MITIGATING HALLUCINATION IN VISUAL LANGUAGE MODELS WITH PARTIAL TRANSPORT KL-CONSTRAINED DECODING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) frequently describe objects that are not present in the image, limiting their use in settings where faithfulness matters. Existing inference-time mitigation methods typically rely on contrastive decoding, attention manipulation, or representation steering, but suppressing hallucinations can also remove valid visual content. We introduce POTION (Partial Optimal Transport Inference with KL ProjectiON), a training-free inference-time method that steers autoregressive VLM decoding toward visually grounded outputs. At each generation step, POTION solves a null-penalized partial optimal transport problem between top-K candidate-token embeddings and cached image-patch hidden states. The partial formulation automatically routes visual support mass to grounded candidates while assigning unsupported candidates to a null sink, yielding per-candidate evidence and risk scores that produce a logit correction. This correction is then projected onto a KL-ball centered at the base model's distribution, preventing aggressive hallucination suppression from collapsing object coverage. Experiments on MSCOCO and AMBER show consistent reductions in object hallucination across Qwen2.5-VL and Llama-3.2-Vision. In particular, POTION achieves state-of-the-art performance in mitigating object hallucinations, substantially reducing CHAIR scores while maintaining comparable object recall.

1 INTRODUCTION

Vision-language models (VLMs), such as LLaVA, InstructBLIP, and Qwen2.5-VL, make visual information accessible through natural-language interaction, supporting tasks from image description to visual question answering (Liu et al., 2023; Dai et al., 2023; Bai et al., 2025). Their usefulness, however, depends on whether their answers reflect the image rather than merely a plausible scene. A persistent failure is *object-existence hallucination*: mentioning objects that are absent from the input image (Rohrbach et al., 2019; Li et al., 2023). Prior work links such hallucinations to language priors, object co-occurrence biases, and imperfect visual grounding (Leng et al., 2023; Liu et al., 2024). Consequently, a model may generate contextually plausible object mentions that are unsupported by the image.

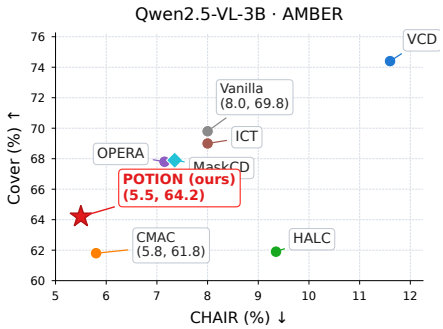


Figure 1: AMBER results for Qwen2.5-VL-3B-Instruct; lower CHAIR scores and higher coverage indicate better performance.

Addressing these hallucinations typically involves either model retraining or inference-time intervention. While retraining with curated data is effective, it is computationally expensive and difficult to transfer across different model families. Consequently, training-free inference methods emerged as a highly practical alternative. Existing approaches—such as contrastive decoding (Leng et al., 2023), attention manipulation (Huang et al., 2024; Li et al., 2026), and visual grounding (Chen et al., 2024b)—aim to steer the frozen model away from hallucinatory text directly. Yet, a fundamental challenge remains: simply suppressing fake objects is insufficient. Directly modifying the

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.

decoding process often hurts overall generation quality and accidentally suppresses valid visual details (Min et al., 2025). This highlights the need for a better mechanism—one that can accurately spot unsupported objects while precisely controlling the magnitude of the decoder's alteration.

We introduce **POTION** (Partial Optimal Transport Inference with KL ProjectiON), a training-free method that corrects object-related next-token predictions while constraining the size of the correction. During decoding, POTION evaluates object-related tokens among the top-K next-token candidates. It matches these candidates to visual features using *partial optimal transport* (Caffarelli & McCann, 2010; Figalli, 2010). A standard, balanced optimal transport would force every candidate to align with visual features—even when the object is absent from the image—resulting in spurious visual correspondences. POTION therefore introduces a null alternative that allows unsupported candidate demand to remain unmatched. The proportion of mass routed to this null sink provides direct signal of weak visual support, quantifiable signal of hallucination, which we then convert into candidate-level logit penalties.

