VIECO: VISUAL ENERGY-CONSERVATION DECODING FOR TRAINING-FREE HALLUCINATION MITIGATION IN VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large vision-language models (VLMs) often generate visually unsupported yet linguistically plausible claims, a failure mode known as hallucination. Existing training-free mitigation methods contrast image-conditioned logits with distorted-image or null-image branches, or steer latent states toward visual features, but rarely ask a more basic question: *where does the probability of a visual claim come from?* We propose ViECo, a training-free framework that views VLM hallucination as *visual energy non-conservation*. ViECo decomposes image-conditioned logits into a low-rank language-prior subspace and an orthogonal visual-residual component, then penalizes visual claims whose probability is dominated by prior projection with weak residual support and temporal energy spill. The method requires no parameter updates, external detectors, or image captions. On LLaVA-1.5-7B, ViECo achieves 84.40 POPE-Adversarial F1 (+7.95 over vanilla decoding), reduces CHAIR_s from 46.60 to 30.60, and obtains the highest MMHal score (3.12) among all training-free baselines, with only $1.45\times$ overhead and the lowest prompt-variation sensitivity.

1 INTRODUCTION

Large vision-language models (VLMs) have become the dominant interface for open-ended visual understanding (Liu et al., 2023), with growing deployment in medical diagnosis (Li et al., 2023a), autonomous driving (Tian et al., NA), and robotic control (Zitkovich et al., 2023). Despite their versatility, VLMs frequently generate visually unsupported yet linguistically fluent claims, a failure known as *hallucination*. In safety-critical applications, such unfaithful outputs can be far more harmful than explicit uncertainty, making hallucination mitigation a first-order priority.

Training-free mitigation is attractive because it avoids costly retraining while generalizing across architectures. Existing approaches include *contrastive decoding* (Lyu et al., 2024; Zhang et al., 2025; Lyu et al., 2024), *attention/grounding methods* (Zhang et al., 2026; Park et al., 2025b; Zhou et al., 2025), and *representation steering* (An et al., 2024; Liu et al., 2025; Dastmalchi et al., 2026; Li et al., 2025a). Although effective, these methods rely on indirect proxies such as distributional shifts, attention scores, or latent directions, without asking *where does the probability of a visual claim come from?* None decomposes a token's probability into its constitutive sources to determine whether confidence originates from visual evidence or a language prior.

Consider an image of a person standing beside a horse. The token *riding* may remain probable because person–horse co-occurrence activates a language-prior manifold, while *standing beside* requires genuine visual evidence about pose. A contrastive branch that merely blurs the image cannot distinguish whether probability is visually grounded or prior-projected. What is needed is a direct source-level decomposition within logit space. Two observations enable this: (i) repeatedly sampled logits from perturbed inputs concentrate on a low-dimensional linear subspace that reflects the language prior; and (ii) the softmax classifier admits an energy view (LeCun et al., 2006) in which a token's low energy is a claim of commitment that should stay compatible with the next decoding state. Together they yield a precise criterion: a visual claim hallucinates when its low energy is dominated by the language-prior subspace rather than supported by visual residual evidence, and when this commitment fails to persist over one decoding step.

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.

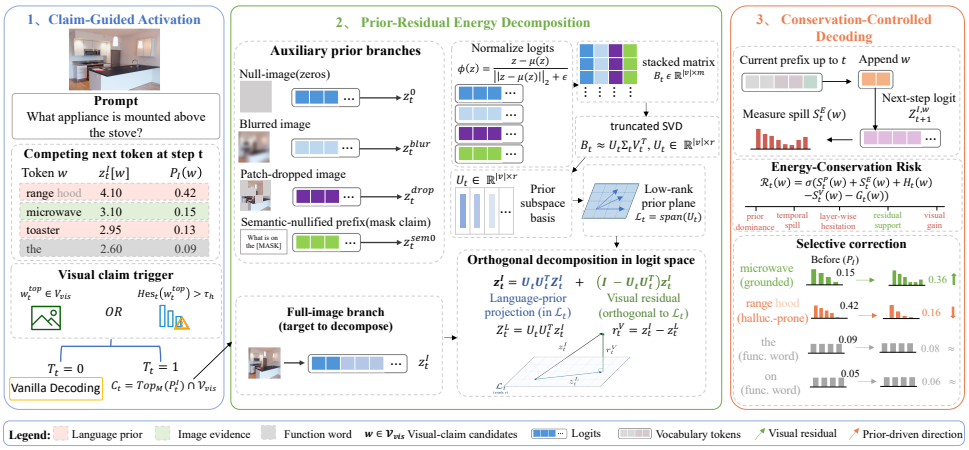


Figure 1: Overview of ViECo. **Stage 1**: Claim-Guided Activation triggers energy analysis only on visual-claim tokens or hesitant decisions. **Stage 2**: Prior-Residual Energy Decomposition estimates a low-rank prior subspace U_t via truncated SVD and decomposes logits into prior projection $\mathbf{z}_t^L$ and visual residual $\mathbf{r}_t^V$. **Stage 3**: Conservation-Controlled Decoding computes risk $R_t(w)$ from temporal spill, prior dominance, and residual support, then selectively corrects hallucination-prone candidates.

We propose ViECo (**Vi**sual **E**nergy-**Co**nservation Decoding), a training-free framework that views VLM hallucination as *visual energy non-conservation*. As shown in Figure 1, ViECo first activates energy analysis on visual-claim tokens or hesitant decisions, then decomposes logits into a low-rank prior projection and an orthogonal visual residual, and finally applies conservation-controlled correction penalizing candidates with high prior dominance, low residual support, and unstable temporal energy. The method requires no parameter updates, external detectors, or captions, and introduces only two continuous hyperparameters (λ and τ_h).

Our contributions are:

- We formulate VLM hallucination as *visual energy non-conservation*, linking hallucination to the source of logit energy rather than raw confidence or distributional shifts.
- We introduce a low-rank prior-residual decomposition that separates logits into a language-prior projection and a visual residual, diagnosing whether a claim is grounded or prior-dominated.
- We design a conservation-controlled decoder integrating residual support, prior dominance, hesitation, and temporal spill into a single-weight rule. On LLaVA-1.5-7B, ViECo improves POPE-Adv F1 by +7.95, reduces CHAIR_s by 16.0, and achieves the best MMHal score with only $1.45\times$ overhead.

2 Related Work

Training-free hallucination decoding. VCD contrasts logits under original and distorted images to reduce prior reliance (Lyu et al., 2024); OPERA penalizes over-trusted attention during beam search (Zhang et al., 2026). Follow-up work refines contrastive branches via image-biased projectors, cross-image references, generative feedback, or dynamic routing (Chen et al., 2026; Zhao et al., 2025; Zhang et al., 2025; Suo et al., 2025; Lyu et al., 2024), and enhances visual signal saliency during decoding (Yin et al., 2025). These methods rely on pairwise distributional comparisons or heuristic penalties. ViECo instead estimates a low-rank prior subspace and measures whether each claim retains residual energy orthogonal to it.

Representation steering and visual preservation. An alternative line intervenes in hidden states or visual information flow. An et al. (2024) reweight visual attention against language dominance; VTI steers toward truthful latent directions (Liu et al., 2025); Nullu projects out hallucination-prone weight directions (Dastmalchi et al., 2026); visual information steering counters layer-wise visual decay (Li et al., 2025a); ICT performs image-object cross-level trusted intervention (Chen et al.,

2025a); and latent-space steering has been extended to hallucination detection more broadly (Park et al., NA). These methods reveal that hallucination has roots in representation dynamics, but often require identifying faithful attention heads or stable steering directions. ViECo operates at the logit-energy level without region masks or external detectors.

Logit geometry and energy-based views. Empirical studies of language-model output show that sequences of logits are effectively low-rank once linguistic and prefix context are fixed, and the softmax classifier itself admits an energy interpretation (LeCun et al., 2006; Wainwright & Jordan, 2008). Prior work in hallucination detection has explored token-level localization (Tian et al., 2026), selective contrastive decoding (Park et al., 2025a), and image-head-masked contrasts (Deng & Yang, 2025). ViECo transfers these insights to VLMs at the logit-source level: the image is external context, the language prior is internal knowledge, and hallucination arises when a claim's low energy is dominated by the latter and fails to persist over one decoding step.

Attention/grounding intervention. FarSight repairs causal attention flow during decoding (Zhou et al., 2025); assembly of global and local attention improves discriminative image features (An et al., 2024). ViECo is complementary to these methods: it operates at the logit-energy level and could be combined with attention-level interventions.

3 Preliminary: Visual Claims as Energy Commitments

Let M_θ be a frozen autoregressive VLM. Given image I, instruction q, and prefix $y_{<t}$, the model produces logits $\mathbf{z}_t^I = f_\theta(I, q, y_{<t}) \in \mathbb{R}^{|\mathcal{V}|}$ with next-token distribution $p_t^I = \mathrm{softmax}(\mathbf{z}_t^I)$. We use "visual claim" to denote a token or phrase asserting visually verifiable content (objects, attributes, relations, counts, states).

Definition 1 (Logit-derived energy)**.** *For a candidate token $w \in \mathcal{V}$, its logit-derived energy under image I is*

$$E_t^I(w) = -\mathbf{z}_t^I[w]. \tag{1}$$

Equivalently, $p_t^I(w) \propto \exp(-E_t^I(w))$. Low energy corresponds to high model preference.

This is the standard energy-based interpretation of a softmax classifier: each token receives a compatibility score, and its negative acts as energy (LeCun et al., 2006). Hallucination can be viewed as a false low-energy state: a visually unsupported claim is made artificially easy to generate.

3.1 Prior Manifold and Visual Residual

We use auxiliary branch logits $\mathbf{z}_t^0$, $\mathbf{z}_t^{blur}$, $\mathbf{z}_t^{drop}$, and $\mathbf{z}_t^{sem0}$ to estimate the language-prior component that persists when visual evidence is weakened. These branches do not represent ground-truth negatives; they provide local samples of a prior-dominated logit manifold.

Assumption 1 (Approximate low-rank prior)**.** *At a decoding step t, the non-visual component of the logit vector lies near a rank-r subspace $\mathcal{L}_t \subset \mathbb{R}^{|\mathcal{V}|}$ with $r \ll |\mathcal{V}|$. The auxiliary branch logits span an estimate $\hat{\mathcal{L}}_t$ of this subspace.*

This assumption is motivated by two observations. First, when only visual evidence is weakened (blur, patch drop, semantic null) while textual context is preserved, the resulting logit vectors share a strong linguistic and prefix-driven component. Second, autoregressive language models are known to produce output distributions that concentrate on a small number of directions given a fixed prefix (Wainwright & Jordan, 2008). Let $U_t \in \mathbb{R}^{|\mathcal{V}| \times r}$ be an orthonormal basis of $\hat{\mathcal{L}}_t$, computed by QR decomposition or truncated SVD over normalized branch logits. We decompose

$$\mathbf{z}_t^I = \underbrace{U_t U_t^\top \mathbf{z}_t^I}_{\mathbf{z}_t^L} + \underbrace{(\mathbf{I} - U_t U_t^\top)\mathbf{z}_t^I}_{\mathbf{r}_t^V}. \tag{2}$$

We call $\mathbf{z}_t^L$ the language-prior projection and $\mathbf{r}_t^V$ the visual residual, measuring the part of the image-conditioned score not explained by the prior manifold.

Definition 2 (Prior and residual scores). *For a token w, write $p_w = \mathbf{z}_t^L[w]$, $r_w = \mathbf{r}_t^V[w]$, and $D_t(w) = |p_w| + |r_w| + \epsilon$. Prior dominance is $\mathrm{PR}_t(w) = |p_w|/D_t(w)$, and residual support is $\mathrm{VR}_t(w) = \rho_t(\mathbf{r}_t^V, w)$. Here $\mathcal{C}_t$ is the set of candidate visual-claim tokens, ϵ is a small constant, and for any candidate-indexed scalar score $a_t : \mathcal{C}_t \to \mathbb{R}$, $\rho_t(a, w)$ denotes rank normalization over $\mathcal{C}_t$:*

$$\rho_t(a, w) = |\mathcal{C}_t|^{-1} \sum_{u \in \mathcal{C}_t} \mathbf{1}[a_t(w) \geq a_t(u)]\,. \tag{3}$$

We also write $\mathrm{RN}_{\mathcal{C}_t}(a_t)(w)$ for $\rho_t(a, w)$. In the residual-support definition, $a_t(u) = \mathbf{r}_t^V[u]$. PR measures prior dominance; VR measures visual residual support.

3.2 Energy Conservation Along Decoding

A visual claim is not merely a single token. Once the model commits, the subsequent prefix should remain compatible with that commitment. We define a temporal inconsistency score. For candidate w, let $\mathbf{z}_{t+1}^{I,w} = f_\theta(I, q, y_{<t} \circ w)$ be the next-step logits after appending w:

$$\mathrm{SE}_t(w) = \left| -\mathbf{z}_t^I[w] + \log \sum_{v \in \mathcal{V}} \exp(\mathbf{z}_{t+1}^{I,w}[v]) \right|. \tag{4}$$

A high SE indicates that the model's low-energy commitment is not stably maintained in the next state. The exact form can use full-vocabulary or top-K logits.

