TAVCD: TARGET-AWARE VISUAL CONTRASTIVE DECODING FOR HALLUCINATION MITIGATION IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Vision-Language Models (LVLMs) frequently hallucinate objects absent from the visual input. Visual Contrastive Decoding (VCD) mitigates this by contrasting positive-branch logits (original image) against a negative branch (degraded image), where the adjustment is scaled by a fixed scalar α. We identify a fundamental limitation of this paradigm, which we formalise as the *Contrastive Correctability Condition* (C^3). No fixed positive α can correct a hallucinated token h whenever h simultaneously dominates the ground-truth token g in both the baseline logit distribution and the visual contrastive signal $D(v) = \mathrm{pos}(v) - \mathrm{neg}(v)$. We further prove a *universal insufficiency* result: for any fixed $\alpha_0 \in \mathbb{R}$, there exist realisable decoding configurations under which α_0 either fails to correct or actively amplifies hallucination. Empirically, across five backbones on MSCOCO, we observe C^3 violations at approximately 26% of object-generating steps under fixed-α VCD, confirming it is counterproductive at roughly one in four object-generating steps. Building on this analysis, we propose **TAVCD**, a lightweight per-step adaptive contrastive decoder. At each decoding step, TAVCD builds a compact feature vector from observable logit statistics and passes it through a shared-trunk MLP with gate and magnitude heads ($\approx$12 K parameters). The gate head decides whether contrastive intervention is warranted; the magnitude head selects the signed strength from a fine-grained 40-point α grid. TAVCD is trained offline with oracle supervision from ground-truth annotations and adds negligible wall-clock overhead at inference. On MSCOCO and PASCAL VOC captioning across five diverse LVLMs, TAVCD reduces hallucination on all five backbones. The largest gains, on PASCAL VOC, reach 24.0 pp in CHAIR_i over standard decoding and 5.8 pp in CHAIR_s over fixed-α VCD.

1 INTRODUCTION

Large Vision-Language Models (LVLMs) (Liu et al., 2023; Dai et al., 2023; Bai et al., 2023; Ye et al., 2024; Zhu et al., 2024) have made substantial progress in multimodal understanding, yet remain susceptible to *object hallucination* (Rohrbach et al., 2018; Li et al., 2023c; Gunjal et al., 2024; Liu et al., 2024a), generating objects absent from the visual input. Object hallucination poses potential risks in safety-critical applications.

Among training-free mitigation strategies, Visual Contrastive Decoding (VCD) frameworks (Leng et al., 2024; Lee & Song, 2025; Deng & Yang, 2025) have emerged as principled and lightweight approaches. At each decoding step, VCD runs two forward passes (one with the original image and one with a noise-corrupted version of the input image) and adjusts the next-token logit as

$$L_\alpha(v) = \mathrm{pos}(v) + \alpha\big(\mathrm{pos}(v) - \mathrm{neg}(v)\big), \tag{1}$$

where $\alpha \geq 0$ is a global hyperparameter.

Across all VCD variants, the contrastive coefficient α is treated as a fixed scalar shared across decoding steps and backbone models. We argue that this static treatment is the binding constraint, not the choice of negative view.

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.

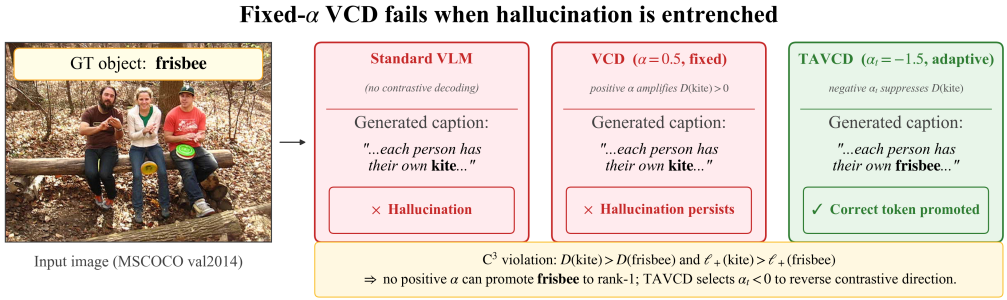

Figure 1: A concrete C^3 violation: in this MSCOCO image containing a frisbee, the hallucinated token *kite* dominates the ground-truth token *frisbee* in both baseline probability and contrastive signal D. A fixed positive α amplifies kite's contrastive advantage, leaving the hallucination intact. TAVCD selects $\alpha_t < 0$, reversing the contrastive direction and promoting *frisbee* to rank-1.

We formalise this claim through a theoretical analysis of when fixed-α VCD can succeed at a single decoding step. Let g be the ground-truth token and h be a hallucination candidate; define $\delta_h = D(g) - D(h)$ and $\Delta_h = \text{pos}(h) - \text{pos}(g)$. Theorem 1 shows that whenever $\delta_h \leq 0$ and $\Delta_h \geq 0$ (a condition we call a *Contrastive Correctability Condition* (C^3) violation), no positive α can place g above h. Corollary 1 proves a strictly stronger *universal insufficiency* result: for any fixed $\alpha_0 \in \mathbb{R}$, there exist realisable decoding configurations under which α_0 fails to promote the target token. The optimal α is therefore *step-dependent*.

Empirically, across five backbones on MSCOCO, we observe C^3 violations at approximately 26% of object-generating decoding steps under fixed-α VCD, confirming it is counterproductive at roughly one in four object-generating steps.

We address this gap with **TAVCD** (Target-Aware VCD), a lightweight per-step adaptive contrastive decoder. TAVCD wraps a frozen LVLM with a compact MLP ($\approx$12 K parameters) that observes per-step logit statistics and emits a signed, scaled coefficient α_t. The *gate head* determines whether contrastive intervention should be applied at step t; the *magnitude head* selects the strength from a 40-point grid spanning $[-4.0, 4.0]$. At inference, TAVCD computes $\alpha_t = \sigma(s_t) \times \mathbb{E}[m_t]$ and clips α_t to a fixed symmetric range. The MLP is trained offline on logged decoding statistics with per-step oracle supervision derived from ground-truth object annotations. No LVLM parameters are updated.

We validate TAVCD on MSCOCO 2014 and PASCAL VOC 2007 captioning across five diverse LVLMs (BLIP-2, Moondream2, Qwen2-VL-2B, LLaVA-1.5, and Qwen-VL-Chat). TAVCD achieves the lowest or tied-for-lowest CHAIR_i and CHAIR_s on all five backbones on both datasets. Gains in CHAIR_i reach up to 24.0 pp over standard decoding; gains in CHAIR_s reach up to 5.8 pp over fixed-α VCD. These results confirm that coefficient adaptation and negative-view engineering address orthogonal sources of failure.

Contributions:

1. **Theoretical analysis.** We formalise the C^3 condition and prove that no fixed $\alpha \in \mathbb{R}$ is universally corrective for VCD (Section 3). The analysis identifies two distinct failure modes (*blocking* and *window collapse*) and characterises the per-step feasible promotion window $\mathcal{A}^*(g)$.
2. **Empirical violation analysis.** Across five backbones on MSCOCO, we measure an approximately 26% C^3 violation rate at object-generating steps, confirming that fixed-α VCD actively harms target promotion at roughly one in four object-generating steps.
3. **TAVCD.** We propose a lightweight gate+magnitude MLP predictor that selects α_t per step from observable logit statistics, trained offline with no LVLM parameter updates. TAVCD

achieves the lowest or tied-for-lowest CHAIR_i and CHAIR_s simultaneously across all five LVLMs on two benchmark datasets.

2 Related Work

Large Vision-Language Models. The rapid advancement of LLMs (Brown et al., 2020; Touvron et al., 2023; Chowdhery et al., 2023) has driven the development of LVLMs capable of complex multimodal reasoning. Foundational architectures (Alayrac et al., 2022; Chen et al., 2023) align frozen visual encoders with language models through large-scale pre-training. Visual instruction tuning (Liu et al., 2023; Dai et al., 2023) has emerged as a dominant paradigm; representative LVLMs include LLaVA (Liu et al., 2023), BLIP-2 (Li et al., 2023a), and Qwen-VL (Bai et al., 2023). Despite strong benchmark performance (Fu et al., 2023; Liu et al., 2024b), these models remain vulnerable to hallucination when visual grounding is insufficient (Li et al., 2023c; Zhou et al., 2024; Wang et al., 2023).