However, aggressively penalizing tokens based on these imperfect visual-support estimates can inadvertently suppress valid object mentions, degrading the model's coverage. POTION therefore projects the corrected distribution onto a Kullback–Leibler (KL) ball centered on the base distribution, renormalized over the top-K candidates. It retains the full correction when the resulting distribution lies within the KL budget and scales the correction down otherwise. *In summary, partial transport estimates what to change, while KL control limits how far decoding may move*.

We evaluate POTION on MSCOCO (Lin et al., 2015) and AMBER (Wang et al., 2024) using Qwen2.5-VL-3B and Llama-3.2-Vision-11B. Figure 1 shows that POTION significantly improves the hallucination–coverage trade-off. On Qwen2.5-VL-3B (AMBER), POTION reduces the CHAIR metric by 31.3% relative to regular decoding (8.0% vs. 5.5%). Furthermore, it outperforms CMAC—the current state-of-the-art baseline—by achieving both lower hallucination rates (5.5% vs. 5.8%) and higher coverage (64.2% vs. 61.8%). This success translates across architectures: on MSCOCO with Llama-3.2-Vision-11B, POTION reduces CHAIR_I to 3.09%, a substantial 60.1% relative reduction from regular decoding.

In summary, our contributions are:

- **Partial-transport visual grounding.** We propose a null-penalized partial optimal transport formulation to evaluate next-token candidates. This allows unsupported candidates to remain unmatched, quantifying hallucinatory risks without forcing false visual alignments.
- **KL-constrained decoding.** We combine transport-based candidate rescoring with a KL projection that adaptively limits changes to the candidate-conditioned base distribution, without updating model parameters.
- **State-of-the-art hallucination mitigation.** Experiments on MSCOCO and AMBER show POTION reduces the CHAIR metric by up to 60.1% relative to standard decoding. Our component analysis further demonstrates that the KL constraint is vital, recovering 17.3 percentage points of object coverage lost by unrestricted rescoring.

2 RELATED WORK

2.1 OBJECT HALLUCINATION IN VLMS

Object-existence hallucination, objects absent from the input image, has been studied in both image captioning and VLMs (Rohrbach et al., 2019; Li et al., 2023). Rather than relying on costly model fine-tuning or instruction-data curation to address this, we focus on training-free interventions. These approaches operate at inference time, aiming to improve visual faithfulness by dynamically adjusting the generation process to suppress hallucinatory language priors in favor of true visual grounding.

A major line of work generates a modified prediction distribution and contrasts it with the original distribution to discourage hallucinated content. Visual Contrastive Decoding (VCD) (Leng et al., 2023) contrasts predictions from the original image with those from a noisy image, motivated by the amplification of language priors and statistical biases under visual uncertainty. HALC (Chen et al., 2024b) introduces adaptive focal-contrast decoding, combining local visual grounding with global

beam search to correct hallucinated content during generation. Image Head Masked Contrastive Decoding (MaskCD, (Deng & Yang, 2025) instead constructs the contrastive branch by masking attention heads associated with visual information. Cross-Modal Attention Calibration (CMAC) (Li et al., 2026) targets spurious inter-modality correlations and positional biases by combining image-token position calibration with contrastive decoding based on masked cross-modal value-vectors.

Complementary approaches exploit internal attention and representations: OPERA (Huang et al., 2024) penalizes excessive reliance on summary tokens during decoding, whereas ICT (Chen et al., 2024a) intervenes on attention-head activations carrying image- and object-level information. Beyond suppressing hallucinations, SumGD (Min et al., 2025) addresses generation quality by summarizing the text context and selectively adjusting image-related part-of-speech tokens. POTION shares this concern but separates visual-support estimation from explicit control of the resulting distribution change.