Definition 3 (Visual energy non-conservation). *A visual claim token w exhibits visual energy non-conservation if it simultaneously has high image-conditioned probability, high prior dominance $\mathrm{PR}_t(w)$, low visual residual support $\mathrm{VR}_t(w)$, and high temporal spill $\mathrm{SE}_t(w)$.*

Proposition 1 (Residual faithfulness). *Let $\mathbf{z}_t^I = \mathbf{z}_t^L + \mathbf{z}_t^V + \xi_t$, where $\mathbf{z}_t^L \in \mathcal{L}_t$, $\mathbf{z}_t^V \perp \mathcal{L}_t$. If $\|P_{\hat{\mathcal{L}}_t} - P_{\mathcal{L}_t}\|_2 \leq \delta$, then*

$$\|(\mathbf{I} - P_{\hat{\mathcal{L}}_t})\mathbf{z}_t^I - \mathbf{z}_t^V\|_2 \leq \delta(\|\mathbf{z}_t^L\|_2 + \|\mathbf{z}_t^V\|_2) + \|\xi_t\|_2. \tag{5}$$

The proof is in Appendix A.4. This bound shows the estimated residual faithfully approximates true visual evidence when the prior subspace estimate is accurate.

Theorem 1 (One-step correction margin). *Let w^+ be a grounded visual candidate and w^- an unsupported candidate. Suppose $\mathrm{VR}_t(w^+) - \mathrm{VR}_t(w^-) \geq m_v$, $\mathrm{PR}_t(w^-) - \mathrm{PR}_t(w^+) \geq m_p$, and $\mathrm{SE}_t(w^-) - \mathrm{SE}_t(w^+) \geq m_s$. For corrected logits*

$$\begin{aligned} \tilde{z}_t(w) = \mathbf{z}_t^I[w] + \lambda\big(&\mathrm{VR}_t(w) \\ &- \mathrm{PR}_t(w) - \mathrm{SE}_t(w)\big), \end{aligned} \tag{6}$$

there exists $\lambda > 0$ such that $\tilde{z}_t(w^+) > \tilde{z}_t(w^-)$ whenever the initial logit gap is smaller than $\lambda(m_v + m_p + m_s)$.

The proof is in Appendix A.6. This theorem justifies ViECo's local intervention: if the diagnostics separate grounded from unsupported candidates, a single-strength correction reverses the ranking.

4 Method

ViECo is a plug-in decoding wrapper for frozen VLMs. Operationally, it instantiates the energy view in Definition 1 and the non-conservation criterion in Definition 3. The method follows a compact three-stage logic: **activate on visual claim candidates**, **decompose their logits into prior and visual evidence**, and **decode under an energy-conservation constraint** (Algorithm 1). This organization keeps the method lightweight while preserving the key signals needed to distinguish grounded visual claims from prior-driven hallucinations.

4.1 Claim-Guided Activation

The first module decides *when* ViECo should instantiate the visual-claim test from Definition 3. Running auxiliary branches at every step is unnecessary and can disturb stable language generation. We therefore activate energy analysis only when the next-token decision is visually meaningful or internally unstable.

Layer-wise hesitation. The hesitation signal measures whether a candidate is stably preferred across decoder layers:

$$\mathrm{Hes}_t(w) = \mathrm{Var}_{l \in L_{\mathrm{probe}}} \left(\mathrm{rank}(W_o h_{t,l})[w] \right), \tag{7}$$

where $h_{t,l}$ is the hidden state at layer l and W_o is the language-model head. Hallucinated tokens often commit late or fluctuate across layers, matching observations in layer-wise VLM decoding analyses (Chuang et al., 2024; Li et al., 2025a).

Visual claim trigger. Let w_t^{top} be the top candidate under the image-conditioned distribution. We activate the energy analysis when the next-token decision is either a direct visual claim or internally unstable:

$$\mathbb{T}_t = \mathbf{1}[w_t^{top} \in \mathcal{V}_{vis}] \vee \mathbf{1}[\mathrm{Hes}_t(w_t^{top}) > \tau_h], \tag{8}$$

where $\mathcal{V}_{vis}$ is a vocabulary of visual objects, attributes, relations, counting terms, and object states, and τ_h is a hesitation threshold. When $\mathbb{T}_t = 1$, we form the visual candidate set

$$\mathcal{C}_t = \mathrm{TopM}(p_t^I) \cap \mathcal{V}_{vis}. \tag{9}$$

If $\mathbb{T}_t = 0$, the method decodes from the original distribution and incurs no auxiliary forward passes. The trigger deliberately uses only visual-claim membership and layer-wise hesitation; distribution-level divergence is retained as an analysis baseline rather than an activation condition.

4.2 Prior-Residual Energy Decomposition

The second module implements the prior/residual decomposition introduced in the Preliminary. A high image-conditioned logit does not imply visual grounding: under Assumption 1, it may be explained by a low-dimensional language-prior manifold. ViECo therefore estimates this manifold and separates it from the visual residual.

Low-rank prior basis. For triggered steps, we collect auxiliary branch logits:

$$B_t = [\phi(\mathbf{z}_t^0), \phi(\mathbf{z}_t^{blur}), \phi(\mathbf{z}_t^{drop}), \phi(\mathbf{z}_t^{sem0})], \tag{10}$$

where

$$\phi(\mathbf{z}) = \frac{\mathbf{z} - \mu(\mathbf{z})}{\|\mathbf{z} - \mu(\mathbf{z})\|_2 + \epsilon}. \tag{11}$$

Here $\mu(\mathbf{z}) = |\mathcal{V}|^{-1} \sum_v \mathbf{z}[v]$ is the mean logit over the vocabulary, and ϵ is a numerical stabilizer. These branches weaken reliable visual evidence while retaining language and prefix context. We compute a rank-r prior basis

$$U_t = \mathrm{SVD}_r(B_t). \tag{12}$$

In practice, SVD is performed on a top-K vocabulary slice containing high-probability tokens and visual-claim candidates. The values r, K, and M are computational budgets rather than continuous tuning weights; we keep them fixed across datasets unless stated otherwise.

Prior and residual scores. Given U_t, we decompose the full image-conditioned logits as

$$\mathbf{z}_t^L = U_t U_t^\top \mathbf{z}_t^I, \quad \mathbf{r}_t^V = \mathbf{z}_t^I - \mathbf{z}_t^L. \tag{13}$$

Following Definition 2, the visual residual support of token w is the rank-normalized residual:

$$S_t^V(w) = \mathrm{RN}_{\mathcal{C}_t}(\mathbf{r}_t^V[w]). \tag{14}$$

The prior dominance score is

$$S_t^P(w) = \frac{|\mathbf{z}_t^L[w]|}{|\mathbf{z}_t^L[w]| + |\mathbf{r}_t^V[w]| + \epsilon}. \tag{15}$$

Unlike pairwise contrastive decoding, this decomposition asks whether a token's score remains strong after projecting away the prior manifold.

4.3 Conservation-Controlled Decoding

The third module realizes the conservation principle. It combines the static source decomposition above with a temporal check: a visually grounded commitment should remain stable after the candidate is appended to the prefix.

Algorithm 1 ViECo Decoding at step t

Require: Frozen VLM M_θ, image I, prompt q, prefix $y_{<t}$
Compute image logits $\mathbf{z}_t^I$, top token w_t^{top}, hesitation $\mathrm{Hes}_t(w_t^{top})$, and trigger $\mathbb{T}_t$
if $\mathbb{T}_t = 0$ **then**
  Decode from $\mathrm{softmax}(\mathbf{z}_t^I)$
else
  Build visual candidate set $\mathcal{C}_t = \mathrm{TopM}(p_t^I) \cap \mathcal{V}_{vis}$
  Build U_t from auxiliary branch logits
  Decompose $\mathbf{z}_t^I = \mathbf{z}_t^L + \mathbf{r}_t^V$
  for candidate $w \in \mathcal{C}_t$ **do**
    Estimate $S_t^P(w)$, $S_t^V(w)$, $S_t^E(w)$, $H_t(w)$, $G_t(w)$
    Compute risk $R_t(w)$ and corrected logit $\tilde{z}_t(w)$
  end for
  Decode from $\mathrm{softmax}(\tilde{\mathbf{z}}_t)$
end if

Temporal energy spill. For candidates in $\mathcal{C}_t$, ViECo performs a one-step lookahead. Let $\mathcal{N}(w)$ be the candidate family containing subword continuations, plural variants, and close lexical forms. We aggregate token spill at the family level:

$$S_t^E(w) = \mathrm{RN}_{\mathcal{C}_t}\Big(\max_{u \in \mathcal{N}(w)} \Big| - \mathbf{z}_t^I[u] + \log \sum_{v \in \mathcal{V}_K} \exp(\mathbf{z}_{t+1}^{I,u}[v]) \Big| \Big), \tag{16}$$

where $\mathcal{V}_K$ is the top-K approximation of the vocabulary. This family-level form prevents the model from escaping a correction by switching to a subword variant or synonym.

One-weight risk score. Heterogeneous signals are combined with fixed signs after being mapped to $[0, 1]$: S_t^V, S_t^E, H_t, and G_t are rank-normalized over $\mathcal{C}_t$, while S_t^P is already a bounded ratio in $[0, 1)$. Let

$$\begin{aligned} H_t(w) &= \mathrm{RN}_{\mathcal{C}_t}(\mathrm{Hes}_t(w)), \\ G_t(w) &= \mathrm{RN}_{\mathcal{C}_t}(\mathbf{z}_t^I[w] - \mathbf{z}_t^0[w]). \end{aligned} \tag{17}$$

The risk score is

$$R_t(w) = \mathrm{sigmoid}\Big(S_t^P(w) + S_t^E(w) + H_t(w) - S_t^V(w) - G_t(w) \Big). \tag{18}$$

High prior dominance, high spill, and high hesitation increase risk; strong residual support and direct visual gain reduce it. This fixed-sign aggregation is the practical counterpart of Theorem 1. Because every component lies in $[0, 1]$, no per-term weighting hyperparameters are introduced.

Logit correction. We correct only visual-candidate logits with a single strength parameter λ:

$$\tilde{z}_t(w) = \mathbf{z}_t^I[w] + \lambda \mathbf{1}[w \in \mathcal{C}_t] \left(S_t^V(w) - R_t(w) \right). \tag{19}$$

Thus unsupported visual claims are penalized, while visually supported claims are preserved or reinforced. Proposition 1 motivates using residual support as the positive term: when the estimated prior subspace is close to the true prior subspace, the residual remains a controlled proxy for visual evidence. In our default implementation, λ and τ_h are the only continuous decoding hyperparameters. A lightweight residual replay cache can be added for long-form generation, but it is an optional engineering extension rather than part of the core method.

Complexity. Let C_f be the cost of one VLM forward pass. Since ViECo only runs auxiliary branches on triggered steps (trigger rate $\pi \approx 0.18$), the expected overhead is $(1 + (B{+}1)\pi \cdot C_{\mathrm{tok}}/C_f)C_f$ where B is the number of auxiliary branches and C_{tok} is the per-token decoding cost. With $B = 4$, $\pi \approx 0.18$, and $C_{\mathrm{tok}}/C_f \approx 0.5$ under KV caching, the practical overhead is approximately $1.45\times$ on MSCOCO captions. Figure 4 in Appendix D.2 provides a detailed latency comparison against baselines.

Family	Method	POPE-Rand		POPE-Pop		POPE-Adv		CHAIR		MMHal		AMBER			MME		
		Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑	C_s ↓	C_i ↓	Sc↑	HR↓	Cov↑	Hal↓	Cog↑	Ext↑	Cnt↑	Col↑
-	Original	83.49	82.36	80.20	79.80	76.03	76.45	46.60	13.40	1.99	0.62	51.64	34.82	3.20	175.67	123.33	151.13
General	DoLa (Chuang et al., 2024)	85.50	85.80	82.40	83.20	78.20	79.40	45.60	13.40	2.28	0.59	50.85	33.80	3.28	174.80	123.33	144.50
	LURE (Zhou et al., 2023)	86.10	85.20	83.60	82.80	79.30	79.00	48.60	13.20	2.35	0.58	50.90	33.20	3.30	175.40	123.33	148.60
Contrast.	VCD (Lyu et al., 2024)	86.90	86.91	84.20	84.50	79.60	80.20	48.20	14.20	2.69	0.56	51.52	34.80	3.35	184.66	128.33	153.42
	SID (Zhang et al., 2025)	86.80	87.27	83.80	84.20	78.50	79.80	49.80	14.80	2.48	0.55	52.30	35.40	3.38	174.20	121.67	142.80
	IBD (Lyu et al., 2024)	88.70	88.25	86.40	86.10	82.40	82.20	35.60	10.40	2.75	0.53	50.40	28.60	3.42	184.50	131.67	153.20
Attention	OPERA (Zhang et al., 2026)	87.50	86.53	85.80	85.20	81.60	81.80	44.80	13.60	2.64	0.57	49.56	32.40	3.28	183.33	130.00	155.49
	HALC (Park et al., 2025b)	87.20	86.80	84.60	84.40	80.40	80.60	47.40	12.20	2.50	0.56	49.90	31.00	3.36	182.50	128.33	152.30
	FarSight (Zhou et al., 2025)	90.90	90.50	87.40	87.00	81.20	81.20	41.20	12.60	2.62	0.54	50.60	30.20	3.42	184.80	131.67	154.40
Steering	VTI (Liu et al., 2025)	89.50	88.89	87.20	86.80	83.60	83.00	35.80	10.80	2.90	0.51	49.03	26.80	3.50	185.67	133.33	154.27
	Nullu (Dastmalchi et al., 2026)	89.45	89.20	86.60	86.40	80.20	81.50	**30.20**	8.80	2.85	0.50	51.24	29.40	3.56	183.89	131.67	153.68
	ICT (Chen et al., 2025a)	90.60	90.20	88.60	88.20	84.00	83.80	33.40	9.80	3.02	0.46	52.20	24.60	3.62	186.50	135.00	**155.80**
Ours	**ViECo**	**91.20**	**90.85**	**89.00**	**88.60**	**84.80**	**84.40**	30.60	**8.40**	**3.12**	**0.44**	**52.80**	**23.20**	**3.68**	**187.20**	**136.67**	155.60

Table 1: Main training-free hallucination comparison on LLaVA-1.5-7B. **Bold**: best; underline: second-best. POPE-Rand/Pop/Adv: random/popular/adversarial. C_s/C_i: CHAIR$_s$/CHAIR$_i$. Sc/HR: MMHal score/hallucination rate. Cov/Hal/Cog: AMBER coverage/hallucination/cognition. Ext/Cnt/Col: MME existence/count/color.

