b) = (W, H)$ and $p_x = p_y = 0$. The resized dimensions are

$$R_w = \begin{cases} S, & W_b \le H_b, \\ \lfloor S W_b / H_b \rfloor, & W_b > H_b, \end{cases} \qquad R_h = \begin{cases} S, & H_b \le W_b, \\ \lfloor S H_b / W_b \rfloor, & H_b > W_b, \end{cases} \tag{16}$$

with center-crop offsets $o_x = \lfloor (R_w - C_w)/2 \rfloor$ and $o_y = \lfloor (R_h - C_h)/2 \rfloor$.

For component $g$, let $c_g^-$ and $c_g^+$ denote its minimum and maximum column indices, and let $r_g^-$ and $r_g^+$ denote its minimum and maximum row indices. Undoing the crop translation, resizing, and padding offsets gives the floating-point boundaries in the original image:

$$\begin{aligned} \bar{x}_0 &= \left( c_g^- \frac{C_w}{m} + o_x \right) \frac{W_b}{R_w} - p_x, \quad \bar{x}_1 = \left( (c_g^+ + 1) \frac{C_w}{m} + o_x \right) \frac{W_b}{R_w} - p_x, \\ \bar{y}_0 &= \left( r_g^- \frac{C_h}{m} + o_y \right) \frac{H_b}{R_h} - p_y, \quad \bar{y}_1 = \left( (r_g^+ + 1) \frac{C_h}{m} + o_y \right) \frac{H_b}{R_h} - p_y. \end{aligned} \tag{17}$$

Rounding the lower boundaries down and the upper boundaries up, then clipping to the image boundaries, gives

$$b_g = (x_0, y_0, x_1, y_1) = \big( \max(0, \lfloor \bar{x}_0 \rfloor), \max(0, \lfloor \bar{y}_0 \rfloor), \min(W, \lceil \bar{x}_1 \rceil), \min(H, \lceil \bar{y}_1 \rceil) \big). \tag{18}$$

The tuple $b_g$ denotes the half-open rectangle $[x_0, x_1) \times [y_0, y_1)$. We discard boxes with $x_1 \le x_0$ or $y_1 \le y_0$. The same mapping and nonempty-box test are applied to individual grid cells when constructing $\mathcal{V}_{\text{valid}}$. Duplicate nonempty boxes are removed, retaining their first occurrence in component order.

**Context expansion.** For a valid raw box $b_g$, let $w_g = x_1 - x_0$ and $h_g = y_1 - y_0$. We set the target extent to

$$e_g = \max \left( w_g, h_g, \left\lceil \min(W, H) \frac{6}{m} \right\rceil \right). \tag{19}$$

For $m = 24$, the last term sets a minimum target extent of one quarter of the shorter image side, rounded up to an integer number of pixels.

For an interval $[u_0, u_1)$ along an image axis of length $A$, we expand it to $[a_A, a_A + L_A)$, where

$$L_A = \min(A, e_g), \qquad a_A = \max \left( 0, \min \left( A - L_A, \left\lfloor \frac{u_0 + u_1 - L_A}{2} \right\rfloor \right) \right). \tag{20}$$

The start position is obtained by centering the interval on the raw box, up to integer rounding, and clamping it to $[0, A - L_A]$. This preserves the length $L_A$ while keeping the interval within the image.

Applying Eq. 20 to $[x_0, x_1)$ with $A = W$ and to $[y_0, y_1)$ with $A = H$ yields the final region $\mathcal{R}_g$. The construction preserves the entire raw box. The final width and height are $\min(W, e_g)$ and $\min(H, e_g)$, respectively, so the region is square when $e_g \le \min(W, H)$. We deduplicate the regions again after expansion, since distinct raw boxes may produce identical final crops.

**Crop encoding and source bookkeeping.** Each final region is cropped from the original image. Before native visual preprocessing, nonsquare crops are padded to a square using the mean color specified by the visual processor. Each crop is then encoded separately by the frozen visual encoder and multimodal projector. We concatenate the resulting token sequences in region order and associate each token with its source region for the contribution calculation in Eq. 8.

A request-local least-recently-used cache with a capacity of 64 entries reuses encodings when the same final crop box recurs within a request. The cache is cleared before each request, so cached encodings are not reused across requests. If no selected component yields a valid image region, the implementation raises an error rather than substituting the full image.