2.2 OPTIMAL TRANSPORT FOR HALLUCINATION MITIGATION

Optimal transport (OT) (Villani, 2009; Peyré & Cuturi, 2020) compares two distributions through a minimum-cost coupling with prescribed marginals, so that correspondences are resolved jointly rather than by independent nearest-neighbor matching. In vision–language learning, it has served mainly as a training-time alignment objective between words and regions (Chen et al., 2020), prompts and local features (Chen et al., 2023), or patches and tokens (Pramanick et al., 2023). These formulations are balanced, so every unit of mass must be matched. Partial OT (POT) (Caffarelli & McCann, 2010; Figalli, 2010) relaxes this by transporting only a fraction of the mass. It can be solved by adding a dummy point that absorbs the unmatched mass (Chapel et al., 2020), a device also used to model null word alignment (Arase et al., 2023).

OT has also been applied to hallucination. Guerreiro et al. (2023) detect hallucinations in machine translation using the Wasserstein distance between source-attention mass and a reference distribution. Onyshchuk et al. (2026) show that this signal is reliable under source disengagement but weak for errors arising downstream of attention. For VLMs, AIR (Zhu et al., 2026) uses entropic OT between hidden states and patch embeddings to select patches to inject into feed-forward layers. CoFi-Dec (Cao et al., 2025) fuses next-token distributions from multiple visual conditions with a Wasserstein-based rule. Scalpel (Shi et al., 2026) uses entropic OT to map hallucination-associated attention activations toward trusted ones.

In these methods, the transport plan assigns every unit of mass, so OT can rank, fuse, or map, but cannot express that a concept lacks visual support. POTION instead adds a null sink to the transport between object-related top-K candidates and cached visual features. The fraction of each candidate's demand routed to the sink serves as direct evidence of missing support. POTION leaves the forward pass and KV cache unmodified, uses the transport statistics only to rescore the assessed candidates, and bounds the resulting shift from the base distribution with a KL projection.

3 BACKGROUND

We consider a frozen vision-language model (VLM) with parameters θ, L decoder layers, and hidden dimension d. Given an image I and a text query q, the model generates a response $y = (y_1, \ldots, y_T)$ autoregressively:

$$p_\theta(y \mid I, q) \;=\; \prod_{t=1}^{T} p_\theta(y_t \mid x_t), \tag{1}$$

where $x_t = (I, q, y_{<t})$ denote the current context, and the prefix $y_{<t} = (y_1, \ldots, y_{t-1})$ contains the previously generated tokens. At step t, let $\mathbf{z}_t \in \mathbb{R}^{|\mathcal{V}|}$ be the corresponding next-token logits, where $\mathcal{V}$ is the tokenizer vocabulary.

In this work, we focus on *object-existence hallucination*, in which the generated response asserts the presence of an object that is absent from the image. Let $\mathcal{O}(I)$ denote the set of objects present in I and $\mathcal{O}(y)$ the object claims appearing in the response. An object-existence hallucination occurs whenever $\mathcal{O}(y) \setminus \mathcal{O}(I) \neq \varnothing$. The central difficulty is that $\mathcal{O}(I)$ is unavailable during inference. We therefore seek an inference-time signal that measures how strongly a prospective object mention

is supported by the model's visual representations, while preventing this signal from overriding the base decoder too aggressively.

To be efficient, our intervention operates only on the K most likely next-token candidates, $\mathcal{K}_t = \mathrm{TopK}(\mathbf{z}_t)$. Given decoding temperature $\tau > 0$, we define the corresponding renormalized base distribution

$$p_t(w) = \frac{\exp(z_t(w)/\tau)}{\sum_{v \in \mathcal{K}_t} \exp(z_t(v)/\tau)}, \qquad w \in \mathcal{K}_t. \tag{2}$$

This distribution serves as the reference decoder that our method will modify. In particular, the KL constraint introduced later controls the departure of the corrected distribution from $\mathbf{p}_t$.