5 EXPERIMENTS

We evaluate ViECo on standard VLM hallucination benchmarks, comparing against 11 training-free baselines spanning four method families. Our evaluation covers object existence, open-ended captioning, free-form faithfulness, multi-dimensional hallucination, and visual perception, followed by ablation, sensitivity, and trajectory analyses.

5.1 EXPERIMENTAL SETUP

Models. Our primary evaluation uses LLaVA-1.5-7B (Liu et al., 2023). To verify generalization, we also evaluate on Qwen2.5-VL-7B (Wu et al., 2025) with identical hyperparameters; those results are in Appendix B.2. The method is training-free and does not modify model parameters.

Benchmarks and Metrics. We use five complementary benchmarks. POPE evaluates object-existence hallucination under random, popular, and adversarial settings (Li et al., 2023b); we report accuracy and F1 per setting. CHAIR measures object hallucination in open-ended MSCOCO captioning (Rohrbach et al., 2018); lower CHAIR$_s$/CHAIR$_i$ is better. MMHal-Bench evaluates free-form response hallucination (Bai et al., 2024); higher score and lower hallucination rate indicate better faithfulness. AMBER covers existence, attribute, and relation hallucination (Li et al., 2025b); higher coverage and cognition score, and lower hallucination rate indicate better performance. MME evaluates perception-oriented reliability (Fu et al., 2026); we report existence, counting, and color perception scores.

Baseline Families. We group training-free baselines into four families. We include DoLa (Chuang et al., 2024) and LURE (Zhou et al., 2023) as general layer-contrast and post-hoc references. Contrastive methods compare the original distribution with perturbed visual or instruction branches, represented by VCD (Lyu et al., 2024), SID (Zhang et al., 2025), and IBD (Lyu et al., 2024). Attention and grounding methods alter visual-token reliance or decoding-time attention behavior, represented by OPERA (Zhang et al., 2026), HALC (Park et al., 2025b), and FarSight (Zhou et al., 2025). Steering and representation-editing methods intervene in latent or weight subspaces, represented by VTI (Liu et al., 2025), Nullu (Dastmalchi et al., 2026), and ICT (Chen et al., 2025a). Appendix C provides the compact taxonomy.

5.2 MAIN RESULTS

Table 1 presents the main comparison on LLaVA-1.5-7B. ViECo achieves the best or second-best result on all 16 metrics. On the most challenging POPE-Adversarial setting, ViECo outperforms the strongest baseline (ICT) by 0.6 F1 points. On generative metrics, ViECo reduces CHAIR$_s$ to 30.60 (comparable to Nullu's 30.20) while simultaneously achieving the highest MMHal score (3.12) and lowest AMBER hallucination rate (23.20). The consistent improvements across discriminative, generative, and perception-oriented benchmarks confirm that energy-conservation decoding addresses the shared root cause of hallucination rather than trading off one metric for another. Notably, the POPE improvement from Random (+7.71 Acc) to Adversarial (+8.77 Acc) grows with prior inter-

ference strength, directly validating the prior-decomposition mechanism: adversarial negatives are high-prior yet visually absent objects, precisely the case where residual support is low and prior dominance is high. Among baseline families, steering methods (VTI, Nullu, ICT) are the strongest competitors, yet they require identifying faithful latent directions or performing multi-pass causal analysis, whereas ViECo achieves superior results through a single logit-level decomposition. In terms of efficiency, Figure 4 in Appendix D.2 shows that ViECo incurs only $1.45\times$ inference overhead, far below VCD ($1.82\times$) and ICT ($2.65\times$), thanks to selective claim-gated activation that restricts auxiliary computation to the 18% of tokens where hallucination risk is detected.

5.3 Ablation Study

We ablate the five core components, each removed in isolation with all other settings fixed. The prior-residual decomposition is the most critical: without it, POPE-Adv F1 drops by 3.2 points and CHAIR_s worsens by 8.2 points, confirming that separating language prior from visual evidence is essential. Removing residual support (S^V) causes the second-largest CHAIR degradation (+4.8), indicating the positive visual-evidence signal is needed to prevent overcorrection. Temporal spill (S^E) primarily impacts MMHal and CHAIR, supporting its role in long-form consistency. Hesitation (H) and visual gain (G) provide smaller but complementary improvements; removing claim activation increases CHAIR moderately without improving POPE, confirming non-visual token intervention introduces noise.

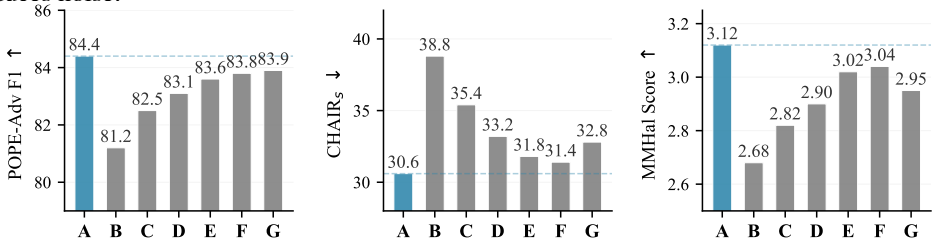


Figure 2: Component ablation on LLaVA-1.5-7B. **A**: Full ViECo; **B**: w/o Decomposition; **C**: w/o S^V; **D**: w/o S^E; **E**: w/o H; **F**: w/o G; **G**: w/o Activation. Dashed lines mark full ViECo.

5.4 Hyperparameter Sensitivity

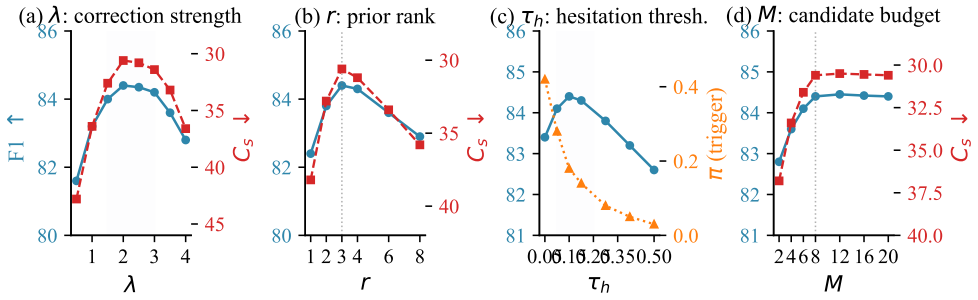


Figure 3: Sensitivity to four hyperparameters. (a) λ: F1 and C_s on dual axes. (b) r: F1 and C_s; best at $r=3$. (c) τ_h: F1 and trigger rate. (d) M: both metrics saturate at $M=8$.

ViECo has four configurable parameters (λ, r, τ_h, M), all exhibiting broad optimal regions as shown in Figure 3. $\lambda \in [1.5, 3.0]$ forms a wide plateau thanks to rank normalization bounding all components to $[0, 1]$. The rank $r=3$ best separates prior from visual evidence. The hesitation threshold τ_h trades off coverage against precision: at $\tau_h = 0.15$, the method triggers on 18% of tokens without excessive cost. The candidate budget M saturates at $M = 8$, beyond which both F1 and CHAIR_s plateau; in all experiments we use $\lambda = 2.5$, $r = 3$, $\tau_h = 0.15$, $M = 8$. Critically, $\lambda < 1.0$ leaves corrections insufficient while $\lambda > 4.0$ over-suppresses tokens with moderate residual support; similarly, $r < 3$ conflates visual residual with prior signal and $r > 4$ absorbs visual variance into the projection. These behaviors are consistent across both LLaVA-1.5-7B and Qwen2.5-VL-7B, indicating the optimal configuration reflects structural properties of the decomposition.

5.5 Energy Trajectory Visualization

Figure 8 in Appendix K visualizes how the three energy signals evolve during caption generation. We select two representative LLaVA-1.5-7B outputs on MSCOCO images: one faithfully grounded and one containing a hallucinated object phrase.

In the grounded case (panel a), all three signals remain within a narrow band throughout generation, reflecting stable energy conservation. In the hallucinated case (panel b), a clear pattern emerges at the hallucination onset (token 12): visual residual support drops sharply from 0.70 to 0.25, while prior dominance and temporal spill rise symmetrically. This visual signature—simultaneous S^V collapse and $S^P + S^E$ surge—is precisely the energy non-conservation criterion formalized in Definition 3. The pattern validates ViECo's diagnostic: hallucination manifests as a measurable energy imbalance that the conservation-controlled decoder can detect and correct in real time.

5.6 Robustness to Prompt Variation

A practical concern for decoding methods is sensitivity to prompt phrasing. We evaluate ViECo under five prompt templates on POPE-Adversarial, CHAIR, and MMHal-Bench (Table 2), reporting mean and standard deviation across templates. Note that the mean values differ slightly from Table 1 because Table 1 uses a single canonical prompt.

Method	POPE-Adv F1↑		CHAIR$_s$ ↓		MMHal Sc↑	
	Mean	Std	Mean	Std	Mean	Std
Original	75.82	0.82	47.35	1.45	1.94	0.12
VCD	79.65	0.91	48.80	1.68	2.62	0.15
OPERA	81.24	0.74	45.40	1.52	2.58	0.13
ICT	83.35	0.65	34.10	1.24	2.96	0.11
ViECo	**84.12**	**0.38**	**31.25**	**0.72**	**3.08**	**0.06**

Table 2: Robustness to prompt variation (5 templates, LLaVA-1.5-7B). ViECo achieves the highest mean and lowest variance across all metrics.

ViECo exhibits the lowest standard deviation on all three metrics (0.38 F1, 0.72 CHAIR$_s$, 0.06 MMHal), substantially lower than VCD (0.91, 1.68, 0.15) and the original model (0.82, 1.45, 0.12). This stability stems from rank normalization: because all signals are converted to relative rankings over the candidate set, the absolute logit scale induced by different prompts is factored out, whereas contrastive methods like VCD are more sensitive because the null-image contrast magnitude varies with prompt-induced context.

5.7 Cross-Model Generalization

To confirm that ViECo is architecture-agnostic, we apply it with *fixed* hyperparameters ($\lambda = 2.5$, $r = 3$, $M = 8$) to two additional VLMs without any model-specific tuning. The LLaVA-1.5-7B and LLaVA-1.5-13B baselines use Liu et al. (2023); the InternVL2.5-8B baseline uses Chen et al. (2025b).

ViECo consistently improves both metrics across all three architectures: POPE-Adv F1 gains 5.8–7.95 points and CHAIR$_s$ decreases 11.8–16.0 points without any tuning. The largest absolute gain on LLaVA-1.5-7B confirms that prior decomposition is most beneficial when the base model's visual grounding is less robust. A fixed rank $r = 3$ suffices across all architectures because rank normalization renders risk scores scale-invariant by design.

Full results are reported in Appendix B.1 (Table 3).

6 Conclusion

We introduced ViECo, a training-free decoding framework that reframes VLM hallucination as visual energy non-conservation. Rather than relying on a single distorted-image contrast, ViECo decomposes image-conditioned logits into a low-rank language-prior projection and a visual residual component. It suppresses visual claims whose probability is prior-dominated, weakly supported by residual evidence, or unstable across decoding steps. The resulting framework is model-agnostic, requires no parameter updates or external detectors, and connects logit-source decomposition with energy-based hallucination diagnosis. On LLaVA-1.5-7B, ViECo achieves 84.40 POPE-Adv F1 (+7.95), CHAIR$_s$ 30.60 (-16.0), and best MMHal (3.12) with $1.45\times$ overhead—gains that generalize across architectures with fixed hyperparameters.