Object hallucination in LVLMs. Object hallucination (Rohrbach et al., 2018; Li et al., 2023c; Gunjal et al., 2024; Liu et al., 2024a; Lovenia et al., 2024) is well documented. POPE (Li et al., 2023c) converts hallucination detection into a binary classification problem; CHAIR (Rohrbach et al., 2018) measures hallucination at sentence and instance levels in open-ended captioning. Post-hoc revision methods (Zhou et al., 2024) correct hallucinations at inference time; factually augmented RLHF (Sun et al., 2024) reduces them via targeted retraining. Dataset construction and instruction tuning (Gunjal et al., 2024; Liu et al., 2024a) reduce hallucination rates through improved supervision. We focus on inference-time decoding, which requires no model retraining.

Contrastive decoding for hallucination mitigation. Inference-time contrastive decoding was introduced for NLP to penalise undesirable generation artefacts (Liu et al., 2021; Li et al., 2023b; O'Brien & Lewis, 2023). In the multimodal domain, VCD (Leng et al., 2024) adapts this principle without model retraining. Subsequent work has moved beyond simple noise injection to construct more discriminative negative views. R-VCD (Lee & Song, 2025) uses retrieved AI-generated single-concept images as the negative branch. OA-VCD (Chen et al., 2026) removes the most salient object regions found by self-supervised ViT attention. MaskCD (Deng & Yang, 2025) masks the LVLM's image-specialised attention heads. VACoDe (Kim et al., 2024) selects adaptively among multiple augmentation strategies. A fundamental limitation across this entire family is the reliance on a static, globally fixed contrastive weight α. Our work departs from negative-view engineering and instead dynamically optimises the contrastive coefficient at each decoding step; this constitutes an orthogonal axis of improvement.

3 Theoretical Analysis: When Can VCD Promote the Target Token?

The fundamental goal of Visual Contrastive Decoding is to *promote the target token*: at each decoding step, the contrastive adjustment should raise the ground-truth token g above every hallucination candidate in the output distribution. We formalise this as the *target promotion problem* and derive precise conditions under which it is solvable. Our central finding is that no fixed α, regardless of sign or magnitude, is universally target-promoting.

3.1 Setup, Notation, and the Target Promotion Objective

Let $\mathcal{V}$ denote the model's vocabulary. At a given decoding step, VCD assigns to each token $v \in \mathcal{V}$ the adjusted logit

$$L_\alpha(v) = \text{pos}(v) + \alpha D(v), \qquad D(v) \triangleq \text{pos}(v) - \text{neg}(v), \tag{2}$$

where $\text{pos}(v)$ and $\text{neg}(v)$ are the logits under the original and noise-corrupted visual inputs respectively, and $\alpha \in \mathbb{R}$ is the contrastive coefficient. The signal $D(v)$ measures the visual sensitivity of token v.

Let $g \in \mathcal{V}$ be a ground-truth object token at the current step, and let $\mathcal{H} \subseteq \mathcal{V} \setminus \{g\}$ denote the set of hallucination candidates. For each $h \in \mathcal{H}$, we define

$$\delta_h \triangleq D(g) - D(h), \qquad \Delta_h \triangleq \text{pos}(h) - \text{pos}(g), \tag{3}$$

and the *pairwise margin*

$$M_h(\alpha) \triangleq L_\alpha(g) - L_\alpha(h) = \alpha\delta_h - \Delta_h, \tag{4}$$

which is affine in α. Correcting hallucination h requires $M_h(\alpha) > 0$; correcting *all* candidates simultaneously requires α to lie in the global feasible set

$$\mathcal{A}^+(g) \triangleq \bigcap_{h \in \mathcal{H}} \big\{\alpha > 0 : M_h(\alpha) > 0\big\}. \tag{5}$$

We further denote by $\mathcal{A}^*(g) \triangleq \bigcap_{h\in\mathcal{H}}\{\alpha \in \mathbb{R} : M_h(\alpha) > 0\}$ the analogous set without sign restriction.

Target Promotion Problem. *Given the current logit geometry* $(\mathrm{pos}, \mathrm{neg})$*, find* $\alpha \in \mathcal{A}^+(g)$ *such that* g *ranks above every* $h \in \mathcal{H}$ *in the adjusted distribution.* The set $\mathcal{A}^+(g)$ being empty means no positive α can simultaneously promote the target over all competitors, a *structural failure* of fixed-α VCD determined jointly by the rank of g in the contrastive signal D (the δ_h axis) and in the baseline probability distribution (the Δ_h axis).

3.2 Pairwise Feasible Sets

We first characterise, for each individual hallucination candidate, the set of positive coefficients that suffice to correct it. Define the *crossover point* $\alpha_h^\times \triangleq \Delta_h/\delta_h$ whenever $\delta_h \neq 0$.

Proposition 1 (Pairwise feasibility under positive α)**.** *For each* $h \in \mathcal{H}$*, the pairwise feasible set* $I_h^+ \triangleq \{\alpha > 0 : M_h(\alpha) > 0\}$ *admits the closed form*

$$I_h^+ = \begin{cases} (\alpha_h^\times, +\infty) & \delta_h > 0,\ \Delta_h > 0, \\ (0, +\infty) & \delta_h > 0,\ \Delta_h \le 0, \\ (0, +\infty) & \delta_h = 0,\ \Delta_h < 0, \\ (0, \alpha_h^\times) & \delta_h < 0,\ \Delta_h < 0, \\ \emptyset & \delta_h \le 0 \text{ and } \Delta_h \ge 0. \end{cases} \tag{6}$$

All proofs are deferred to Appendix A.

The fifth case is the structural source of failure: any candidate h with $\delta_h \le 0$ and $\Delta_h \ge 0$ renders I_h^+ empty regardless of α.

3.3 Target Promotion Failure: The C^3 Blocking Condition

Case 5 of Proposition 1 identifies the precise configuration that *structurally blocks* target promotion: when the hallucination candidate h simultaneously dominates the target g in the contrastive signal ($D(h) \ge D(g)$) and in the baseline probability ($\mathrm{pos}(h) \ge \mathrm{pos}(g)$), no positive contrastive amplification can push g above h.

Theorem 1 (C^3 characterisation — Target Promotion Blocking)**.** *For every* $h \in \mathcal{H}$*,*

$$I_h^+ = \emptyset \iff \delta_h \le 0 \text{ and } \Delta_h \ge 0.$$

We call the right-hand side the ***Contrastive Correctability Condition*** *(**C^3**)* ***violation****. Whenever any* $h \in \mathcal{H}$ *violates* C^3*,* $\mathcal{A}^+(g) \subseteq I_h^+ = \emptyset$*.*

The C^3 condition has a concrete geometric interpretation. When this condition is violated, the target token g is outranked by h on *both* axes simultaneously. First, h carries a larger contrastive signal ($D(h) \ge D(g)$): the noisy branch suppresses h less than g relative to language priors, making g appear driven by language priors rather than visual evidence, despite g being the correct token. Second, h has a higher baseline probability ($\mathrm{pos}(h) \ge \mathrm{pos}(g)$), so language priors also favour h. Increasing α only amplifies h's contrastive advantage, deepening g's rank deficit. Together, these conditions define a regime where **VCD's contrastive mechanism actively works against target promotion**.

3.4 WINDOW COLLAPSE: JOINT TARGET PROMOTION FAILS WITHOUT C^3

C^3 is sufficient for target promotion failure but not necessary. Even when each individual competitor h is correctable in isolation (i.e., there exists a positive α that promotes g over h), joint promotion of g over *all* competitors may still fail. This occurs when the per-competitor feasible intervals have an empty intersection.