## D Parameter Sensitivity

Figure 6 reports the sensitivity of VisDelta to the localization budget $K_{\text{loc}}$ and the retrieval visual evidence budget $K_{\text{ret}}$ on LLaVA-1.5-7B. Each budget is varied independently, with the remaining parameters held at their reference values in $(K_{\text{loc}}, K_{\text{ret}}, \alpha_{\max}) = (64, 576, 0.18)$. We report POPE average accuracy and F1, together with $\text{CHAIR}_S$ and $\text{CHAIR}_I$.

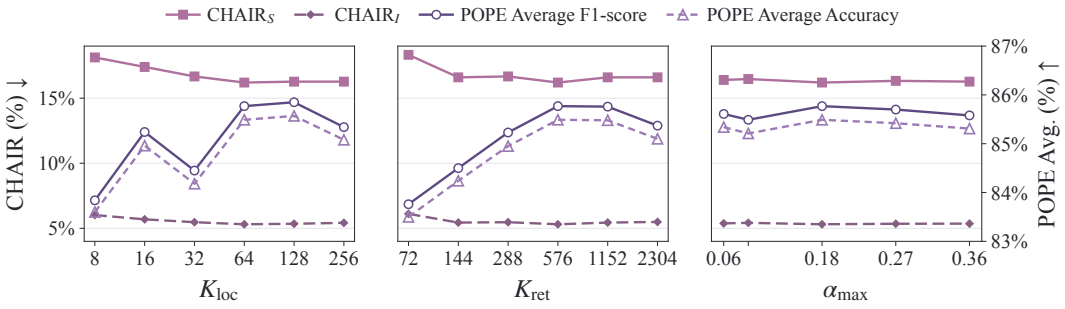


Figure 6: Sensitivity to the localization budget $K_{\text{loc}}$ (left) and retained visual evidence budget $K_{\text{ret}}$ (middle) on LLaVA-1.5-7B. The reference configuration is $(64, 576, 0.18)$. The left axis shows $\text{CHAIR}_S$ and $\text{CHAIR}_I$ under a 64-token generation limit and the right axis shows POPE accuracy and F1 averaged across nine combinations. All metrics are reported as percentages. Lower CHAIR and higher POPE values are preferable.

**Localization budget.** As $K_{\text{loc}}$ increases from small values, POPE average F1 improves and $\text{CHAIR}_S$ decreases, consistent with insufficient spatial coverage at very small localization budgets. Beyond $K_{\text{loc}} = 64$, F1 improves only marginally before declining, while $\text{CHAIR}_S$ increases. Thus, a larger localization budget does not consistently improve object-presence reasoning or reduce object hallucination. The default setting of $K_{\text{loc}} = 64$ achieves near-peak F1 and the lowest $\text{CHAIR}_S$ among the evaluated budgets, supporting its use as a balanced choice.

**Retrieval visual evidence budget.** As $K_{\text{ret}}$ increases from small values, POPE average F1 and accuracy improve, while both $\text{CHAIR}_S$ and $\text{CHAIR}_I$ decrease. These trends are consistent with insufficient visual evidence being retained at very small budgets. Beyond $K_{\text{ret}} = 576$, the POPE metrics plateau before declining slightly, while the CHAIR metrics increase modestly. This limited degradation is consistent with CCR weighting source regions by their contributions, which may suppress noise from less relevant regions as the retained evidence budget grows. Thus, retaining more visual evidence does not consistently improve object-presence reasoning or reduce object hallucination.

**Reinforcement strength.** The trends shown in Figure 6 illustrate limited variation in POPE and CHAIR metrics as $\alpha_{\max}$ changes. POPE average F1 and accuracy peak near $\alpha_{\max} = 0.18$ and decline slightly at larger values, while $\text{CHAIR}_S$ reaches its minimum near the same setting and

$\text{CHAIR}_I$ remains nearly unchanged. This relative stability is consistent with contribution-controlled integration moderating sensitivity to $\alpha_{\max}$ by dynamically adjusting the reinforcement strength. POPE average F1 and accuracy peak near $\alpha_{\max} = 0.18$ and decline slightly at larger values, while $\text{CHAIR}_S$ reaches its minimum near the same setting and $\text{CHAIR}_I$ remains nearly unchanged.

## E Additional Qualitative Cases

Figures 7 and 8 show two cases in which the queried objects are visible. In both cases, LLaVA-1.5-7B answers negatively, while the same backbone with VisDelta gives the correct affirmative answer. These cases show how VisDelta corrects missed object-presence judgments when the relevant visual evidence is available.