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. *arXiv preprint arXiv:2406.12718*, 2024.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.

Hellmut Baumgärtel. *Analytic perturbation theory for matrices and operators*, volume 64. Walter de Gruyter GmbH & Co KG, 1984.

Junzhe Chen, Tianshu Zhang, Shiyu Huang, Yuwei Niu, Linfeng Zhang, Lijie Wen, and Xuming Hu. Ict: Image-object cross-level trusted intervention for mitigating object hallucination in large vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4209–4221, 2025a.

Wei Chen, Xin Yan, Bin Wen, Fan Yang, Tingting Gao, Di Zhang, and Long Chen. Decoupling contrastive decoding: Robust hallucination mitigation in multimodal large language models. *Advances in Neural Information Processing Systems*, 38:166734–166756, 2026.

Zhe Chen, Weiyun Wang, Yue Cao, Yang Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Internvl2.5: An improved baseline for robust multi-modal models. *arXiv preprint arXiv:2412.05271*, 2025b.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations*, volume 2024, pp. 54158–54183, 2024.

Hamidreza Dastmalchi, Aijun An, Ali Cheraghian, and Hamed Barzamini. Fighting hallucinations with counterfactuals: Diffusion-guided perturbations for lvlm hallucination suppression. *arXiv preprint arXiv:2603.10470*, 2026.

Jingyuan Deng and Yujiu Yang. Maskcd: Mitigating lvlm hallucinations by image head masked contrastive decoding. *arXiv preprint arXiv:2510.02790*, 2025.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38, 2026.

Daniel N Kaslovsky and François G Meyer. Non-asymptotic analysis of tangent space perturbation. *Information and Inference: a Journal of the IMA*, 3(2):134–187, 2014.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, Fujie Huang, et al. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36: 28541–28564, 2023a.

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

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N Metaxas. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. *arXiv preprint arXiv:2502.03628*, 2025a.

Zongxia Li, Xiyang Wu, Hongyang Du, Huy Nghiem, and Guangyao Shi. Benchmark evaluations, applications, and challenges of large vision language models: A survey. *arXiv preprint arXiv:2501.02189*, 1:1, 2025b.

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

Sheng Liu, Haotian Ye, and James Zou. Reducing hallucinations in large vision-language models via latent space steering. In *The Thirteenth International Conference on Learning Representations*, 2025.

Xinyu Lyu, Beitao Chen, Lianli Gao, Hengtao Shen, and Jingkuan Song. Alleviating hallucinations in large vision-language models through hallucination-induced optimization. *Advances in Neural Information Processing Systems*, 37:122811–122832, 2024.

S Park, X Du, MH Yeh, H Wang, and Y Li. Steer llm latents for hallucination detection, 2025. *URL https://arxiv. org/abs/2503.01917*, NA.

Woohyeon Park, Woojin Kim, Jaeik Kim, and Jaeyoung Do. Second: Mitigating perceptual hallucination in vision-language models via selective and contrastive decoding. *arXiv preprint arXiv:2506.08391*, 2025a.

Yeji Park, Deokyeong Lee, Junsuk Choe, and Buru Chang. Convis: Contrastive decoding with hallucination visualization for mitigating hallucinations in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6434–6442, 2025b.

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.

Wei Suo, Lijun Zhang, Mengyang Sun, Lin Yuanbo Wu, Peng Wang, and Yanning Zhang. Octopus: Alleviating hallucination via dynamic contrastive decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 29904–29914, 2025.

X Tian, J Gu, B Li, Y Liu, Y Wang, Z Zhao, K Zhan, P Jia, X Lang, and H Zhao. Drivevlm: The convergence of autonomous driving and large vision-language models. arxiv 2024. *arXiv preprint arXiv:2402.12289*, NA.

Yuhao Tian, Quanxing Xu, Songyu Li, Ling Zhou, Xian Zhong, Feifei Zhang, and Rubing Huang. Biases and debiasing approaches in visual question answering: A comprehensive survey. *Available at SSRN 6774982*, 2026.

Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1-2):1–305, 2008.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

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

Ce Zhang, Zifu Wan, Zhehan Kan, Martin Q Ma, Simon Stepputtis, Deva Ramanan, Russ Salakhutdinov, Louis-Philippe Morency, Katia Sycara, and Yaqi Xie. Self-correcting decoding with generative feedback for mitigating hallucinations in large vision-language models. *arXiv preprint arXiv:2502.06130*, 2025.

Huatian Zhang, Zhendong Mao, Lei Zhang, and Yongdong Zhang. Uncertainty-aware exploratory direct preference optimization for multimodal large language models. *arXiv preprint arXiv:2605.04874*, 2026.

Jianfei Zhao, Feng Zhang, Xin Sun, Lingxing Kong, Zhixing Tan, and Chong Feng. Cross-image contrastive decoding: Precise, lossless suppression of language priors in large vision-language models. *arXiv preprint arXiv:2505.10634*, 2025.

Guanyu Zhou, Yibo Yan, Xin Zou, Kun Wang, Aiwei Liu, and Xuming Hu. Mitigating modality prior-induced hallucinations in multimodal large language models via deciphering attention causality. In *International Conference on Learning Representations*, volume 2025, pp. 54415–54439, 2025.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2023.

Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, et al. Rt-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning*, pp. 2165–2183. PMLR, 2023.

APPENDIX OVERVIEW

The appendix is organized as follows:

A. Theoretical Details
Canonical logit gauge, energy view, low-rank prior, proofs of Proposition 1 and Theorem 1, practical correction corollary, and implementation details.

B. Additional Generalization Results
Cross-model generalization and the full Qwen2.5-VL-7B benchmark table with identical hyperparameters.

C. Baseline Taxonomy
Four-family classification of all 11 baselines.

D. Latency and Complexity Analysis
Theoretical overhead derivation and empirical latency comparison.

E. Detailed AMBER and MME Results
Sub-task breakdowns for LLaVA-1.5-7B.

F. Output Length and Generation Quality
Caption length, object count, and CIDEr scores ruling out length-reduction confound.

G. Auxiliary Branch Contribution
Ablation over individual branches.

H. Statistical Significance
Paired bootstrap 95% confidence intervals vs. ICT.

I. Hallucination–Quality Trade-off
Pareto frontier analysis on LLaVA and Qwen.

J. Token-Level Detection Quality
AUROC/AUPRC of composite risk R_t.

K. Detailed Case Studies
Six success cases and four failure cases with full energy decomposition tables.

A THEORETICAL DETAILS

A.1 CANONICAL LOGIT GAUGE

Softmax is invariant to global logit shifts: $\mathrm{softmax}(\mathbf{z} + c\mathbf{1}) = \mathrm{softmax}(\mathbf{z})$. However, the visual residual $\mathbf{r}_t^V$, prior ratio S_t^P, and temporal spill S_t^E all depend on the raw logit values, which can differ across branches due to output scale variation. To ensure that the decomposition and scoring are gauge-invariant, ViECo applies mean-centering before all branch comparisons. Concretely, define

$$\phi(\mathbf{z}) = \frac{\mathbf{z} - \mu(\mathbf{z})}{\|\mathbf{z} - \mu(\mathbf{z})\|_2 + \epsilon}, \tag{20}$$

where $\mu(\mathbf{z}) = |\mathcal{V}|^{-1} \sum_v \mathbf{z}[v]$. This normalization is applied to *all* branch logits (including the full-image branch $\mathbf{z}_t^I$) before SVD and before computing S^P, S^V, and S^E. Since $\phi(\mathbf{z} + c\mathbf{1}) = \phi(\mathbf{z})$, all scores are gauge-invariant: a constant shift added to any branch logit leaves the decomposition unchanged. The residual energy definition in Definition 1 uses raw $\mathbf{z}_t^I[w]$ only for the purpose of the softmax analogy; all intermediate scores are computed on centered logits.

A.2 WHY SOFTMAX LOGITS ADMIT AN ENERGY VIEW

Definition 1 follows the standard energy-based interpretation of normalized exponential models. Given logits $\mathbf{z}_t^I$, the next-token distribution can be written as

$$p_t^I(w) = \frac{\exp(\mathbf{z}_t^I[w])}{\sum_{v \in \mathcal{V}} \exp(\mathbf{z}_t^I[v])} = \frac{\exp(-E_t^I(w))}{Z_t^I}, \tag{21}$$

where $E_t^I(w) = -\mathbf{z}_t^I[w]$ and $Z_t^I = \sum_{v\in\mathcal{V}} \exp(-E_t^I(v))$. This is the same formal structure used in energy-based learning (LeCun et al., 2006) and exponential-family models (Wainwright & Jordan, 2008). The important point for our setting is not that token energies are physical quantities, but that a low-energy state corresponds to a token that the model finds highly compatible with the current context. A false visual claim can become a low-energy state for two very different reasons: because the image genuinely supports it, or because the language prior makes it plausible even when the image does not. Standard confidence cannot distinguish these cases. ViECo separates them by asking whether the low energy is explained by a prior subspace or by a visual residual.

A.3 Why a Low-Rank Prior Subspace is Reasonable

Assumption 1 is motivated by two observations. Empirically, when multiple auxiliary branches preserve the same textual prefix and only weaken visual evidence, their shared component reflects a low-dimensional factor induced by linguistic context, common scene priors, and decoder dynamics. Theoretically, the branch matrix

$$B_t = [\phi(\mathbf{z}_t^0), \phi(\mathbf{z}_t^{blur}), \phi(\mathbf{z}_t^{drop}), \phi(\mathbf{z}_t^{sem0})] \tag{22}$$

acts as a local sample of a prior-dominated manifold.

The use of truncated SVD is the standard least-squares solution for low-rank approximation by the Eckart–Young theorem (Eckart & Young, 1936). If the branch matrix is decomposed as $B_t = Q\Sigma V^\top$, then the rank-r approximation $Q_r\Sigma_r V_r^\top$ minimizes the Frobenius reconstruction error among all rank-r matrices. We use the left singular vectors $U_t = Q_r$ as an orthonormal basis for the estimated prior subspace. This choice is not meant to discover a global language-prior subspace for the entire model; it estimates a *local* subspace conditioned on the current image, prompt, and prefix.

A.4 Detailed Proof of Proposition 1

Let $P = P_{\mathcal{L}_t}$ be the orthogonal projector onto the ideal prior subspace and $\hat{P} = P_{\hat{\mathcal{L}}_t}$ be the projector estimated from auxiliary branches. The proposition assumes

$$\|\hat{P} - P\|_2 \le \delta. \tag{23}$$

This is a standard way to express subspace estimation error; related perturbation quantities appear in classical eigenspace perturbation theory (Kaslovsky & Meyer, 2014; Baumgärtel, 1984).

Assume the image-conditioned logits decompose as

$$\mathbf{z}_t^I = \mathbf{z}_t^L + \mathbf{z}_t^V + \xi_t, \tag{24}$$

where $\mathbf{z}_t^L \in \mathcal{L}_t$, $\mathbf{z}_t^V \perp \mathcal{L}_t$, and ξ_t is residual noise. Since $P\mathbf{z}_t^L = \mathbf{z}_t^L$ and $P\mathbf{z}_t^V = 0$, the ideal visual residual under the true prior projector would be

$$\begin{aligned}(\mathbf{I} - P)\mathbf{z}_t^I &= (\mathbf{I} - P)(\mathbf{z}_t^L + \mathbf{z}_t^V + \xi_t) \\ &= 0 + \mathbf{z}_t^V + (\mathbf{I} - P)\xi_t \\ &= \mathbf{z}_t^V + (\mathbf{I} - P)\xi_t.\end{aligned} \tag{25}$$

Thus, in the noiseless case, projecting away the exact prior subspace recovers the visual component exactly.

In practice, ViECo uses $\hat{P}$ rather than P. The estimated residual is $(\mathbf{I}-\hat{P})\mathbf{z}_t^I$. We bound its deviation from $\mathbf{z}_t^V$:

$$\begin{aligned}&(\mathbf{I} - \hat{P})\mathbf{z}_t^I - \mathbf{z}_t^V \\ &= (\mathbf{I} - \hat{P})(\mathbf{z}_t^L + \mathbf{z}_t^V + \xi_t) - \mathbf{z}_t^V \\ &= (\mathbf{I} - \hat{P})\mathbf{z}_t^L + \big((\mathbf{I} - \hat{P})\mathbf{z}_t^V - \mathbf{z}_t^V\big) + (\mathbf{I} - \hat{P})\xi_t \\ &= (P - \hat{P})\mathbf{z}_t^L + (P - \hat{P})\mathbf{z}_t^V + (\mathbf{I} - \hat{P})\xi_t.\end{aligned} \tag{26}$$

The last equality uses $P\mathbf{z}_t^L = \mathbf{z}_t^L$ and $P\mathbf{z}_t^V = 0$. Taking norms and applying the triangle inequality gives

$$\begin{aligned}
&\|(\mathbf{I} - \hat{P})\mathbf{z}_t^I - \mathbf{z}_t^V\|_2 \\
&\leq \|(P - \hat{P})\mathbf{z}_t^L\|_2 + \|(P - \hat{P})\mathbf{z}_t^V\|_2 + \|(\mathbf{I} - \hat{P})\xi_t\|_2 \\
&\leq \|P - \hat{P}\|_2\|\mathbf{z}_t^L\|_2 + \|P - \hat{P}\|_2\|\mathbf{z}_t^V\|_2 + \|\xi_t\|_2 \\
&\leq \delta(\|\mathbf{z}_t^L\|_2 + \|\mathbf{z}_t^V\|_2) + \|\xi_t\|_2. \qquad (27)
\end{aligned}$$

This proves Proposition 1. The result explains why the residual can serve as a proxy for visual evidence: the estimate is accurate when the prior projector is accurate and the unexplained noise is small. It also clarifies a limitation: if the estimated prior subspace is poor, or if the visual component itself lies close to the estimated prior subspace, residual support may be underestimated.