Theorem 2 (Structure of $\mathcal{A}^+(g)$)**.** *Define*

$$\underline{\alpha}(g) \triangleq \max_{\substack{h\in\mathcal{H}\\ \delta_h>0,\ \Delta_h>0}} \alpha_h^{\times}, \qquad \bar{\alpha}(g) \triangleq \min_{\substack{h\in\mathcal{H}\\ \delta_h<0,\ \Delta_h<0}} \alpha_h^{\times},$$

with $\max\emptyset = 0$ *and* $\min\emptyset = +\infty$*. If no* $h \in \mathcal{H}$ *violates* C^3*, then* $\mathcal{A}^+(g) = \big(\underline{\alpha}(g),\ \bar{\alpha}(g)\big) \cap (0, +\infty)$*. Combining with Theorem 1:*

$$\mathcal{A}^+(g) = \emptyset \iff \underbrace{\exists h \text{ violating } C^3}_{\text{pairwise mode}} \vee \underbrace{\underline{\alpha}(g) \geq \bar{\alpha}(g)}_{\text{window collapse}}. \tag{7}$$

The two failure modes are qualitatively distinct but share the same consequence. C^3 violation is *local*: a single offending pair (g, h) renders g structurally blocked, detectable from (δ_h, Δ_h) alone. Window collapse is *non-local*: every competitor is individually promotable, yet their joint feasibility constraints conflict.

3.5 BEYOND POSITIVE α: UNIVERSAL INSUFFICIENCY

Allowing $\alpha \in \mathbb{R}$ shrinks the failure set but does not eliminate it. Let $I_h \triangleq \{\alpha \in \mathbb{R} : M_h(\alpha) > 0\}$.

Proposition 2 (Pairwise feasibility on $\mathbb{R}$)**.** $I_h = (\alpha_h^{\times}, +\infty)$ *if* $\delta_h > 0$*;* $I_h = (-\infty, \alpha_h^{\times})$ *if* $\delta_h < 0$*;* $I_h = \mathbb{R}$ *if* $\delta_h = 0,\ \Delta_h < 0$*;* $I_h = \emptyset$ *if* $\delta_h = 0,\ \Delta_h \geq 0$*.*

Corollary 1 (Universal insufficiency)**.** *For any fixed* $\alpha_0 \in \mathbb{R}$*, there exist realisable decoding configurations under which* $\alpha_0 \notin \mathcal{A}^*(g)$*. Hence no fixed contrastive coefficient is universally corrective across decoding steps.*

Implication: per-step adaptation is necessary. The corollary rules out any escape via global tuning: even with α unrestricted in sign, no scalar choice is universally corrective. Theorems 1 and 2 show that promotability at each step depends on the *instantaneous rank geometry* of g and its competitors in both pos and D, which shifts as decoding proceeds. A fixed α therefore inevitably meets steps where it either works against target promotion (C^3 violation) or falls outside the feasible promotion window (window collapse). The only remaining degree of freedom is to observe the current-step logit structure and select α_t accordingly, which we term *target-aware* adaptation.

4 TAVCD: TARGET-AWARE VISUAL CONTRASTIVE DECODING

Section 3 establishes that the feasible contrastive window $\mathcal{A}^*(g)$ (the set of α values that promote the target token at step t) is step-dependent and shifts with every update to the logit distribution. Computing it requires knowing the ground-truth token g, which is unavailable at inference. TAVCD addresses this by training a small predictor to *approximate* whether the current step warrants positive amplification, negative suppression, or neutral intervention, using only observable decoding statistics as proxies for the rank geometry of g.

4.1 APPROXIMATING THE CONTRASTIVE FEASIBILITY WINDOW

The core challenge. Theorems 1 and 2 show that target promotion requires $\alpha_t \in \mathcal{A}^*(g)$, an interval whose boundaries depend on the pairwise quantities (δ_h, Δ_h) for every hallucination candidate h, which are undefined without knowing g. The challenge is to approximate the membership decision $\alpha_t \in \mathcal{A}^*(g)$ from *observable* signals only.

Gate+magnitude decomposition. We parameterise the per-step coefficient as

$$\alpha_t = \sigma(s_t) \times \mathbb{E}[m_t], \tag{8}$$

where $\sigma(s_t) \in (0,1)$ is a soft gate (sigmoid of a scalar logit) deciding whether contrastive intervention is warranted at step t, and $\mathbb{E}[m_t] = \sum_k \mathrm{softmax}(\ell_t)[k] \cdot c_k$ is the expected signed magnitude under a learned distribution over a discrete α grid $\mathcal{G} = \{-4.0, -3.8, \ldots, -0.2, 0.2, \ldots, 4.0\}$ (40 values, step size 0.2, excluding zero); the clip bound is $\alpha_{\text{clip}} = 4.0$, matching the grid extremes. This factorisation separates two orthogonal decisions: *whether* to intervene (gate) and *how strongly and in which direction* (magnitude). Both outputs are computed simultaneously by a compact shared-trunk MLP f_θ that reads the observable feature vector $\mathbf{x}_t$:

$$(s_t,\ \ell_t) = f_\theta(\mathbf{x}_t), \tag{9}$$

where s_t is a scalar gate logit and $\ell_t \in \mathbb{R}^{40}$ are magnitude class logits. A final symmetric clip on $|\alpha_t|$ prevents excessively aggressive intervention. Figure 2 illustrates the complete inference pipeline.

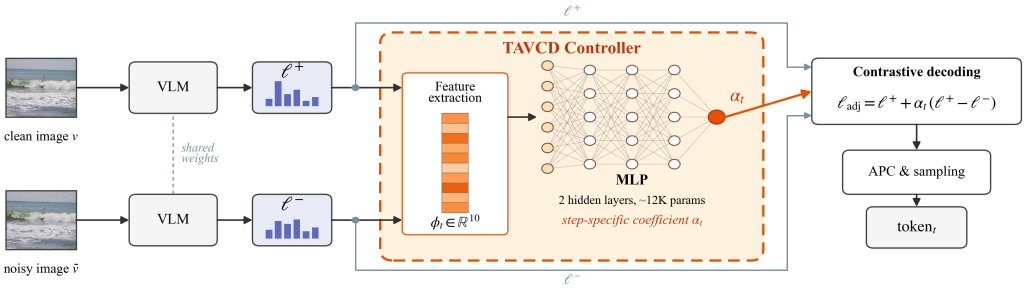

Figure 2: TAVCD inference pipeline at decode step t. The *TAVCD Controller* (dashed box) computes a 10-dimensional feature vector $\mathbf{x}_t$ from the shared top-K support of both branches and passes it through a shared-trunk MLP with two hidden layers (FC 128, FC 64, ReLU, dropout 0.1). The gate head produces gate value $\sigma(s_t) \in (0,1)$ and the magnitude head produces a distribution over the 40-point grid $\mathcal{G}$, yielding expected value $\mathbb{E}[m_t]$. Their product $\alpha_t{=}\sigma(s_t) \times \mathbb{E}[m_t]$, clipped to a fixed symmetric range, enters the contrastive formula (Eq. 1); APC then masks tokens with $\mathrm{pos}(v) < \log\beta + \max_w \mathrm{pos}(w)$ before sampling. When $\sigma(s_t) \approx 0$, the formula recovers plain pos.

4.2 Proxy Features for Step-Level Logit Statistics

At inference, $\mathbf{x}_t$ must be computable without ground-truth labels. We therefore use proxy statistics from current logits and recent decoding history. Specifically, we build a 10-dimensional feature vector from two sources:

Branch statistics (7 features). Taking the top-K tokens ($K = 50$) from pos as a common support, we compute seven statistics. These include the Shannon entropies H_{pos} and H_{neg} of both branches, their difference $\Delta H = H_{\text{pos}} - H_{\text{neg}}$, and the KL divergence $\mathrm{KL}(p_{\text{pos}} \| p_{\text{neg}})$. We also record the peak positive-branch probability $p_{\text{pos}}^{\max}$, the negative-branch probability at the top-1 positive token $p_{\text{neg}}^{\max}$, and the logit gap between rank-1 and rank-2 tokens in the positive branch, gap_1.