Candidate object tokens. Rather than assess every token in $\mathcal{K}_t$, we intervene only when a candidate may introduce or complete an object mention. A fixed object vocabulary together with a token-suffix matcher will select a subset $\mathcal{A}_t \subseteq \mathcal{K}_t$ of tokens to be assessed against the image. Let $m_t := |\mathcal{A}_t|$ denote the number of such candidates, and index them as

$$\mathcal{A}_t = \{c_{t,j}\}_{j=1}^{m_t}. \tag{3}$$

Here, $c_{t,j} \in \mathcal{V}$ denotes the jth candidate token selected for visual assessment at decoding step t. For example, the token *light* may be selected when the current suffix allows it to complete the object phrase *traffic light*. Importantly, $c_{t,j}$ always denotes the *next token* whose logit is modified; the method does not assign a probability directly to the complete object phrase. Candidates in $\mathcal{K}_t \setminus \mathcal{A}_t$ remain available for decoding but receive zero transport correction. If $\mathcal{A}_t = \varnothing$, our method reduces to the reference decoder in equation 2.

Visual representations. We denote by $\mathcal{V}_I = \{v_i\}_{i=1}^{n},\ v_i \in \mathbb{R}^d$, the visual representations extracted from the VLM for image I. The precise extraction is architecture-dependent, and we use each model's native vision–language interface rather than assuming that visual features occupy explicit positions in the decoder sequence. For example, in Llama-3.2-Vision, we extract the projected image features provided to the language model as cross-attention memory (Meta AI, 2024); in Qwen2.5-VL, we use the merged visual-token representations produced by its vision encoder and vision–language merger (Bai et al., 2025). We cache these representations for decoding and, unless stated otherwise, use the representations corresponding to the final available visual feature layer.

4 Partial Optimal Transport Inference with KL Projection

Our method, POTION (Partial Optimal Transport Inference with KL ProjectiON), modifies the base decoder in two stages. First, we compare each assessed object candidate with the cached visual atoms through a partial transport problem. The resulting transport plan determines how much visual support each candidate receives and produces a raw logit correction. Second, we constrain this correction by projecting the resulting proposal distribution onto a KL ball around the base decoder. Thus, partial transport determines *what* the visual evidence suggests changing, whereas the KL constraint controls *how far* the decoder is allowed to move. For clarity, we suppress the decoding-step index t inside the transport problem. A diagram that summarizes our method can be found in Figure 2.

Notation. For a finite set A, $\Delta(A)$ denotes the probability simplex on A; Δ^n is the simplex of length-n probability vectors. We write $\mathbf{1}$ for an all-ones vector of the required dimension and $\langle \mathbf{C}, \mathbf{P} \rangle = \sum_{i,j} C_{ij} P_{ij}$. For distributions with a common support, their KL divergence is defined as $D_{\mathrm{KL}}(\mathbf{p} \| \mathbf{q}) = \sum_w p(w) \log(p(w)/q(w))$.

4.1 Visual Support through Partial Transport

Visual–candidate ground cost. For each assessed candidate $c_{t,j} \in \mathcal{A}_t$, let $e_j \in \mathbb{R}^d$ denote its corresponding row in the VLM output-embedding matrix. We measure the incompatibility between visual atom v_i and candidate $c_{t,j}$ using cosine distance:

$$C_{ij} = 1 - \frac{v_i^\top e_j}{\|v_i\|_2 \|e_j\|_2}, \qquad i = 1, \ldots, n, \quad j = 1, \ldots, m_t. \tag{4}$$