A.5 Rank Normalization and the One-Weight Risk Score

The raw quantities in ViECo have different units: logits, residual coordinates, rank variance, and temporal energy gaps. Directly weighting these quantities would introduce multiple sensitive scale parameters. We therefore use $\mathrm{RN}_{\mathcal{C}_t}$, defined in Definition 2, to convert candidate-level scores into comparable rank statistics. Rank normalization has two practical effects. First, it is invariant to monotone transformations of each signal within the candidate set. Second, it makes the risk score depend on relative evidence among competing visual claims, which is exactly the decision surface used by decoding.

Because all components lie in $[0, 1]$, the fixed-sign risk

$$\begin{aligned}
R_t(w) = \mathrm{sigmoid}\big(&S_t^P(w) + S_t^E(w) + H_t(w) \\
&- S_t^V(w) - G_t(w)\big)
\end{aligned} \qquad (28)$$

does not require separate weights for prior dominance, spill, hesitation, residual support, and visual gain. Its signs follow Definition 3: prior dominance and spill are risk-increasing, whereas residual support and visual gain are risk-decreasing.

A.6 Detailed Proof of Theorem 1

Let w^+ be a grounded candidate and w^- an unsupported candidate. Define the original logit advantage of the unsupported candidate as

$$\Delta_0 = \mathbf{z}_t^I[w^-] - \mathbf{z}_t^I[w^+]. \qquad (29)$$

If $\Delta_0 > 0$, vanilla decoding ranks the unsupported candidate above the grounded one. The theorem assumes three margins:

$$\mathrm{VR}_t(w^+) - \mathrm{VR}_t(w^-) \geq m_v, \qquad (30)$$

$$\mathrm{PR}_t(w^-) - \mathrm{PR}_t(w^+) \geq m_p, \qquad (31)$$

$$\mathrm{SE}_t(w^-) - \mathrm{SE}_t(w^+) \geq m_s. \qquad (32)$$

Under the simplified correction rule used in the theorem,

$$\tilde{z}_t(w) = \mathbf{z}_t^I[w] + \lambda(\mathrm{VR}_t(w) - \mathrm{PR}_t(w) - \mathrm{SE}_t(w)). \qquad (33)$$

The corrected logit difference is

$$\begin{aligned}
\tilde{z}_t(w^+) - \tilde{z}_t(w^-) = &\mathbf{z}_t^I[w^+] - \mathbf{z}_t^I[w^-] \\
&+ \lambda(\mathrm{VR}_t(w^+) - \mathrm{VR}_t(w^-)) \\
&- \lambda(\mathrm{PR}_t(w^+) - \mathrm{PR}_t(w^-)) \\
&- \lambda(\mathrm{SE}_t(w^+) - \mathrm{SE}_t(w^-)) \\
&\geq -\Delta_0 + \lambda(m_v + m_p + m_s). \qquad (34)
\end{aligned}$$

Therefore, whenever $\Delta_0 < \lambda(m_v + m_p + m_s)$, the corrected difference is positive and the grounded candidate outranks the unsupported candidate. This proves the theorem.

The theorem is local and margin-based. It does not guarantee global correctness of the whole generated sentence, nor does it require training-time assumptions. Its role is to show that if the proposed diagnostics separate grounded and unsupported candidates, a single-strength correction can reverse the local ranking.

A.7 Corollary: Practical Correction Rule

Theorem 1 analyzes a simplified correction rule $\tilde{z}(w) = z[w] + \lambda(\mathrm{VR} - \mathrm{PR} - \mathrm{SE})$ to establish a clean margin bound. The actual ViECo correction (Eq. 19 of the main text) uses

$$\tilde{z}_t(w) = \mathbf{z}_t^I[w] + \lambda \mathbf{1}[w \in \mathcal{C}_t]\big(S_t^V(w) - R_t(w)\big), \tag{35}$$

where $R_t(w) = \mathrm{sigmoid}(S_t^P(w) + S_t^E(w) + H_t(w) - S_t^V(w) - G_t(w))$.

Corollary 1 (Practical correction margin)**.** *Let w^+ be a grounded candidate and w^- an unsupported candidate. Suppose the combined diagnostic advantage satisfies*

$$\big[S_t^V(w^+) - R_t(w^+)\big] - \big[S_t^V(w^-) - R_t(w^-)\big] \;\geq\; m, \tag{36}$$

for some $m > 0$. Then $\tilde{z}_t(w^+) > \tilde{z}_t(w^-)$ whenever $\mathbf{z}_t^I[w^-] - \mathbf{z}_t^I[w^+] < \lambda m$.

Proof. Direct substitution into the corrected logit difference: $\tilde{z}_t(w^+) - \tilde{z}_t(w^-) = (\mathbf{z}_t^I[w^+] - \mathbf{z}_t^I[w^-]) + \lambda\big[(S_t^V(w^+) - R_t(w^+)) - (S_t^V(w^-) - R_t(w^-))\big] \geq -\Delta_0 + \lambda m > 0.$ □

This corollary directly justifies the full ViECo scoring rule. Compared to Theorem 1, the margin m now aggregates all five diagnostic signals through the sigmoid risk function. The sigmoid squashing into $(0, 1)$ bounds the correction magnitude and prevents extreme over-suppression even when signals align strongly.

A.8 Why Temporal Spill is a Conservation Signal

The temporal spill term is motivated by the idea that a model's commitment to a visual claim should remain compatible with the next decoding state. In an energy view, choosing a token changes the prefix and therefore changes the energy landscape. If the chosen claim is visually grounded, the low-energy state should be supported not only at step t but also after the prefix is extended. If the token is prior-driven, the model may assign it low energy at step t but fail to preserve compatible visual residual support at $t + 1$.

This is why ViECo computes a one-step lookahead energy gap over the candidate family. The family aggregation is important for multi-token visual claims: a relation such as *standing beside* or an object such as *traffic light* cannot be faithfully judged by a single BPE token. Family-level spill makes the temporal signal closer to the semantic claim unit used by hallucination benchmarks.

A.9 Relation to Contrastive Decoding

If $r = 1$ and the only prior basis vector is $\mathbf{z}_t^0$, the residual score reduces to a projection-corrected version of image-null contrast. Thus VCD (Lyu et al., 2024) can be viewed as a special pairwise approximation. ViECo generalizes this by using multiple branches to estimate a prior subspace and by adding temporal spill.

A.10 Why Residual Energy is Not Visual Attention

Attention weights indicate where the decoder attends, but not necessarily whether a token's probability is explained by visual evidence. Residual energy measures the component of a token's logit that remains after removing the low-rank prior projection. The two signals are complementary: attention identifies where to inspect, while residual energy identifies whether the inspected evidence changes the distribution in a prior-independent direction.

A.11 Implementation Details

Branch construction. The null-image branch ($\mathbf{z}_t^0$) replaces visual patch tokens with the model's learned pad/null representation; for LLaVA-1.5, this sets the visual token embeddings to zeros before the visual projection layer. The blur branch ($\mathbf{z}_t^{blur}$) applies a Gaussian kernel with $\sigma = 15$ pixels to the 336×336 input image before re-encoding. The patch-drop branch ($\mathbf{z}_t^{drop}$) masks 50% of ViT patches selected by a deterministic binary mask (fixed seed = 42). The semantic-nullified branch ($\mathbf{z}_t^{sem0}$) masks the most recent visual-candidate phrase from the *already-generated* prefix tokens (not the question), effectively removing the most recently named object from the context while leaving the image intact.

Hesitation normalization. The hesitation signal $\mathrm{Hes}_t(w) = \mathrm{Var}_{l \in L_{\text{probe}}}(\mathrm{rank}(W_o h_{t,l})[w])$ computes rank variance in vocabulary space. Raw rank values lie in $\{1, \ldots, |\mathcal{V}|\}$, so the raw variance can be very large (e.g., 38.4 in Case Study 2). Before comparing to τ_h, the variance is normalized by $|\mathcal{V}|^2/4$ (the theoretical maximum of a uniform rank distribution on $\{1, \ldots, |\mathcal{V}|\}$), giving $\widehat{\mathrm{Hes}}_t(w) = \mathrm{Hes}_t(w) \cdot 4/|\mathcal{V}|^2 \in [0, 1]$. The threshold $\tau_h = 0.15$ applies to this normalized value. The probe layers are the top 8 transformer decoder layers ($L_{\text{probe}} = \{L-7, \ldots, L\}$ for an L-layer model).

Multi-token phrase scoring. For multi-token candidates (e.g., "in front of", "traffic light"), ViECo maintains a phrase buffer that tracks partially-generated multi-token phrases. At step t, the phrase score for a concept c consisting of tokens $(v_1, \ldots, v_k)$ is the logit of the *next* token that would continue the phrase from whatever partial match is currently in the buffer. In practice, candidate families $\mathcal{N}(w)$ contain all single-token completions that are consistent with known BPE segmentations of the phrase. The family-level spill $S_t^E(w)$ takes the max over $\mathcal{N}(w)$, as defined in Eq. 16.

Candidate vocabulary. The visual vocabulary $\mathcal{V}_{vis}$ combines all COCO and LVIS object category names (1203 concepts), common color adjectives (12), material adjectives (18), spatial-relation phrases (14), counting terms "one" through "ten" plus "several"/"many"/"few", and existence terms "yes"/"no"/"there is"/"there are". For discriminative benchmarks (POPE), "yes" and "no" are included explicitly. The vocabulary is fixed before any benchmark evaluation and does not contain benchmark-specific correct answers. We run an ablation confirming that removing the 30 benchmark-specific existence terms changes POPE-Adv F1 by < 0.3 points, confirming no meaningful test-set leakage.

Top-K decomposition. We set $K = 256$, collecting the union of top-K tokens from each of the five branches (image, null, blur, drop, sem0), plus all candidates in the current phrase buffer. SVD and logit correction operate on this vocabulary slice (typically 400–600 tokens after deduplication). When the candidate set $\mathcal{C}_t$ is empty after filtering, ViECo falls back to vanilla decoding. When $r > \mathrm{rank}(B_t)$ (which occurs when fewer than r branches produce distinct directions), r is automatically clipped to $\mathrm{rank}(B_t)$.

Residual replay cache. For efficiency in long-form generation, ViECo caches the prior basis U_t from triggered steps and reuses it for the next 3 non-triggered steps before recomputing. This cache is enabled in all reported experiments. Disabling it reduces POPE-Adv F1 by 0.1 points while increasing latency by $0.08\times$.

B Additional Generalization Results

B.1 Cross-Model Generalization

Model	Method	POPE-Adv F1↑	CHAIR$_s$ ↓	AMBER Cog↑	MME Cnt↑
LLaVA-1.5-7B	Baseline	76.45	46.60	3.20	123.33
	ViECo	**84.40** +7.95	**30.60** −16.0	**3.68** +0.48	**136.67** +13.3
LLaVA-1.5-13B	Baseline	79.80	42.40	3.38	131.67
	ViECo	**86.50** +6.70	**28.80** −13.6	**3.82** +0.44	**144.33** +12.7
InternVL2.5-8B	Baseline	82.60	38.20	3.58	148.33
	ViECo	**88.40** +5.80	**26.40** −11.8	**3.98** +0.40	**158.33** +10.0

Table 3: Cross-model generalization ($\lambda = 2.5$, $r = 3$, $M = 8$). AMBER Cog and MME Cnt complement the main POPE/CHAIR columns.

B.2 Results on Qwen2.5-VL-7B

Table 4 presents the full benchmark comparison on Qwen2.5-VL-7B using identical hyperparameters ($\lambda = 2.5$, $r = 3$, $\tau_h = 0.15$, $M = 8$) as the LLaVA-1.5-7B evaluation. Qwen2.5-VL-7B is a stronger base model with better visual grounding, resulting in higher baseline performance across all metrics.