Short-horizon trajectory (3 features). To capture temporal instability before hallucination, we maintain a sliding window of $W = 5$ steps and compute the rolling mean of ΔH and KL, plus the relative step position t/T.

All features are z-score normalised with per-feature mean μ and standard deviation $\hat{\sigma}$ computed from the training split, and stored alongside the model checkpoint.

4.3 Empirical Supervision via Object-Token Grid Search

Oracle labels are generated by an offline pass over COCO training captions decoded with fixed-α VCD. Each step is classified as a *hallucinated-object*, *correct-object*, or *non-object* step using GT object annotations; a grid search over the 40-point $\mathcal{G}$ then assigns the largest α^* that promotes a ground-truth token g^* to rank 1, the target promotion objective of Section 3. Figure 3 illustrates the three-way decision logic; the full procedure (including step-type definitions, grid-search criteria, and class statistics) is given in Appendix B.

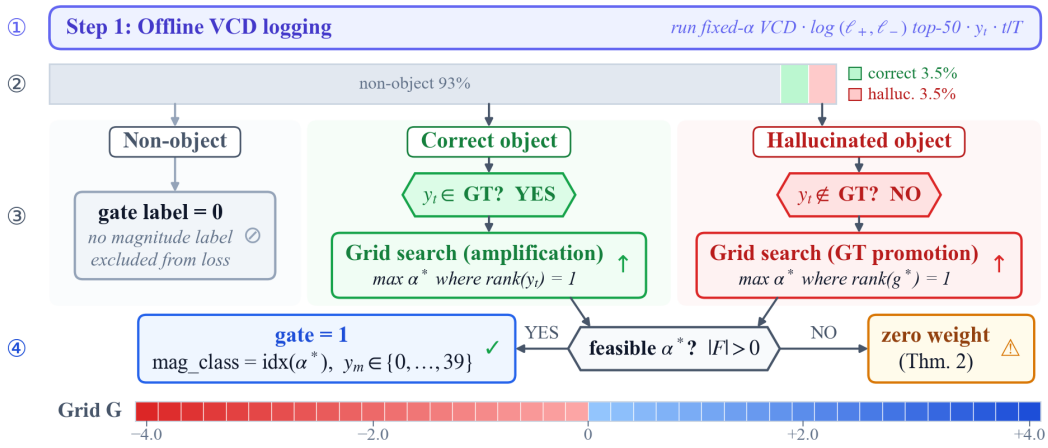

Figure 3: Oracle α-labelling pipeline. Each logged fixed-α VCD step passes two checks: *object token* (is y_t the first token of a COCO category?) and *GT presence* (is that category annotated?). Non-object steps receive gate label y_g=0. Object-generating steps undergo a 40-point grid search for the maximum α^* that promotes a ground-truth token g^* to rank-1: g^*=y_t for correct objects, and the highest-pos GT token in the top-K for hallucinated objects. Steps with no feasible grid point are structurally uncorrectable (Theorem 2) and receive zero loss weight.

4.4 Architecture and Training

Predictor architecture. The target-aware predictor f_θ is a shared-trunk MLP with ≈12 K parameters:

$$\mathbf{h} = \mathrm{ReLU}(\mathrm{Dropout}(\mathbf{W}_2\,\mathrm{ReLU}(\mathrm{Dropout}(\mathbf{W}_1\mathbf{x}_t + \mathbf{b}_1)) + \mathbf{b}_2)),$$
$$s_t = \mathbf{w}_g^\top \mathbf{h} + b_g, \quad \ell_t = \mathbf{W}_m\mathbf{h} + \mathbf{b}_m,$$

where $\mathbf{W}_1 \in \mathbb{R}^{128\times 10}$, $\mathbf{W}_2 \in \mathbb{R}^{64\times 128}$, $\mathbf{w}_g \in \mathbb{R}^{64}$ (gate head), $\mathbf{W}_m \in \mathbb{R}^{40\times 64}$ (magnitude head), with dropout rate 0.1.

Objective. The gate head uses focal BCE to handle the 93% non-object class imbalance ($\gamma = 2$, positive-class weight $w_{\text{pos}} = 0.95$). The magnitude head uses cross-entropy with label smoothing ($\epsilon = 0.1$), applied only to object-generating steps with a valid oracle label.

$$\mathcal{L} = \mathcal{L}_{\text{gate}} + \lambda_{\text{mag}} \cdot \mathcal{L}_{\text{mag}}, \quad \lambda_{\text{mag}} = 1.0. \tag{10}$$

We train with Adam (lr $= 10^{-3}$, weight decay $= 10^{-4}$), batch size 512, cosine annealing over 12 epochs with early stopping (patience $= 3$ epochs).

Training data. Training samples are obtained by running fixed-coefficient VCD offline on 5,000 MSCOCO training images captioned with Qwen-VL-Chat and logging per-step top-50 logit pairs. This logging procedure yields ≈600,000 total decoding steps, of which ≈41,024 are object-generating steps used for label generation. We use a hash-based split (80% train / 10% val / 10% test) stratified by image ID.

4.5 Inference and Overhead

At each decoding step t, TAVCD first builds $\mathbf{x}_t$ from the current-step logit pair and a rolling 5-step buffer. It then evaluates f_θ in a single forward pass and substitutes α_t (Eq. 8, clipped) into the VCD contrastive adjustment (Eq. 1). We additionally apply Adaptive Plausibility Constraints (APC) (Leng et al., 2024): any token v with $p_{\text{pos}}(v) < \beta \max_u p_{\text{pos}}(u)$, i.e. $\mathrm{pos}(v) < \log\beta + \max_u \mathrm{pos}(u)$ ($\beta = 0.1$), is masked to $-\infty$, preventing contrastive amplification of intrinsically implausible tokens. The MLP forward pass over $\mathbf{x}_t \in \mathbb{R}^{10}$ adds less than 1 % wall-clock overhead relative to a single LVLM decoder layer. The LVLM itself is never updated; TAVCD wraps it as a frozen black box.

5 EXPERIMENTS

We empirically validate the analysis of Section 3 (measuring C^3 violation rates) and compare TAVCD against fixed-α VCD and alternative decoding strategies on two benchmark datasets across five LVLMs.

5.1 SETUP

Backbones. We evaluate five LVLMs spanning 2B to 7B parameters: BLIP-2 OPT-2.7B (Li et al., 2023a), Moondream2 (Korrapati, 2024), Qwen2-VL-2B-Instruct (Wang et al., 2024a), LLaVA-1.5-7B (Liu et al., 2023), and Qwen-VL-Chat 7B (Bai et al., 2023). All backbones are held frozen across all experiments.

Datasets. We evaluate on **MSCOCO 2014** (Lin et al., 2014), using object-presence ground truth from the MSCOCO 2014 segmentation annotations, and **PASCAL VOC 2007** (Everingham et al., 2010) test split, using class annotations from the detection task. Each image is captioned with the prompt "Please describe this image in detail." We evaluate on 500 images from each dataset, using greedy decoding with random seed 42 for reproducibility.

Metrics. We report the CHAIR metric (Rohrbach et al., 2018). $\text{CHAIR}_s = \frac{|\{\text{sentences with} \geq 1 \text{ hallucinated object}\}|}{|\{\text{sentences}\}|}$ measures the sentence-level hallucination rate (lower is better). $\text{CHAIR}_i = \frac{|\{\text{hallucinated object mentions}\}|}{|\{\text{all object mentions}\}|}$ measures the instance-level hallucination rate (lower is better). We additionally report Recall (fraction of ground-truth objects mentioned) and Precision (fraction of mentioned objects present in the ground-truth annotation).

Methods compared. Regular: standard autoregressive decoding. **VCD** (Leng et al., 2024): fixed contrastive coefficient $\alpha = 0.5$, noise step $T = 500$, $\beta = 0.1$. We verified that α=0.5 yields the best aggregate CHAIR across a sweep of $\alpha \in \{0.1, 0.2, \ldots, 1.0\}$ on a held-out validation split. **ICD** (Wang et al., 2024b): contrasts against a disturbed-instruction branch, with fixed α. **SID** (Huo et al., 2025): self-introspective contrastive decoding. **TAVCD** (ours): target-aware gate+magnitude MLP, no GT labels at inference time.