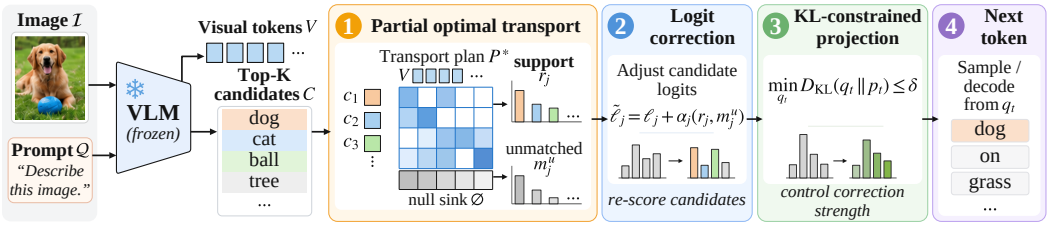


Figure 2: **Overview of POTION.** Partial optimal transport assigns visual support to object candidates while routing unsupported mass to a null sink. KL projection then limits the resulting correction before next-token decoding.

The resulting matrix $\mathbf{C} = (C_{ij}) \in \mathbb{R}^{n \times m_t}$ specifies the cost of supporting each assessed candidate with each visual atom.

Null-penalized partial transport. We assign uniform masses to the visual atoms and candidate tokens: $\mathbf{a} = 1/n\ \mathbf{1}_n,\ \mathbf{b} = 1/m_t\ \mathbf{1}_{m_t}$, where $\mathbf{a} \in \Delta^n$ is the visual mass distribution and $\mathbf{b} \in \Delta^{m_t}$ is the candidate mass distribution. A balanced transport problem would force every unit of candidate demand to be explained by some visual atom. This is undesirable for hallucination mitigation: an unsupported object candidate should instead be allowed to remain unmatched. We therefore introduce a null-mass vector $\mathbf{s} = (s_1, \ldots, s_{m_t})^\top \in \mathbb{R}_+^{m_t}$, where s_j denotes the unmatched mass associated with candidate $c_{t,j}$.

Our hallucination mitigation partial transport problem is then

$$\begin{aligned} \min_{\mathbf{P} \geq 0,\ \mathbf{s} \geq 0} \quad & \langle \mathbf{C}, \mathbf{P} \rangle + \kappa\, \mathbf{1}^\top \mathbf{s} \\ \text{s.t.} \quad & \mathbf{P}\mathbf{1} \leq \mathbf{a}, \\ & \mathbf{P}^\top \mathbf{1} + \mathbf{s} = \mathbf{b}. \end{aligned} \tag{5}$$

Here, $\mathbf{P} \in \mathbb{R}_+^{n \times m_t}$ is the transport plan and $\kappa \geq 0$ is the unit cost of assigning candidate demand to the null alternative. The entry P_{ij} is the amount of mass transported from visual atom v_i to candidate $c_{t,j}$, whereas s_j is the portion of candidate j that is left unsupported by real visual atoms.

In our formulation, the quantity κ acts as a null-matching threshold: a real visual transport is favored only when its cost is competitive with assigning the corresponding mass to the null alternative. The marginal constraint $\mathbf{P}\mathbf{1} \leq \mathbf{a}$ also couples the candidates considered at the current decoding step: different candidate continuations compete for a finite amount of visual support.

4.2 From Transport to a Candidate-Score Correction

The optimal transport solution provides both a measure of visual support and a visual representation associated with each candidate. For each candidate $c_{t,j}$, define the three transport statistics

$$m_j := \sum_{i=1}^{n} P_{ij}^\star, \quad u_j := \frac{s_j^\star}{b_j}, \quad r_j := \frac{\sum_{i=1}^{n} P_{ij}^\star v_i}{m_j + \varepsilon_m}, \qquad j = 1, \ldots, m_t. \tag{6}$$

The quantity m_j is the visual mass assigned to candidate j, while u_j is the fraction of its demand routed to the null alternative. Therefore, a large u_j indicates weak visual support. The vector r_j summarizes the visual atoms transported to candidate j. We add a small quantity ε_m in the calculation of r_j to avoid numerical instability when $m_j = 0$.