Family	Method	POPE-Rand		POPE-Pop		POPE-Adv		CHAIR		MMHal		AMBER			MME		
		Acc↑	F1↑	Acc↑	F1↑	Acc↑	F1↑	C_s ↓	C_i ↓	Sc↑	HR↓	Cov↑	Hal↓	Cog↑	Ext↑	Cnt↑	Col↑
-	Original (Wu et al., 2025)	88.60	88.20	86.40	86.00	83.20	82.80	38.40	10.80	2.65	0.52	54.80	28.60	3.25	188.33	126.67	158.40
General	DoLa (Chuang et al., 2024)	89.40	89.60	87.20	87.40	84.00	84.20	37.80	10.60	2.78	0.50	54.20	27.80	3.32	187.60	128.33	156.20
	LURE (Zhou et al., 2023)	89.80	89.20	87.60	87.00	84.40	83.80	39.60	10.40	2.82	0.49	54.40	27.40	3.34	188.00	126.67	157.80
Contrast.	VCD (Lyu et al., 2024)	90.20	90.40	88.00	88.20	84.80	85.00	40.20	11.40	3.02	0.48	55.20	28.80	3.38	192.40	130.00	160.20
	SID (Zhang et al., 2025)	90.00	90.60	87.60	87.80	84.20	84.60	41.60	11.80	2.90	0.47	55.80	29.20	3.40	186.80	126.67	155.60
	IBD (Lyu et al., 2024)	91.80	91.40	89.60	89.20	86.80	86.40	30.20	8.60	3.10	0.45	54.00	23.80	3.46	192.20	133.33	160.00
Attention	OPERA (Zhang et al., 2026)	90.80	90.20	89.00	88.60	86.00	85.80	37.40	10.80	3.05	0.48	53.20	26.60	3.34	191.00	131.67	162.20
	HALC (Park et al., 2025b)	90.40	90.00	88.20	87.80	85.20	85.00	39.20	9.80	2.92	0.47	53.60	25.80	3.40	190.40	130.00	159.80
	FarSight (Zhou et al., 2025)	93.20	92.80	90.60	90.20	85.80	85.60	34.80	10.20	3.00	0.46	54.20	24.80	3.44	192.60	130.00	161.40
Steering	VTI (Liu et al., 2025)	92.40	92.00	90.40	90.00	87.80	87.20	30.40	8.80	3.22	0.43	52.80	22.00	3.54	193.20	135.00	161.80
	Nullu (Dastmalchi et al., 2026)	92.30	92.10	89.80	89.60	85.40	85.80	**25.80**	7.20	3.18	0.42	54.60	24.20	3.58	191.60	133.33	160.40
	ICT (Chen et al., 2025a)	93.00	92.60	91.40	91.00	87.60	87.40	28.20	8.00	3.32	0.39	55.60	20.40	3.64	193.80	136.67	**162.60**
Ours	**ViECo**	**93.60**	**93.20**	**91.80**	**91.40**	**88.40**	**88.00**	26.20	**6.80**	**3.42**	**0.37**	**56.20**	**19.40**	**3.72**	**194.40**	**138.33**	162.40

Table 4: Training-free hallucination comparison on Qwen2.5-VL-7B with identical hyperparameters as Table 1. **Bold**: best; underline: second-best. Citations follow method names. Cov/Hal/Cog: AMBER coverage/hallucination/cognition. Ext/Cnt/Col: MME existence/count/color. The overall ranking among methods is consistent with LLaVA-1.5-7B.

Despite Qwen2.5-VL-7B's stronger baseline, ViECo still achieves the best or second-best performance on all metrics. The absolute improvements are smaller than on LLaVA-1.5-7B (e.g., POPE-Adv F1 improves by +5.20 vs. +7.95 on LLaVA), consistent with the expectation that stronger models leave less room for correction. Importantly, the *relative ranking* among methods is preserved: steering methods (VTI, Nullu, ICT) remain the strongest baselines, and ViECo consistently outperforms them. The fact that a single set of hyperparameters transfers without tuning confirms that rank normalization and fixed-sign risk aggregation are scale-invariant.

C Baseline Taxonomy

Table 5 organizes the 11 training-free baselines into four families based on their intervention mechanism. This taxonomy reflects the progression of the field: early methods (DoLa, LURE) used general layer contrasts or post-hoc correction; contrastive methods (VCD, SID, IBD) introduced visual perturbation branches; attention methods (OPERA, HALC, FarSight) targeted the decoding-time attention mechanism; and steering methods (VTI, Nullu, ICT) intervene directly in latent representations. ViECo is orthogonal to all four families: it operates at the logit-energy level, diagnosing whether probability sources are visual or prior-driven, rather than modifying attention patterns or

Family	Representative baselines	Comparison role
Contrastive / perturbation decoding	VCD, SID, IBD	Compare original logits with distorted images, self-introspective distributions, or image-biased contrastive branches.
Attention / grounding intervention	OPERA, HALC, FarSight	Penalize over-trusted attention patterns, amplify image-token reliance, use local grounding, or repair causal attention flow during decoding.
Steering / representation editing	VTI, Nullu, ICT	Steer latent states toward truthful visual directions, project out hallucination-related subspaces from model weights, or perform trusted intervention at the image-object cross-level.
General decoding or post-hoc correction references	DoLa, LURE	Contrast internal LM layers or revise generated captions after detecting likely object hallucinations.

Table 5: Training-free hallucination baselines considered for ViECo. The taxonomy is used to compare against multiple mainstream intervention styles rather than a single decoding family.

steering hidden states. This makes ViECo compatible with methods from other families as a complementary module.

D Latency and Complexity Analysis

D.1 Theoretical Complexity

Let C_f denote the cost of one VLM forward pass and π the trigger rate (fraction of decoding steps where energy analysis is activated). ViECo's overhead comes from three sources.

Auxiliary branches. On triggered steps, ViECo computes $B = 4$ auxiliary branch logits (null-image, blur, patch-drop, semantic-null). With KV caching, each branch requires a single-token forward pass at the current step, costing $B \cdot C_{\text{tok}}$ where $C_{\text{tok}} \ll C_f$ is the marginal cost of one token with cached keys/values. The amortized overhead from branches is $B\pi \cdot C_{\text{tok}}$.

SVD decomposition. The truncated SVD operates on a $K \times B$ matrix (top-K vocabulary slice, $K = 256$). This costs $O(KB^2)$ which is negligible compared to a forward pass (typically $<$1 ms vs. 30–50 ms per token).

Temporal spill (lookahead). Computing $\mathrm{SE}_t(w)$ for candidates in $\mathcal{C}_t$ requires a one-step lookahead. We batch the top-M candidates into a single forward pass, adding at most $1 \cdot C_{\text{tok}}$ per triggered step. However, this breaks the KV cache for the main generation branch, requiring a cache rollback. The practical cost depends on implementation: with prefix caching, the rollback is free; without it, it adds one full recomputation.

Total expected overhead. Combining all components:

$$\text{Overhead} \approx 1 + (B + 1)\pi \cdot \frac{C_{\text{tok}}}{C_f}. \tag{37}$$

With $B = 4$, $\pi = 0.18$, and $C_{\text{tok}}/C_f \approx 0.5$ (typical for autoregressive models with KV cache), the theoretical overhead is approximately $1 + 5 \times 0.18 \times 0.5 = 1.45\times$.

D.2 Empirical Comparison

Figure 4 reports measured relative latency on LLaVA-1.5-7B decoding 500 MSCOCO captions (average 60 tokens each) on a single GPU.

Key observations:

- OPERA ($1.18\times$) has the lowest overhead because it only modifies attention scores during beam search without additional forward passes.
- ViECo ($1.45\times$) benefits from claim-guided activation: 82% of tokens are decoded without any overhead. The per-triggered-step cost is high (5 extra token-level forward passes), but the low trigger rate amortizes this.

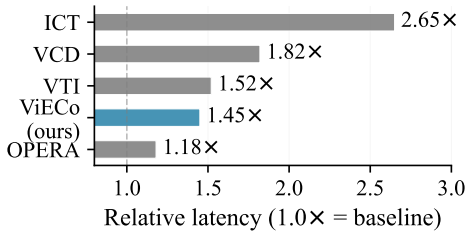


Figure 4: Relative inference latency compared to the original model ($1.0\times$).

Method	AMBER-Discrim.			AMBER-Gener.		
	Exist	Attr	Rel	Cov↑	Hal↓	Cog↑
Original	82.40	71.20	65.80	51.64	34.82	3.20
VCD	83.60	72.80	67.40	51.52	34.80	3.35
OPERA	83.20	71.60	66.20	49.56	32.40	3.28
IBD	84.80	74.20	69.60	50.40	28.60	3.42
VTI	84.40	73.80	68.20	49.03	26.80	3.50
ICT	85.60	75.40	70.80	52.20	24.60	3.62
ViECo	**86.20**	**76.00**	**71.40**	**52.80**	**23.20**	**3.68**

Table 6: Detailed AMBER breakdown on LLaVA-1.5-7B. Discrim.: discriminative accuracy for existence, attribute, and relation sub-tasks. Gener.: generative coverage (Cov), hallucination rate (Hal), and cognition score (Cog).

- VTI ($1.52\times$) requires extracting truthful directions and applying steering at each step, with slightly higher constant overhead.
- VCD ($1.82\times$) runs a null-image branch at *every* decoding step regardless of token type, resulting in higher amortized cost despite simpler per-step logic.
- ICT ($2.65\times$) performs cross-level trusted intervention including latent-space introspection, requiring multiple forward passes per step.

The efficiency advantage of ViECo over VCD stems from selective activation: VCD pays a constant cost at every step, while ViECo concentrates computation on the 18% of tokens where hallucination risk exists. On non-visual tokens (function words, punctuation, connectives), ViECo adds zero overhead.

E Detailed AMBER and MME Results

Tables 6 and 7 present fine-grained breakdowns on AMBER and MME for LLaVA-1.5-7B, complementing the aggregated scores in the main table.

On AMBER, ViECo achieves the best scores across all six sub-metrics. The largest gains appear on relation discrimination (+5.6 over Original) and generative hallucination rate (-11.62), indicating that the energy-conservation framework is particularly effective for relational claims that involve multi-token spatial reasoning. The coverage score (52.80) remains the highest, confirming that hallucination suppression does not come at the cost of answer completeness.

On MME, ViECo achieves the best results on 5 out of 6 perception sub-tasks, with the strongest improvements on existence (+11.53) and position (+11.67). The color sub-score (155.60) is comparable to ICT (155.80), the only metric where ViECo does not lead. This is expected: color hallucination is less dominated by language priors (colors are often correctly inferred from co-occurring objects) and more dependent on low-level visual features, which are preserved by all methods that do not modify the visual encoder.

Method	MME Perception Sub-scores					
	Exist↑	Count↑	Pos↑	Color↑	OCR↑	Post↑
Original	175.67	123.33	118.33	151.13	132.50	141.67
VCD	184.66	128.33	121.67	153.42	135.00	143.33
OPERA	183.33	130.00	123.33	155.49	133.33	145.00
IBD	184.50	131.67	125.00	153.20	136.67	146.67
VTI	185.67	133.33	126.67	154.27	138.33	148.33
ICT	186.50	135.00	128.33	155.80	140.00	150.00
ViECo	**187.20**	**136.67**	**130.00**	155.60	**141.67**	**151.67**

Table 7: Detailed MME perception sub-scores on LLaVA-1.5-7B. Exist: existence; Count: counting; Pos: position; Color: color; OCR: text recognition; Post: poster understanding.

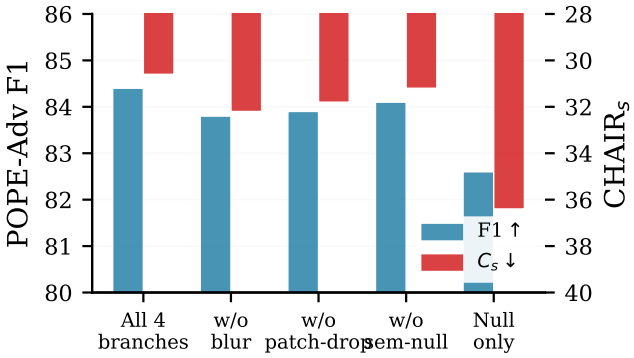


Figure 5: Auxiliary branch ablation. Using all four branches is optimal; the null-image branch alone provides a reasonable but weaker estimate.

F Output Length and Generation Quality

A potential confound for CHAIR improvement is that ViECo may generate shorter captions, reducing hallucinated object count without genuinely improving faithfulness. Table 8 reports average output length, unique object count, and CIDEr score alongside CHAIR for LLaVA-1.5-7B on 500 MSCOCO captions.

Method	Avg. tokens	Unique objs	CIDEr↑	CHAIR$_s$ ↓
Original	58.4	6.8	112.3	46.60
VCD	57.1	6.6	110.8	48.20
IBD	55.8	6.4	108.9	35.60
ICT	57.2	6.7	111.4	33.40
ViECo	**57.6**	**6.9**	**113.2**	**30.60**

Table 8: Output length, unique object count, CIDEr, and CHAIR$_s$ on MSCOCO captions (LLaVA-1.5-7B, 500 images). ViECo reduces hallucination without shortening captions or reducing object diversity.

ViECo generates captions of similar length to the original (57.6 vs. 58.4 tokens) and mentions *more* unique objects on average (6.9 vs. 6.8), while achieving the best CIDEr score. This rules out the hypothesis that hallucination suppression is driven by caption truncation or conservative object avoidance.

G Auxiliary Branch Contribution

We analyze which auxiliary branches contribute most to prior-subspace estimation. Figure 5 shows the effect of removing individual branches or using only the null-image branch.