5.2 C^3 VIOLATION RATE ANALYSIS

We log per-step logit statistics during VCD inference on the 5,000-image MSCOCO training set and identify the worst-case pair (g^*, h^*) at each object-generating step. A step is a C^3 *violation* if $\delta_h \leq 0$ and $\Delta_h \geq 0$.

Across object-generating steps from the five evaluated backbones, we measure a violation rate of approximately **26%**, confirming that fixed-α VCD is counterproductive at roughly one in four critical moments. Figure 1 shows a concrete instance (with logit values in Appendix D, Figure 4) where α=0.5 amplifies the hallucination further, while TAVCD's predicted α_t=-1.5 (a negative coefficient) reverses the contrastive direction and recovers the ground-truth token at rank 1. This 26% violation rate motivates per-step adaptation. A method that correctly identifies and reverts violations can correct a structurally non-trivial share of hallucinations.

5.3 MAIN RESULTS

Table 1 reports CHAIR results on MSCOCO and PASCAL VOC.

Consistency across backbones and domains. TAVCD achieves the lowest or tied-for-lowest CHAIR_i and CHAIR_s on *all five backbones* on both datasets (the only method to do so simultaneously on both metrics). CHAIR_i margins over the next-best method range from 0.0 pp (Qwen-VL-Chat COCO and LLaVA-1.5 VOC, tied) to 1.0 pp (BLIP-2 COCO); CHAIR_s margins over fixed-α VCD reach up to 5.8 pp (Moondream2 VOC). Each per-backbone predictor is trained on MSCOCO only, yet generalises to PASCAL VOC (20 categories vs. COCO's 80). These results confirm that step-level logit statistics are informative proxies for the feasibility geometry even when the target token is unknown at inference time. On VOC, TAVCD and VCD both reduce LLaVA-1.5 CHAIR_i from 38.0 to 14.0, while ICD/SID remain near the baseline, confirming that coefficient adaptation and

Table 1: CHAIR results (500 images per dataset, greedy decoding). **Bold** = best per backbone (ties included). Lower $CHAIR_s$/$CHAIR_i$ is better; higher Recall/Prec is better. †ICD BLIP-2 captions are degenerate prompt-loops; $CHAIR_s$/$CHAIR_i$ imputed from the Regular run for that image (Recall set to observed $\approx 1\%$).

		MSCOCO				PASCAL VOC			
Backbone	**Method**	$CHAIR_s \downarrow$	$CHAIR_i \downarrow$	Rec↑	Prec↑	$CHAIR_s \downarrow$	$CHAIR_i \downarrow$	Rec↑	Prec↑
BLIP-2	Regular	9.6	6.4	47.9	93.3	19.6	16.6	70.9	79.7
	VCD	9.0	6.1	46.5	93.1	18.0	16.4	61.6	79.3
	ICD†	9.2	6.5	1.0	93.8	20.8	17.0	4.4	75.6
	SID	8.4	6.7	41.4	92.9	14.8	**16.2**	61.0	82.6
	TAVCD	**8.2**	**5.1**	45.6	93.1	**14.2**	**16.2**	49.0	79.0
Moondream2	Regular	33.6	15.5	70.9	81.1	45.2	23.3	89.3	63.5
	VCD	35.6	14.1	70.9	82.0	47.4	23.4	89.1	62.6
	ICD	34.2	16.1	68.0	80.5	44.6	23.5	85.9	63.7
	SID	34.4	15.8	70.8	80.5	44.6	22.6	89.4	64.3
	TAVCD	**31.0**	**14.0**	66.7	82.4	**41.6**	**22.5**	89.5	64.8
Qwen2-VL-2B	Regular	53.0	15.0	75.9	72.4	55.8	21.5	93.5	55.5
	VCD	56.6	15.9	75.7	70.5	62.4	23.7	92.5	53.5
	ICD	54.0	14.8	75.1	71.1	55.2	22.2	93.1	54.9
	SID	55.6	14.8	75.6	71.5	59.4	22.9	94.1	53.7
	TAVCD	**52.6**	**14.6**	74.4	72.9	**53.4**	**21.3**	93.5	58.3
LLaVA-1.5	Regular	83.6	36.2	72.9	51.1	81.2	38.0	95.5	43.1
	VCD	82.2	30.8	81.0	55.4	36.0	**14.0**	88.0	71.0
	ICD	85.6	31.1	77.3	55.4	79.6	36.2	94.6	41.3
	SID	86.0	31.3	80.5	55.6	80.6	36.6	95.3	41.2
	TAVCD	**82.0**	**30.5**	81.6	55.2	**35.6**	**14.0**	88.3	71.2
Qwen-VL-Chat	Regular	88.8	41.9	61.7	45.7	86.4	43.9	88.6	41.1
	VCD	74.6	22.8	81.1	58.9	61.2	30.8	90.8	48.2
	ICD	73.4	24.8	80.4	57.6	**57.4**	29.2	89.4	49.4
	SID	71.2	**22.7**	79.8	60.7	57.6	29.3	90.0	49.7
	TAVCD	**69.4**	**22.7**	77.3	60.5	**57.4**	**28.9**	91.1	49.4

negative-view engineering each target a distinct, orthogonal source of failure. See Appendix D for extended analysis.

6 Conclusion

We presented TAVCD, a target-aware per-step adaptive variant of Visual Contrastive Decoding. We formalised the C^3 condition and proved that fixed-α VCD fails to promote the target token whenever the condition is violated. We measured a 26% violation rate at object-generating steps across the evaluated backbones on MSCOCO. TAVCD (a lightweight gate+magnitude MLP, $\approx$12 K parameters) trained offline on logged decoding statistics reduces $CHAIR_i$ and $CHAIR_s$ relative to fixed-α VCD on both datasets, achieving the lowest or tied-for-lowest scores simultaneously across all five tested LVLMs on MSCOCO and PASCAL VOC.

Limitations. The MLP predictor requires offline-logged decoding statistics with ground-truth annotations for training. Our C^3 analysis addresses single-step pairwise correctability and does not directly account for compounding errors across the autoregressive trajectory. Each per-backbone predictor is trained on MSCOCO images, and while it generalises to VOC, broader evaluation across additional datasets and instruction-tuning styles is left for future work.

Future work. Promising directions include making the predictor trajectory-aware via a small recurrent module, replacing GT-supervised labels with unsupervised proxies, and pairing TAVCD's coefficient adaptation with engineered negative views (e.g. R-VCD (Lee & Song, 2025)) to address both orthogonal sources of failure jointly.

AI USE STATEMENT

In this work, we used generative AI tools to assist with literature search, LaTeX formatting, and code debugging. We have not used generative AI tools for core theoretical derivations, experimental design, or data analysis. All AI-assisted work has been reviewed by the authors. We take responsibility for the final content of this paper.

REPRODUCIBILITY STATEMENT

All experimental results use frozen LVLM backbones with greedy decoding and random seed 42. The TAVCD predictor architecture, feature specification, α grid, and training hyperparameters are fully described in Section 4 and Appendix C. Complete proofs of all theoretical results are provided in Appendix A.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23716–23736, 2022.

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, 2023. arXiv preprint arXiv:2308.12966.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020.

Boqi Chen, Xudong Liu, and Jianing Qiu. Mask what matters: Mitigating object hallucinations in multimodal large language models with object-aligned visual contrastive decoding, 2026. arXiv preprint arXiv:2602.11737.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. PaLI: A jointly-scaled multilingual language-image model. In *International Conference on Learning Representations*, 2023.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. PaLM: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

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.

Jingyuan Deng and Yujiu Yang. MaskCD: Mitigating LVLM hallucinations by image head masked contrastive decoding. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025.