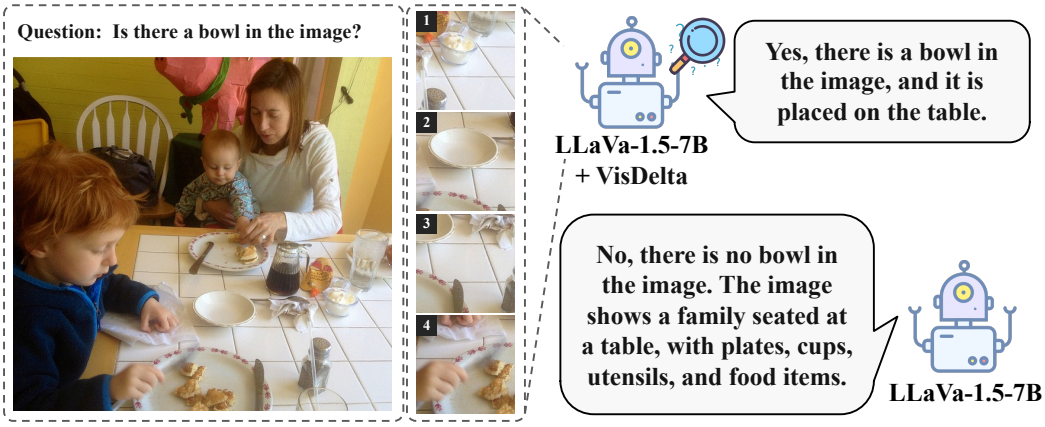


Figure 7: Bowl-presence example on LLaVA-1.5-7B. The baseline answers negatively despite two visible bowls, whereas VisDelta answers affirmatively. The enlarged views show the small filled bowl and the shallow white bowl.

**Bowl presence.** Figure 7 shows two bowls among other tableware. The baseline describes the dining scene but incorrectly denies the presence of bowls. With VisDelta, the same backbone correctly affirms their presence.

**Sandwich presence.** Figure 8 shows a meat-filled roll cut into portions on a plate. The baseline mentions a possible roll and meat but incorrectly answers the sandwich-presence question negatively. The visible bread and filling provide local visual evidence supporting an affirmative answer in this case, which the same backbone gives with VisDelta.

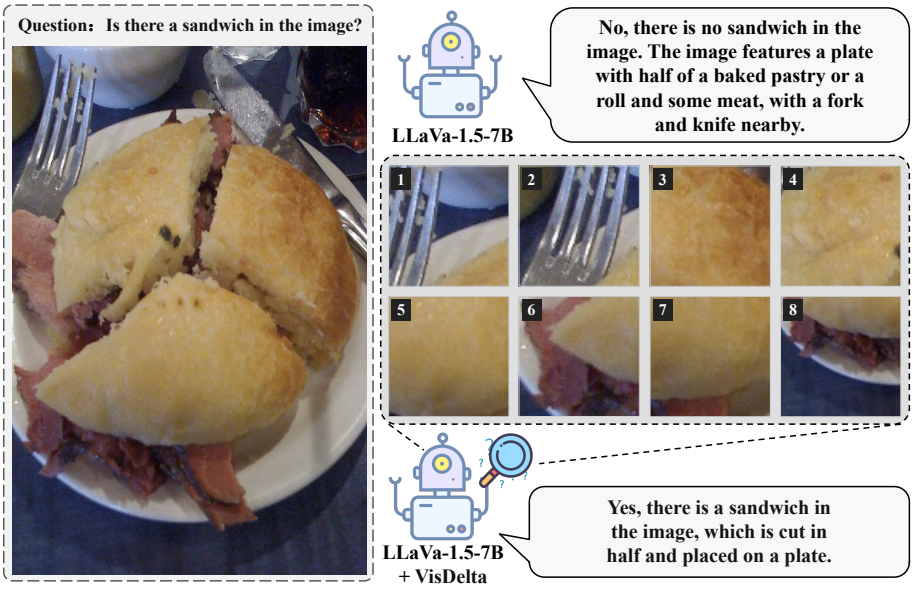


Figure 8: Sandwich-presence example on LLaVA-1.5-7B. The baseline incorrectly denies the presence of the visible sandwich, whereas VisDelta correctly affirms its presence. The enlarged views show the food and nearby utensils.