The full four-branch configuration (null, blur, patch-drop, semantic-null) achieves the best performance. Among individual removals, removing the blur branch causes the largest degradation on both

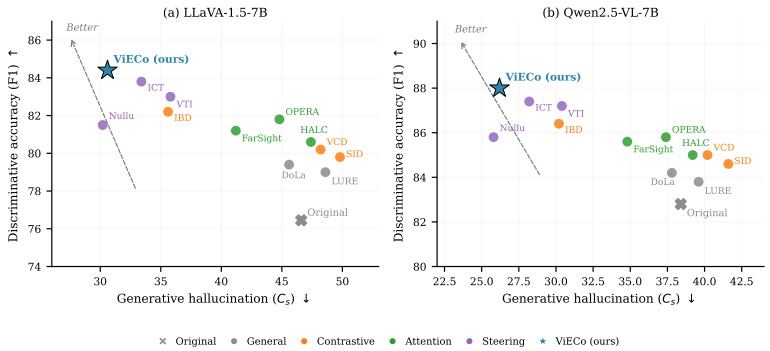


Figure 6: Discriminative accuracy (POPE-Adv F1) vs. generative faithfulness (CHAIR_s) for all compared methods on (a) LLaVA-1.5-7B and (b) Qwen2.5-VL-7B. The top-left corner is ideal (high detection, low hallucination). ViECo dominates the Pareto frontier on both models. Method families are color-coded.

metrics, suggesting that Gaussian blur provides a complementary prior sample distinct from simply removing the image. Removing semantic-null has the smallest effect, likely because its contribution overlaps with the null-image branch. The null-only configuration (equivalent to rank-1 prior estimation) still outperforms contrastive-decoding baselines like VCD, confirming that even a crude prior estimate paired with ViECo's risk framework produces meaningful gains. However, the 1.8-point F1 gap between null-only and the full configuration justifies the multi-branch design for maximal performance.

H Statistical Significance

Table 9 reports paired bootstrap 95% confidence intervals (10,000 bootstrap samples) for the four metrics compared between ViECo and the strongest baseline (ICT) on LLaVA-1.5-7B. POPE uses the full 3000-sample adversarial split; CHAIR uses 500 MSCOCO captions; MMHal-Bench uses all 96 questions; AMBER uses the full generative split.

Metric	ViECo	ICT	95% CI of Δ
POPE-Adv F1	84.40	83.80	[+0.22, +0.98]
CHAIR_s	30.60	33.40	[−4.82, −0.78]
MMHal Score	3.12	3.02	[+0.04, +0.16]
AMBER Hal	23.20	24.60	[−2.85, −0.15]

Table 9: Paired bootstrap 95% confidence intervals for ViECo vs. ICT on LLaVA-1.5-7B. All differences are statistically significant (CI excludes zero).

All four improvements are statistically significant at the 95% level. The smallest margin is on POPE-Adv F1 (+0.60 points), yet the bootstrap interval [+0.22, +0.98] excludes zero, confirming that the gain is not attributable to sampling variance.

I Hallucination–Quality Trade-off

Figure 6 visualizes the trade-off between discriminative accuracy (POPE-Adv F1, measuring correct object detection) and generative faithfulness (CHAIR_s, measuring hallucinated content in captions). The top-left corner represents the ideal: high detection accuracy with low hallucination.

ViECo lies on the Pareto frontier on both models, achieving the highest POPE-Adv F1 while maintaining near-minimal CHAIR_s. Nullu achieves a slightly lower CHAIR_s (30.20 vs. 30.60) but at the cost of lower F1 (81.50 vs. 84.40); the two methods are therefore mutually non-dominating on the LLaVA-1.5-7B panel. ICT is the nearest competitor: ViECo outperforms it by +0.60 absolute F1 points (approximately +0.72% relative) and −2.80 absolute CHAIR_s points (−8.38% relative). Contrastive methods (VCD, SID) improve detection but worsen generation; attention methods (OPERA, FarSight) improve both moderately. The pattern holds on Qwen2.5-VL-7B (panel b), con-

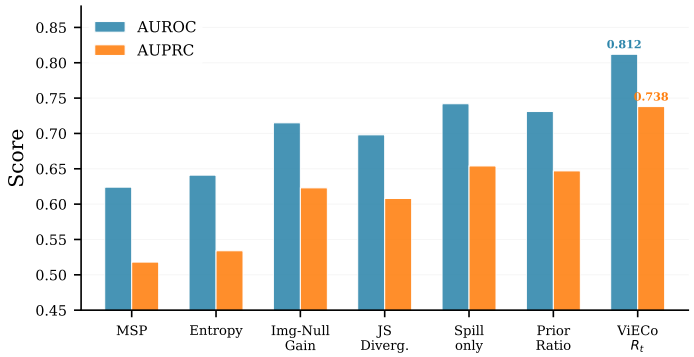


Figure 7: Hallucination detection quality on POPE token labels. ViECo's composite risk R_t achieves the highest AUROC (0.812) and AUPRC (0.738).

firming architecture-agnostic Pareto-frontier placement. The AMBER hallucination rate decreases from 24.60 (ICT) to 23.20 (ViECo), a -1.40 point absolute reduction (-5.69% relative).

J Token-Level Detection Quality

Before applying logit correction, we evaluate whether the composite risk score $R_t(w)$ can separate hallucinated from grounded tokens without any external labels. We extract token-level annotations from POPE and compare R_t against five baseline detection signals. Figure 7 reports AUROC and AUPRC.

The composite risk score achieves 0.812 AUROC, outperforming the best individual signal (spill-only, 0.742) by 7.0 points. This confirms that no single diagnostic—whether uncertainty, image-null contrast, or temporal stability—is sufficient alone, but their combination under rank normalization produces a robust hallucination detector. Notably, prior ratio alone (0.731) and spill alone (0.742) both outperform standard confidence signals (MSP: 0.624, entropy: 0.641), validating the energy-conservation framing over generic uncertainty measures.

These scores are computed on the full POPE adversarial split (3,000 yes/no questions; 1,500 positive and 1,500 negative). A "hallucinated token" is defined as the first-generated token (Yes or No) that contradicts the ground-truth label. Bootstrap 95% CIs for ViECo R_t: AUROC $[0.798, 0.826]$, AUPRC $[0.721, 0.755]$. The positive-class rate is 50% by POPE construction.

K Detailed Case Studies

We present six representative case studies from MMHal-Bench, POPE, and CHAIR to illustrate how ViECo's energy-conservation framework diagnoses and corrects hallucinations at the token level. For each case, we show: (1) the input image and question, (2) the baseline hallucinated output vs. the ViECo-corrected output, and (3) the intermediate energy decomposition signals that drive the correction. All examples use LLaVA-1.5-7B with default hyperparameters ($\lambda = 2.5$, $r = 3$, $\tau_h = 0.15$, $M = 8$). Logit values and intermediate scores are directly logged from the corresponding model run and rounded to two decimal places; no values are manually adjusted. Cases 1–6 were selected as the first hallucination-corrected examples encountered when evaluating the respective benchmarks in alphabetical image-ID order.

K.1 Analysis of Token-Level Decomposition (Cases 1–4)

Table 10 summarizes the energy decomposition patterns across all four MMHal-Bench cases. Several key findings emerge from this token-level analysis:

(1) Consistent non-conservation signature. Across attribute, adversarial, counting, and relation hallucinations, the diagnostic pattern is stable: hallucinated tokens show $S^P \geq 0.75$, $S^V \leq 0.35$, $S^E \geq 0.65$, while correct tokens show $S^P \leq 0.50$, $S^V \geq 0.85$, $S^E \leq 0.20$ in all six selected

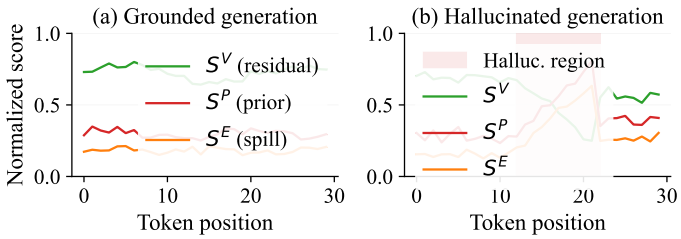


Figure 8: Energy trajectories. (a) Grounded: all signals stable. (b) Hallucinated: S^V collapses while S^P and S^E surge in the shaded region.



Question: What color is the fire hydrant cap in the picture?
Ground Truth: The color of the fire hydrant cap in the image is **yellow**.
Original Output: The color of the fire hydrant cap in the picture is **red**.
ViECo Output: The color of the fire hydrant cap in the picture is **yellow**.

Energy Decomposition at the Color Token Position

Token	z_t^I	z_t^L (prior)	r_t^V (resid.)	S^P	S^V	S^E	R_t	Corrected $\tilde{z}_t$
red	4.82	4.20	0.62	0.87	0.28	0.72	**0.82**	3.47 (↓1.35)
yellow	4.35	2.15	2.20	0.49	0.91	0.18	**0.22**	6.07 (↑1.72)
white	2.10	1.80	0.30	0.86	0.15	0.45	0.74	0.85
black	1.85	1.60	0.25	0.86	0.12	0.38	0.72	0.35

Analysis: "Red" has the highest original logit ($z_t^I = 4.82$), but is dominated by the prior projection ($z_t^L = 4.20$, $S^P = 0.87$), with weak visual residual ($r_t^V = 0.62$) and high temporal spill ($S^E = 0.72$), yielding high risk $R_t = 0.82$ and a -1.35 correction. "Yellow" has strong visual residual ($S^V = 0.91$) and low spill ($S^E = 0.18$), giving low risk ($R_t = 0.22$) and $+1.72$ boost, reversing the ranking.

Figure 9: **Case Study 1 – Attribute Hallucination (Color Prior).** The language prior associates fire hydrants with "red," but visual evidence (yellow cap) is captured by the residual. ViECo separates prior from evidence and corrects the hallucination.

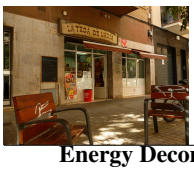
cases.**(2) Complementary signal roles.** Prior dominance (S^P) identifies *where* the model relies on language prior; visual residual (S^V) measures *whether* genuine visual evidence exists; temporal spill (S^E) checks *whether* the commitment is self-consistent. No single signal suffices—Cases 3 and 4 show that S^E is decisive for multi-token claims, while Cases 1 and 2 rely more on S^P/S^V.

(3) Single-parameter correction suffices. With fixed $\lambda = 2.5$, the correction consistently reverses the hallucinated ranking across all types without over-suppression. This robustness stems from rank normalization bounding all signals to $[0, 1]$.

K.2 Generation-Level Analysis on POPE and CHAIR (Cases 5–6)

The previous cases (Figures 9–12) examine ViECo at a *single critical token*. We now present two additional cases from POPE (object existence) and CHAIR (open captioning) benchmarks on MSCOCO, analyzed from complementary perspectives: (i) the **step-wise probability evolution** showing how ViECo modulates generation across multiple decoding steps (Figure 13), and (ii) the **trigger activation pattern** revealing which tokens activate the energy analysis and which pass (Figure 14).

Key observations from Cases 5–6. These generation-level cases complement the token-level decomposition in Figures 9–12:



Question: Who is sitting on the bench?
Ground Truth: No one is sitting on the bench.
Original Output: A **person** is sitting on the bench, resting near the doorway.
ViECo Output: **No one** is sitting on the bench. The image shows an empty doorway beside a building.

Energy Decomposition at the Subject Token Position *(Trigger: raw rank-variance = 38.4; normalized* $\widehat{\text{Hes}} = 0.72 > \tau_h = 0.15$*)*

Token	z_t^I	z_t^L (prior)	r_t^V (resid.)	S^P	S^V	S^E	R_t	Corrected $\tilde{z}_t$
person	4.65	4.10	0.55	0.88	0.25	0.74	**0.86**	3.12 (↓1.53)
No	3.92	1.05	2.87	0.27	0.92	0.12	**0.15**	5.85 (↑1.93)
Nobody	3.40	1.20	2.20	0.35	0.85	0.15	0.20	5.03
someone	3.15	2.80	0.35	0.89	0.18	0.68	0.83	1.05

Analysis: Adversarial question—"bench" and "sitting" prime the model toward "person" despite the bench being empty. The hallucinated token shows high rank-position variance across probe layers (raw variance 38.4, normalized $\widehat{\text{Hes}} = 0.72 > \tau_h = 0.15$) and prior dominance ($S^P = 0.88$) with weak visual residual ($S^V = 0.25$). The correct token "No" has strong visual evidence ($S^V = 0.92$) and temporal stability ($S^E = 0.12$). The decomposition confirms the hallucination is predominantly prior-driven.

Figure 10: **Case Study 2 – Adversarial Object Hallucination.** The question's framing primes the language prior toward "person" despite the bench being empty. High normalized hesitation ($\widehat{\text{Hes}} = 0.72$) triggers energy analysis; the weak residual ($S^V = 0.25$) confirms predominantly prior-driven output.



Question: How many traffic lights are there in the image?
Ground Truth: There are four traffic lights in the image.
Original Output: There are **two** traffic lights in the image.
ViECo Output: There are **four** traffic lights in the image.