Mark Everingham, Luc Van Gool, Christopher K I Williams, John Winn, and Andrew Zisserman. The Pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*, 88 (2):303–338, 2010.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. MME: A comprehensive evaluation benchmark for multimodal large language models, 2023. arXiv preprint arXiv:2306.13394.

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

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*, 2025.

Sihyeon Kim, Boryeong Cho, Sangmin Bae, Sumyeong Ahn, and Se-Young Yun. VACoDe: Visual augmented contrastive decoding, 2024. arXiv preprint arXiv:2408.05337.

Vikhyat Korrapati. Moondream2: A small vision language model. `https://moondream.ai`, 2024.

Jihoon Lee and Min Song. Retrieval visual contrastive decoding to mitigate object hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.

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*, 2024.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*, 2023a.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023b.

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*, 2023c.

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

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021.

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 *International Conference on Learning Representations*, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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 *European Conference on Computer Vision*, 2024b.

Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. Negative object presence evaluation (NOPE) to measure object hallucination in vision-language models. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, 2024.

Sean O'Brien and Mike Lewis. Contrastive decoding improves reasoning in large language models, 2023. arXiv preprint arXiv:2309.09117.

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, 2018.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented RLHF. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13088–13110, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and efficient foundation language models, 2023. arXiv preprint arXiv:2302.13971.

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. Evaluation and analysis of hallucination in large vision-language models, 2023. arXiv preprint arXiv:2308.15126.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution, 2024a. arXiv preprint arXiv:2409.12191.

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*, 2024b.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mPLUG-Owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 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*, 2024.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations*, 2024.

A Proofs of Theoretical Results

A.1 Proof of Proposition 1

For convenience, recall

$$M_h(\alpha) = \alpha\delta_h - \Delta_h, \qquad I_h^+ = \{\alpha > 0 : M_h(\alpha) > 0\}. \tag{11}$$

Since $M_h(\alpha)$ is affine in α, the feasible set is determined by the sign of δ_h and the positivity constraint $\alpha > 0$. We split on the sign of δ_h; the case numbers below refer to Proposition 1.

Sign $\delta_h > 0$ (strictly increasing in α). $M_h(\alpha) > 0 \iff \alpha > \alpha_h^\times = \Delta_h/\delta_h$.

- Case 1 ($\Delta_h > 0$): $\alpha_h^\times > 0$, so $I_h^+ = (\alpha_h^\times, +\infty)$.
- Case 2 ($\Delta_h \le 0$): $\alpha_h^\times \le 0$, so the constraint $\alpha > \alpha_h^\times$ holds for all $\alpha > 0$, giving $I_h^+ = (0, +\infty)$.

Sign $\delta_h = 0$ (constant margin). $M_h(\alpha) \equiv -\Delta_h$ for all α.

- Case 3 ($\Delta_h < 0$): $M_h(\alpha) = -\Delta_h > 0$ for all $\alpha > 0$, giving $I_h^+ = (0, +\infty)$.
- Case 5 with $\delta_h = 0$ ($\Delta_h \ge 0$): $M_h(\alpha) \le 0$ for all $\alpha > 0$, giving $I_h^+ = \emptyset$. This includes $(\delta_h, \Delta_h) = (0, 0)$, where $M_h \equiv 0$ and the strict inequality never holds.

Sign $\delta_h < 0$ (strictly decreasing in α). $M_h(\alpha) > 0 \iff \alpha < \alpha_h^\times = \Delta_h/\delta_h$.

- Case 4 ($\Delta_h < 0$): $\alpha_h^\times = \Delta_h/\delta_h > 0$ (negative divided by negative), so $I_h^+ = (0, \alpha_h^\times)$.
- Case 5 with $\delta_h < 0$ ($\Delta_h \geq 0$): $\alpha_h^\times = \Delta_h/\delta_h \leq 0$, so no positive α satisfies $\alpha < \alpha_h^\times$, giving $I_h^+ = \emptyset$.

The three signs of δ_h are exhaustive, and the subcases above are exactly the five cases of Proposition 1. □

A.2 Proof of Theorem 1

The five cases of Proposition 1 partition $\{(\delta_h, \Delta_h)\}$, and only case 5 ($\delta_h \leq 0$ and $\Delta_h \geq 0$) yields $I_h^+ = \emptyset$; cases 1–4 each give a non-empty interval. Therefore:

$$I_h^+ = \emptyset \iff \delta_h \leq 0 \text{ and } \Delta_h \geq 0.$$

The second claim follows immediately: if any $h \in \mathcal{H}$ violates C^3, then $\mathcal{A}^+(g) \subseteq I_h^+ = \emptyset$, so $\mathcal{A}^+(g) = \emptyset$. □

A.3 Proof of Theorem 2

Without a C^3 violation, each I_h^+ falls into one of the non-blocking cases of Proposition 1 (cases 1–4):

- Case 1 ($\delta_h > 0, \Delta_h > 0$): $I_h^+ = (\alpha_h^\times, +\infty)$, imposing lower bound $\alpha > \alpha_h^\times$.
- Case 2 ($\delta_h > 0, \Delta_h \leq 0$) and Case 3 ($\delta_h = 0, \Delta_h < 0$): $I_h^+ = (0, +\infty)$, imposing no constraint beyond $\alpha > 0$.
- Case 4 ($\delta_h < 0, \Delta_h < 0$): $I_h^+ = (0, \alpha_h^\times)$, imposing upper bound $\alpha < \alpha_h^\times$.

The intersection $\mathcal{A}^+(g) = \bigcap_{h \in \mathcal{H}} I_h^+$ is therefore:

$$\mathcal{A}^+(g) = \Big(\max_{h:\delta_h>0,\Delta_h>0} \alpha_h^\times,\ \min_{h:\delta_h<0,\Delta_h<0} \alpha_h^\times \Big) \cap (0, +\infty) = (\underline{\alpha}(g), \bar{\alpha}(g)) \cap (0, +\infty).$$

The disjunction in Eq. 7 is the negation of $\mathcal{A}^+(g) \neq \emptyset$: either some h causes $I_h^+ = \emptyset$ (pairwise blocking, Theorem 1) or the interval is empty due to $\underline{\alpha} \geq \bar{\alpha}$ (window collapse). □

A.4 Proof of Proposition 2

Without the sign restriction $\alpha > 0$, the feasible set $I_h = \{\alpha \in \mathbb{R} : M_h(\alpha) > 0\}$. Since $M_h(\alpha) = \alpha\delta_h - \Delta_h$:

- $\delta_h > 0$: M_h is increasing; $M_h(\alpha) > 0 \iff \alpha > \alpha_h^\times$. So $I_h = (\alpha_h^\times, +\infty)$.
- $\delta_h < 0$: M_h is decreasing; $M_h(\alpha) > 0 \iff \alpha < \alpha_h^\times$. So $I_h = (-\infty, \alpha_h^\times)$.
- $\delta_h = 0, \Delta_h < 0$: $M_h(\alpha) \equiv -\Delta_h > 0$ for all α. So $I_h = \mathbb{R}$.
- $\delta_h = 0, \Delta_h \geq 0$: $M_h(\alpha) \equiv -\Delta_h \leq 0$ for all α. So $I_h = \emptyset$.

□

A.5 Proof of Corollary 1

Fix any $\alpha_0 \in \mathbb{R}$. We construct a witness decoding configuration in which $\alpha_0 \notin \mathcal{A}^*(g)$.

Case $\alpha_0 > 0$. Choose $\delta_h < 0$ and $\Delta_h > 0$, so $\alpha_h^\times = \Delta_h/\delta_h < 0 < \alpha_0$. By Proposition 2, $I_h = (-\infty, \alpha_h^\times)$, so $\alpha_0 \notin I_h$, hence $\alpha_0 \notin \mathcal{A}^*(g) \subseteq I_h$.

Case $\alpha_0 \leq 0$. Choose $\delta_h > 0$ and $\Delta_h > 0$ with $\alpha_h^\times > \alpha_0$ (e.g., $\alpha_h^\times = \alpha_0 + 1$). By Proposition 2, $I_h = (\alpha_h^\times, +\infty)$, so $\alpha_0 < \alpha_h^\times$ implies $\alpha_0 \notin I_h$; since $\mathcal{A}^*(g) \subseteq I_h$, we have $\alpha_0 \notin \mathcal{A}^*(g)$.