To evaluate how this visual evidence affects the candidate representation, we temporarily append $c_{t,j}$ to the current context and extract its contextual hidden state

$$h_j = H^{(\ell)}(x_t, c_{t,j})_{\text{last}} \in \mathbb{R}^d. \tag{7}$$

All candidate branches share the same prefix and can be evaluated in a single batch; only the token ultimately selected by the decoder is appended to the live generation state. We then form the auxiliary visually adjusted state

$$\widetilde{h}_j = h_j + \eta(r_j - h_j), \tag{8}$$

where $\eta \geq 0$ controls the strength of the visual interpolation. Let $g_\theta(h)_w$ denote the logit assigned to token w by the frozen output head when applied to representation h, our steering quantity is then defined as

$$\boxed{\Delta_t(w) = \begin{cases} g_\theta(\widetilde{h}_j)_{c_{t,j}} - g_\theta(h_j)_{c_{t,j}} - \lambda u_j, & w = c_{t,j} \in \mathcal{A}_t, \\ 0, & w \in \mathcal{K}_t \setminus \mathcal{A}_t, \end{cases}} \tag{9}$$

where $\lambda \geq 0$ controls the penalty assigned to unsupported mass. Intuitively, Δ_t adjusts each assessed candidate according to its agreement with the transported visual evidence, while explicitly penalizing unsupported mass through u_j. Candidates outside $\mathcal{A}_t$ are left unchanged, so the intervention remains localized to object-relevant alternatives.

4.3 KL-Constrained Distributional Correction

Applying the transport correction Δ_t without restriction can strongly suppress object mentions, including visually valid ones. We therefore treat Δ_t as a *proposal correction* and constrain how far it may move the decoder from the reference distribution $\mathbf{p}_t$.

More formally, Δ_t defines the proposal distribution

$$\widetilde{p}_t(w) = \frac{p_t(w)\exp(\Delta_t(w)/\tau)}{\sum_{v\in\mathcal{K}_t} p_t(v)\exp(\Delta_t(v)/\tau)}, \qquad w \in \mathcal{K}_t. \tag{10}$$

To prevent this correction from moving too far from the base decoder, we project the proposal onto a KL trust region (with the radius $\delta > 0$):

$$\boxed{\mathbf{p}_t^\star = \arg\min_{\mathbf{p}\in\Delta(\mathcal{K}_t)} D_{\mathrm{KL}}(\mathbf{p}\|\widetilde{\mathbf{p}}_t) \quad \text{s.t.} \quad D_{\mathrm{KL}}(\mathbf{p}\|\mathbf{p}_t) \leq \delta.} \tag{11}$$

The solution lies on the exponential-tilt path

$$p_t^{(\alpha)}(w) = \frac{p_t(w)\exp(\alpha\Delta_t(w)/\tau)}{\sum_{v\in\mathcal{K}_t} p_t(v)\exp(\alpha\Delta_t(v)/\tau)}, \qquad \alpha \in [0,1], \tag{12}$$

with $\mathbf{p}_t^\star = \mathbf{p}_t^{(\alpha_t)}$. If $D_{\mathrm{KL}}(\widetilde{\mathbf{p}}_t\|\mathbf{p}_t) \leq \delta$, then $\alpha_t = 1$; otherwise, $\alpha_t \in (0,1)$ is the unique solution of

$$D_{\mathrm{KL}}\left(\mathbf{p}_t^{(\alpha_t)}\middle\|\mathbf{p}_t\right) = \delta, \tag{13}$$

which can be obtained efficiently by one-dimensional bisection. Thus, the KL constraint preserves the direction of the transport-based correction while adaptively scaling its magnitude through α_t. Small δ keeps decoding close to the original model, whereas sufficiently large δ recovers the full correction.

The complete POTION procedure is given in Algorithm 1.

5 Experiments

We evaluate whether POTION reduces object-existence hallucination and how this reduction affects the amount of correctly described visual content. We report hallucination, object coverage, and realized response length jointly, then use a component ablation to examine the role of the KL constraint. Throughout, we report observations with their source-run provenance rather than treating them as a controlled speed benchmark.