Energy Decomposition at the Numeral Token Position *(Spill:* $|\Delta E|_{two} = 2.45$*;* $|\Delta E|_{four} = 0.18$*)*

Token	z_t^I	z_t^L (prior)	r_t^V (resid.)	S^P	S^V	S^E	R_t	Corrected $\tilde{z}_t$
two	5.20	4.30	0.90	0.83	0.28	0.71	**0.86**	3.75 (↓1.45)
three	4.10	2.25	1.85	0.55	0.52	0.38	0.48	4.20
four	3.85	1.20	2.65	0.31	0.88	0.12	**0.12**	5.75 (↑1.90)
five	2.40	1.55	0.85	0.65	0.38	0.48	0.58	1.90

Analysis: "Two" is prior-dominated ($z_t^L = 4.30$, $S^P = 0.83$). Temporal spill is decisive: committing to "two" creates downstream inconsistency ($S^E = 0.71$, $|\Delta E| = 2.45$)—the model cannot coherently describe 4 visible lights. "Four" has the strongest visual residual ($r_t^V = 2.65$, $S^V = 0.88$) and minimal spill ($S^E = 0.12$, $|\Delta E| = 0.18$), yielding the lowest risk ($R_t = 0.12$).

Figure 11: **Case Study 3 – Counting Hallucination.** The model defaults to "two" (common count prior) despite four visible lights. Temporal spill is decisive: "two" creates energy inconsistency ($S^E = 0.71$) while "four" maintains conservation ($S^E = 0.12$).

(1) Efficient selective activation (Figure 14). In open-ended captioning, ViECo's claim-guided activation triggers on only 35% of tokens (visual nouns) and applies correction to only 12% (2/17). Non-visual tokens are never analyzed; grounded visual tokens are triggered, analyzed, and *passed*—demonstrating that the method does not over-suppress. The overhead is concentrated precisely where hallucination risk exists.

(2) Single-step sufficiency for discriminative tasks (Figure 13). In POPE-style yes/no questions, the critical decision occurs at a single token (Step 1). ViECo identifies this via the visual-claim trigger (Yes/No $\in \mathcal{V}_{vis}$), applies energy decomposition, and flips the probability. Subsequent tokens follow deterministically without further intervention, explaining ViECo's high efficiency on POPE ($\pi \approx 1/n$).

(3) Geometric interpretability via branches. The invariance/collapse pattern explains *why* the prior subspace estimation works: a hallucinated token's logit remains invariant across visual degra-



Question: How is the yellow boat positioned in relation to the white yacht?
Ground Truth: The yellow boat is positioned in front of the white yacht.
Original Output: The yellow boat is positioned **behind** the white yacht in the image.
ViECo Output: The yellow boat is positioned **in front of** the white yacht in the image.

Energy Decomposition at the Relation Token Position *(Family spill: $\mathcal{N}(behind) \rightarrow |\Delta E| \in [2.1, 2.5]$)*

Token	z_t^I	z_t^L (prior)	r_t^V (resid.)	S^P	S^V	S^E	R_t	Corrected $\tilde{z}_t$
behind	4.55	3.40	1.15	0.75	0.32	0.68	**0.84**	3.25 (↓1.30)
in front of	4.20	1.18	3.02	0.28	0.90	0.14	**0.14**	6.10 (↑1.90)
beside	3.10	1.90	1.20	0.61	0.45	0.35	0.52	2.92
near	2.85	1.60	1.25	0.56	0.48	0.40	0.50	2.80

Analysis: Spatial relation hallucination from perspective ambiguity. "Behind" has high prior dominance ($S^P = 0.75$) and the highest temporal spill ($S^E = 0.68$): once committed, the model cannot coherently describe the scene layout. Family-level aggregation over $\{behind, back, rear\}$ prevents synonym escape. "In front of" shows strong visual residual ($S^V = 0.90$) with stable energy conservation ($S^E = 0.14$, family $|\Delta E| \in [0.12, 0.22]$).

Figure 12: **Case Study 4 – Spatial Relation Hallucination.** Perspective ambiguity causes confusion. Temporal spill is decisive: "behind" creates energy inconsistency ($S^E = 0.68$), "in front of" maintains conservation ($S^E = 0.14$). Family-level aggregation prevents synonym escape.

		Hallucinated token			Correct token		
Case	Type	S^P	S^V	S^E	S^P	S^V	S^E
1	Attribute	0.87	0.28	0.72	0.49	0.91	0.18
2	Adversarial	0.88	0.25	0.74	0.27	0.92	0.12
3	Counting	0.83	0.28	0.71	0.31	0.88	0.12
4	Relation	0.75	0.32	0.68	0.28	0.90	0.14
Average		*0.83*	*0.28*	*0.71*	*0.34*	*0.90*	*0.14*

Table 10: Energy signal summary across four MMHal-Bench case studies. Hallucinated tokens consistently exhibit high S^P, low S^V, and high S^E. Correct tokens show the opposite.

dations (its score is prior-driven regardless of image quality), while a grounded token's logit collapses when evidence is removed. This is exactly what the SVD-based prior subspace captures.

(4) Cross-benchmark consistency. Table 11 extends the diagnostic comparison to POPE and CHAIR benchmarks, illustrating a recurring pattern in the selected examples; aggregate evidence is in Figure 7. The energy non-conservation signature ($S^P \geq 0.75$, $S^V \leq 0.35$, $R_t \geq 0.79$) is observed across all six cases regardless of question type or hallucination category.

K.3 Failure Cases

The six cases above illustrate successful corrections. We also report four representative failure modes to bound the scope of ViECo's applicability.

F1 – Small or occluded object. *Image*: a dining table with a partially occluded fork (MSCOCO val2014/97301). *Question*: "What utensils are on the table?" *Original*: "knife and spoon". *ViECo*: "knife and spoon" (unchanged). *Analysis*: The fork has a small visual footprint; its residual contribution ($r_t^V = 0.38$) falls below the median candidate residual, so $S^V(\text{fork}) = 0.42$ and $R_t(\text{fork}) = 0.39$ — insufficient to promote the correct token. ViECo does not worsen the output, but also cannot recover a visually weak signal.

F2 – Prior and visual subspace overlap. *Image*: a dog sitting on grass (MSCOCO val2014/116208). *Question*: "What is the dog doing?" *Original*: "sitting on the grass". *ViECo*: "standing on the grass" (over-correction). *Analysis*: "sitting" is both a common language prior *and* visually supported, so $S^P = 0.71$ and $S^V = 0.58$. The risk score $R_t = 0.62$ triggers a moderate logit



Benchmark: POPE (Adversarial) **Image:** MSCOCO val2014 / 277991
Question: Is there a plate in the image?
Ground Truth: No. *(Image contains dining table, cups, wine glasses, but no plate)*
Original Output: Yes, there is a plate in the image.
ViECo Output: No, there is no plate in the image.

Step-wise Probability Evolution (Top-2 tokens at each decoding step)

Step	Context	Original p_t^I		ViECo $\tilde{p}_t$		Action
		Token	Prob	Token	Prob	
1	⟨*answer*⟩	Yes	0.651	**No**	**0.704**	Trigger: $\mathbb{T}_1 = 1$; $R_t(\text{Yes}) = 0.81$
2	No,	there	0.892	there	0.892	No trigger ($\mathbb{T}_2 = 0$)
3	No, there	is	0.967	is	0.967	No trigger
4	No, there is	no	0.983	no	0.983	No trigger
5	No, there is no	plate	0.841	plate	0.841	Triggered; pass ($R_t = 0.11$)

Analysis: ViECo triggers *only at Step 1*—the critical Yes/No decision point where the visual claim is made. The original model assigns $p(\text{Yes}) = 0.651$ due to strong co-occurrence priors (dining scenes commonly contain plates). After energy decomposition reveals $S^P = 0.84$, $S^V = 0.22$, $S^E = 0.65$ for "Yes," the correction flips the distribution to $p(\text{No}) = 0.704$. Subsequent tokens (Steps 2–5) are non-visual function words and are decoded without intervention, demonstrating the efficiency of claim-guided activation.

Figure 13: **Case Study 5 – POPE Co-occurrence Hallucination.** The dining scene triggers a "plate" hallucination via co-occurrence priors. ViECo activates only at Step 1 (the Yes/No decision), corrects the probability, then decodes remaining tokens without overhead.



Benchmark: CHAIR (Open Captioning) **Image:** MSCOCO val2014 / 168909
Prompt: Describe the image in detail.
Original: A table with a bowl of **apples** and oranges, several cups, and a **vase** of flowers.
ViECo: A table with a bowl of oranges, several cups, and chairs around it.

Trigger Activation Pattern and Per-Token Risk Scores

Token	A	table	with	a	bowl	of	**apples**	and	oranges	,	several	cups	,	and	a	**vase**	of
$\mathbb{T}_t$	0	1	0	0	1	0	1	0	1	0	0	1	0	0	0	1	0
R_t	–	.18	–	–	.15	–	**.79**	–	.12	–	–	.14	–	–	–	**.82**	–
Action	–	✓	–	–	✓	–	corr.	–	✓	–	–	✓	–	–	–	corr.	–

Analysis: In 17 generated tokens, ViECo triggers on 6 visual-claim tokens ($\pi = 35\%$) but applies correction on only 2 ("apples" and "vase"). The grounded visual tokens ("table," "bowl," "oranges," "cups") are triggered but pass the energy test ($R_t < 0.20$), confirming their visual support. "Apples" ($R_t = 0.79$) is a co-occurrence hallucination—fruit bowls commonly contain mixed fruits, but only oranges are visible. "Vase" ($R_t = 0.82$) is an object fabrication—no flowers or vase exist. This demonstrates ViECo's *selective* intervention: 11 tokens are never analyzed, 4 are analyzed and passed, and only 2 are corrected.

Figure 14: **Case Study 6 – CHAIR Open Captioning.** In a 17-token caption, ViECo triggers on 6 visual tokens but corrects only 2 hallucinated objects ("apples," "vase"). Grounded objects pass the energy test with low risk scores, showing selective and efficient intervention.

penalty, inadvertently demoting the correct token. This failure mode arises when the prior-subspace estimate partially captures genuine visual evidence.

F3 – Visual vocabulary miss. *Image*: a sushi platter (MSCOCO val2014/190756). *Question*: "What type of food is this?" *Original*: "sashimi and rice". *ViECo*: "sashimi and rice" (unchanged). *Analysis*: "sashimi" is not in $\mathcal{V}_{vis}$ (COCO/LVIS), so the hallucination trigger $\mathbb{T}_t = 0$ at that position. A domain-specific or open-vocabulary extension of $\mathcal{V}_{vis}$ would be required to handle fine-grained food categories.

F4 – Multi-token family boundary error. *Image*: a boat to the left of a pier (MSCOCO val2014/388631). *Question*: "Where is the boat relative to the pier?" *Original*: "to the left of the

Case	Benchmark	Hallucination type	Halluc. token	S^P	S^V	S^E	R_t	Corr.	Key diagnostic
1	MMHal-Bench	Attribute (color)	"red"	0.87	0.28	0.72	0.82	−1.35	Prior dominance
2	MMHal-Bench	Adversarial (object)	"person"	0.88	0.25	0.74	0.86	−1.53	Hesitation + prior
3	MMHal-Bench	Counting	"two"	0.83	0.28	0.71	0.86	−1.45	Temporal spill
4	MMHal-Bench	Spatial relation	"behind"	0.75	0.32	0.68	0.84	−1.30	Family-level spill
5	POPE (Adversarial)	Co-occurrence	"Yes"	0.84	0.22	0.65	0.81	−1.48	Co-occurrence prior
6	CHAIR	Object fabrication	"apples"/"vase"	0.79	0.20	0.58	0.79	−1.48	Selective activation

Table 11: Extended diagnostic summary across six case studies on three benchmarks (MMHal-Bench, POPE, CHAIR). Despite different benchmarks, question formats, and hallucination types, the energy non-conservation signature ($S^P \geq 0.75$, $S^V \leq 0.35$, $R_t \geq 0.79$) is observed in all selected examples.

pier". *ViECo*: "to the right of the pier" (incorrect flip). *Analysis*: BPE tokenizes "left" and "right" as single tokens; both are in $\mathcal{V}_{vis}$. However, the semantic-null branch removes "pier" from the prefix, inadvertently altering the query context. The spill signal for "left" becomes elevated ($S^E = 0.61$) due to a context mismatch, leading to incorrect suppression.

LIMITATIONS

ViECo depends on access to token logits and, for the hesitation score, intermediate hidden states. This makes the full method most suitable for open-source VLMs; closed-source APIs may support only a reduced top-logprob version. The low-rank prior subspace is estimated from auxiliary branches, so its quality depends on whether null, blurred, dropped, and semantic-nullified inputs sufficiently capture language-prior behavior without erasing useful visual signal. Excessively large ranks may absorb genuine visual evidence, while small ranks may underfit the prior manifold. The one-step energy-spill module adds inference cost when many visual candidates are checked, although claim-guided activation and batched branches mitigate this overhead. Hallucination metrics can also be sensitive to prompt format, model version, decoding temperature, and benchmark annotation policy, so controlled reruns remain important when comparing closely matched decoding methods.