In both cases, the logit geometry (δ_h, Δ_h) is realisable as a concrete decoding step configuration. Therefore no fixed α_0 is universally corrective. □

B ORACLE LABELLING PIPELINE

This section gives the full procedure for generating per-step training labels for the gate and magnitude heads of f_θ. Labels are produced entirely offline from pre-logged decoding data; no additional LVLM inference is needed during the labelling stage.

Step 1: Offline VCD logging. We run fixed-coefficient VCD (α=0.5, β=0.1) on MSCOCO 2014 training images (a split established before any evaluation images were selected, ensuring zero overlap) and record, at every decode step t: the top-K ($K = 50$) token IDs and logit values from both the positive branch pos and the negative branch neg, the greedy-decoded token y_t, the step index t, and the total caption length T. In our experiments this produces $\approx$600,000 total step records, of which $\approx$41,024 are object-generating steps; the remainder are discarded for label-generation purposes but are used to compute rolling-history features during training.

Step 2: Object-token detection and step typing. A step t is an *object-generating step* if y_t is the first sub-word token that begins a COCO object-category string in the partial caption $y_1 \dots y_t$, i.e. the step at which CHAIR's object mention is emitted. Detection uses case-insensitive prefix matching against the 80 COCO category names after stripping punctuation and possessives. Object-generating steps are further divided:

- **Hallucinated-object step:** the matched category is *not* present in the GT instance annotation for the image. The model is generating an object that should not appear.
- **Correct-object step:** the matched category *is* present in the GT instance annotation. The model is faithfully naming a visible object.
- **Non-object step:** all other steps (punctuation, articles, verbs, continuation sub-words, etc.). These receive gate label $y_g = 0$ with no magnitude label; they contribute only to the gate-off class during training.

Step 3: Grid search for the oracle α^*. For each object-generating step, we select a ground-truth target token g^* and evaluate its CD-adjusted rank under every candidate $\alpha \in \mathcal{G}$:

$$L_\alpha^{\mathrm{CD}}[i] = (1+\alpha)\cdot \mathrm{pos}[i] - \alpha \cdot \mathrm{neg}[i], \quad \mathrm{rank}_\alpha(v) \triangleq \left|\{i : L_\alpha^{\mathrm{CD}}[i] \geq L_\alpha^{\mathrm{CD}}[v]\}\right|. \tag{12}$$

Both the positive and negative vectors are evaluated at the stored top-K positions of pos (same 50 positions used for feature extraction). Both step types share one objective, *target promotion*: the label is the largest grid value that places g^* at rank-1. They differ only in the choice of g^*.

- **Hallucinated-object step.** The generated token y_t is not a GT object, so the target is the GT token the model ranks highest: g^* is the GT-object token with the highest pos logit among the stored top-K positions. A top-K token counts as a GT-object token if it is the leading sub-word of an annotated category name or of one of its CHAIR synonyms (e.g. "man" for *person*). Collect the *feasible set* $F_{\mathrm{hal}} = \{\alpha \in \mathcal{G} : \mathrm{rank}_\alpha(g^*) = 1\}$ (values that promote g^* above y_t and every other logged competitor). Select $\alpha^* = \max F_{\mathrm{hal}}$. If no GT-object token appears in the top-K, the step has no target and is loss-masked.
- **Correct-object step.** Here $g^* = y_t$ (the generated token is itself a GT object). Collect the feasible set $F_{\mathrm{cor}} = \{\alpha \in \mathcal{G} : \mathrm{rank}_\alpha(g^*) = 1\}$ (values preserving g^* at rank-1). Select $\alpha^* = \max F_{\mathrm{cor}}$ (the widest amplification that keeps the ground-truth token at rank-1).

If F is empty for an object-generating step, the step is *structurally uncorrectable* (the empty-window scenario of Theorem 2) and receives zero weight in both the gate and magnitude objectives. The step is excluded entirely from the training loss; it is neither a positive nor a negative example for the gate head.

Step 4: Label assignment. Each step receives a label pair $(y_g,\, y_m)$:

- Non-object step: $y_g = 0$, y_m undefined (excluded from magnitude loss).
- Object-generating step with feasible α^*: $y_g = 1$, $y_m = \mathrm{idx}(\alpha^*)$, where $\mathrm{idx}(\alpha^*)$ is the 0-based position of α^* in the sorted 40-value grid $\mathcal{G} = \{-4.0, -3.8, \dots, -0.2, 0.2, \dots, 4.0\}$.
- Object-generating step that is structurally uncorrectable or has no GT-object token in the top-K: loss-masked; excluded from both gate and magnitude objectives.

Class statistics. Non-object steps account for $\approx 93\%$ of all steps, creating severe class imbalance on the gate task. Among the $\approx 41{,}024$ object-generating steps, approximately 55% are hallucinated-object and 45% are correct-object steps. Of hallucinated-object steps, $\approx 85\%$ have at least one feasible grid point; the remaining $\approx 15\%$ are structurally uncorrectable and are masked. This class imbalance motivates the focal BCE loss ($\gamma = 2$, $w_{\text{pos}} = 0.95$) applied to the gate head (Section 4.4).

C Implementation Details

Full feature specification. Table 2 gives the complete definition of the 10 features extracted at each decoding step. All features are computed from the top-K ($K = 50$) tokens of the positive branch; negative-branch statistics are evaluated at the same token positions. Z-score normalisation uses per-feature mean and standard deviation computed from the training split (80% of 41,024 steps).

Table 2: TAVCD feature specification (10 features per step).

Feature	Formula	Captures
H_{pos}	$-\sum_k p_k \log p_k$ (pos top-K)	Positive branch uncertainty
H_{neg}	$-\sum_k q_k \log q_k$ (neg top-K)	Negative branch uncertainty
ΔH	$H_{\text{pos}} - H_{\text{neg}}$	Entropy shift from CD
KL	$\text{KL}(p_{\text{pos}} \Vert p_{\text{neg}})$	Distributional divergence
$p_{\text{pos}}^{\max}$	$\max_k \text{softmax}(\text{pos}_k)$	Top-1 positive confidence
$p_{\text{neg}}^{\max}$	p_{neg} at top-1 pos position	Neg competition at pos argmax
gap_1	$\text{pos}[0] - \text{pos}[1]$	Sharpness of positive top-1
$\mu_{\Delta H}^5$	mean of ΔH over last 5 steps	Rolling entropy trend
μ_{KL}^5	mean of KL over last 5 steps	Rolling divergence trend
t/T	step index / total caption length	Relative position in caption

Inference algorithm. Algorithm 1 gives the full TAVCD inference procedure at a single decoding step.

Algorithm 1 TAVCD inference at decoding step t

Require: Frozen LVLM, image v, noised image v', history $\mathbf{y}_{<t}$, MLP f_θ, feature stats $\mu, \hat{\sigma}$, buffer B of last 5 steps, α_{clip}, β, grid $\mathcal{G} = \{c_1, \ldots, c_{40}\}$.

$\boldsymbol{\ell}_{\text{pos}} \leftarrow \text{LVLM}(v, \mathbf{y}_{<t})$
$\boldsymbol{\ell}_{\text{neg}} \leftarrow \text{LVLM}(v', \mathbf{y}_{<t})$
Extract raw $\mathbf{x}_t$ from $\boldsymbol{\ell}_{\text{pos}}, \boldsymbol{\ell}_{\text{neg}}, B$
$\mathbf{x}_t^z \leftarrow (\mathbf{x}_t - \mu)/\hat{\sigma}$ *(z-score)*
$(s_t, \boldsymbol{\ell}_t) \leftarrow f_\theta(\mathbf{x}_t^z)$
$\alpha_t \leftarrow \sigma(s_t) \times \sum_k \text{softmax}(\boldsymbol{\ell}_t)[k] \cdot c_k$
$\alpha_t \leftarrow \text{clip}(\alpha_t,\ -\alpha_{\text{clip}},\ \alpha_{\text{clip}})$
$\boldsymbol{\ell}_{\text{adj}} \leftarrow \boldsymbol{\ell}_{\text{pos}} + \alpha_t(\boldsymbol{\ell}_{\text{pos}} - \boldsymbol{\ell}_{\text{neg}})$
APC: $\boldsymbol{\ell}_{\text{adj}}(v) \leftarrow -\infty$ for all v with $\text{softmax}(\boldsymbol{\ell}_{\text{pos}})(v) < \beta \max_w \text{softmax}(\boldsymbol{\ell}_{\text{pos}})(w)$
Update buffer $B \leftarrow B \cup \{(\Delta H_t, \text{KL}_t)\}$, drop oldest if $|B| > 5$
$y_t \leftarrow \arg\max(\boldsymbol{\ell}_{\text{adj}})$ *(greedy; replace with softmax sample for stochastic decoding)*

Ensure: Token y_t

Training hyperparameters. Optimiser: Adam, lr $= 10^{-3}$, weight decay $= 10^{-4}$. Scheduler: cosine annealing, $T_{\max} = 12$ epochs, $\eta_{\min} = 10^{-5}$. Early stopping patience: 3 epochs. Batch size: 512. Focal loss: $\gamma = 2.0$, $w_{\text{pos}} = 0.95$. Magnitude cross-entropy label smoothing: $\epsilon = 0.1$. $\lambda_{\text{mag}} = 1.0$. All experiments use random seed 42.

Reproducibility notes. All LVLM backbones are loaded in bfloat16 and held frozen. Inference uses greedy decoding (temperature $= 0$) with fixed seed 42 for all reported results. The α grid $\mathcal{G}$ spans $[-4.0, 4.0]$ in steps of 0.2, excluding 0.0 (40 values total).

D Extended Discussion

A worked C^3 violation. Figure 4 expands the example of Figure 1 with the logged logit values of both branches and the adjusted logits under fixed and adaptive α.

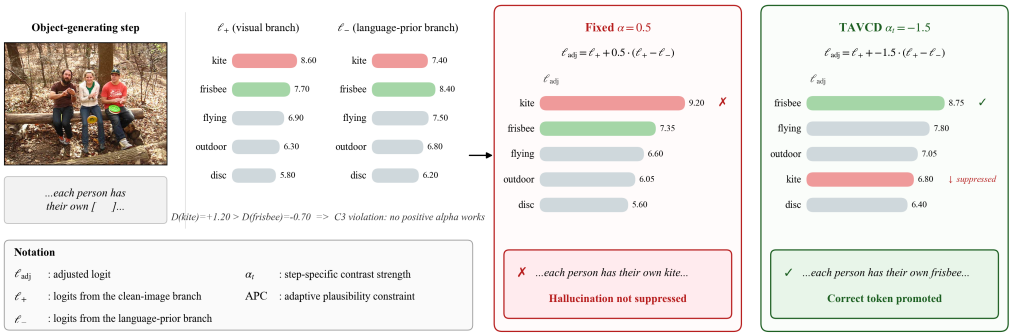

Figure 4: C^3 violation at a representative object-generating step (MSCOCO image 459195, backbone LLaVA-1.5). *Left*: the pos (visual) and neg (language-prior) distributions. The visual branch strongly distinguishes "kite" ($D(\text{kite}) = +1.20$) while the language-prior branch favours "frisbee" ($D(\text{frisbee}) = -0.70$): $\delta_h = D(g) - D(h) < 0$, so the feasible set $I_h^+ = \emptyset$: *no positive α can promote "frisbee" above "kite"*. *Centre*: fixed $\alpha = 0.5$ amplifies the hallucination further. *Right*: TAVCD predicts $\alpha_t = -1.5$ (a negative coefficient), reversing the contrastive direction and promoting "frisbee" to rank 1.

Gate activation and object-generating steps. The gate head outputs values near zero on the majority ($\approx 93\%$) of continuation steps and activates primarily at object-generating steps, where the model first generates a new object token. This is consistent with the C^3 theory: the feasibility geometry is only non-trivial at steps where the positive and negative distributions differ substantially with respect to a candidate hallucination object h.

Coefficient adaptation vs. negative-view engineering. TAVCD uses the standard noise-corrupted negative view from VCD but adapts α per step. On all 10 backbone$\times$dataset combinations, TAVCD outperforms fixed-α VCD. This consistent outperformance confirms that coefficient adaptation and negative-view engineering are orthogonal: a stronger negative view (as in R-VCD (Lee & Song, 2025) or MaskCD (Deng & Yang, 2025)) could be combined with TAVCD's gate+magnitude head to obtain further gains.

E Qualitative Results

We present qualitative examples with the source image for each case. All captions are taken verbatim from the model output; CHAIR_i values are from the per-image CHAIR evaluation. The cases (S1–S5) illustrate language prior suppression.

E.1 Success Cases: Language Prior Suppression



S1: Ocean scene (QID 391214) | Backbone: LLaVA-1.5 | GT: *surfboard*

Regular (CHAIR_i=0.83): "...an *inflatable surfboard* floating on top of the waves. *Two people* are visible in the water...There's also a *small part of land*...". The model hallucinated "inflatable" and inserted extra people, driven by the ocean-scene co-occurrence prior.

TAVCD (CHAIR_i=0.0): "A surfer is riding one of these waves on top of a *white surfboard*...The ocean extends all the way to the horizon...". Only the correct object appears; no COCO co-occurrences were hallucinated.



S2: Food tray (QID 557357) | Backbone: LLaVA-1.5 | GT: *broccoli*

Regular (CHAIR$_i$=0.80): "...broccoli. A single square piece of cooling *aluminum foil*...near the edge of this entree. The items consist of not only the food being served but also possibly their *tools for preparation*..."; the kitchen-utensil prior adds foil and tools absent from the scene.

TAVCD (CHAIR$_i$=0.0): "...a pan with a mix of vegetables, including broccoli, cauliflower, and Brussels sprouts...". No extra COCO objects appear; the tray vegetables are correctly identified.



S3: City street with construction wall (QID 113935) | Backbone: LLaVA-1.5 | GT: *car, parking meter, traffic light*

Regular (CHAIR$_i$=1.00): "The scene features a slimy-looking brick building with illuminated walls...A red light shines through *wall pillar light fixtures*...*People can be seen* nearby...". The glowing red paint panels on the construction hoarding are misread as decorative fixtures, and people are hallucinated from the urban-street prior.

TAVCD (CHAIR$_i$=0.0): "The scene features a city street with a red brick building wall...A black metal pole with a blue sign attached to it stands by the building. Several signs are visible...". The model correctly describes street infrastructure without hallucinating any COCO objects.



S4: Three friends with frisbees (QID 459195) | Backbone: LLaVA-1.5 | GT: *frisbee, person*

Regular (CHAIR$_i$=0.71): "...A green frisbee can be seen attached to one of the men's hat. *Each person has their own kite*, with winds blowing through the area overlooking a field, inviting them to fly their kites together...". The frisbee $\rightarrow$ flying-toy co-occurrence introduces a kite per person; no kites exist in the scene.

TAVCD (CHAIR$_i$=0.0): "The scene features three friends sitting on a tree trunk, with two of them holding frisbees. A man wearing a hat sits between the other two people...There are two frisbees visible...". The model correctly identifies frisbees only; the kite prior is suppressed.



S5: Person on carpeted floor (QID 567886) | Backbone: LLaVA-1.5 | GT: *book, person, teddy bear*

Regular (CHAIR$_i$=0.71): "...there are a *cup, spoon and fork* placed in a corner, a *television* located near the left side of the room..."; the indoor relaxation prior generates kitchen utensils and a TV that are absent from the scene.

TAVCD (CHAIR$_i$=0.0): "...two black flip-flops placed near each other. There are also several *books* scattered on the floor, indicating a potential reading session...". The output describes only what is visible (books, person, teddy bear); the utensil and television co-occurrence priors are suppressed.