5.1 Experimental Setup

Models and baselines. We use two model families: Qwen2.5-VL-3B-Instruct (Bai et al., 2025) and Llama-3.2-Vision-11B (Meta AI, 2024). We compare regular decoding with VCD (Leng et al., 2023), OPERA (Huang et al., 2024), HALC (Chen et al., 2024b), CMAC (Li et al., 2026), ICT (Chen et al., 2024a), and MaskCD (Deng & Yang, 2025). POTION denotes ours (partial-transport and KL-constrained method).

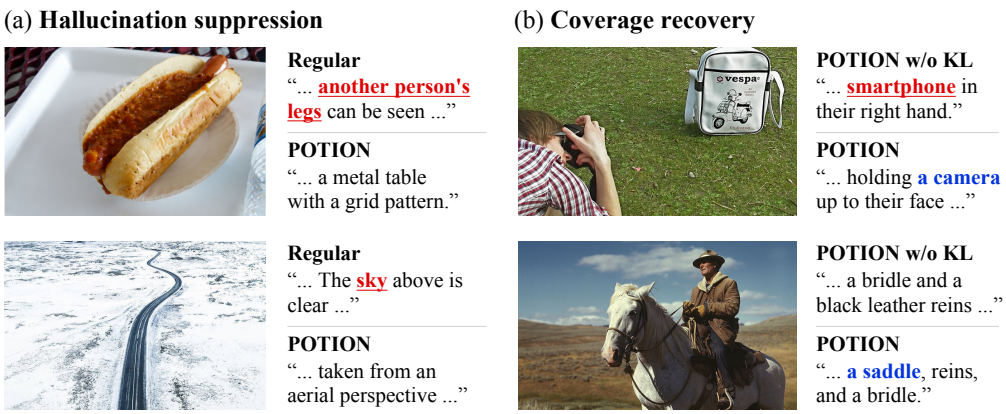


Figure 3: Qualitative comparison of model-generated descriptions on AMBER. (a) POTION avoids the unsupported person mention produced by regular decoding. (b) Adding KL control recovers the visible camera, misidentified as a smartphone without KL. Red marks unsupported mentions; **blue** marks recovered content. Additional examples are provided in Appendix G.

Datasets and metrics. On MSCOCO (Lin et al., 2015), we evaluate 500 validation images with the prompt *"Please describe this image in detail."* We report CHAIR_S and CHAIR_I (Rohrbach et al., 2019) at the caption and object-mention levels, together with object Recall. On the 1,004 generative examples of AMBER (Wang et al., 2024), we use *"Describe this image."* and report CHAIR, Cover, Hal, and Cog. More specifically, CHAIR and Hal quantify object-mention and response-level hallucination; Cover measures coverage of annotated objects; Cog is the benchmark's cognition-based hallucination metric. Each benchmark retains its own object-matching protocol, so similarly named scores are not pooled across datasets.

Generation protocol. Across all methods, we constrain the generation process to a maximum length of 512 tokens. Since models frequently terminate generation before reaching this limit, we additionally report the mean realized output length per image (Avg. Tok.). Observed Tok/s retains the recorded source-run measurements. For both Qwen2.5-VL and Llama3.2-vision, POTION uses a partial-transport cost based on LM-head embedding cosine distance, $\kappa = 0.92$, grounded-state interpolation coefficient $\eta = 0.5$, hallucination penalty $\lambda = 8$, KL radius $\delta_{\text{KL}} = 0.1$, and a candidate pool of top-5 tokens. The hyperparameter are selected on 200 MSCOCO validation images disjoint from the 500 evaluation images and then fixed for all reported settings. We use the final hidden layer, object-claim filtering, candidate-set decoding, greedy generation, and KV caching.

5.2 Results

Results on MSCOCO. POTION substantially reduces object hallucination on both evaluated models (Table 1), with particularly strong results on Llama. It achieves the lowest CHAIR_S and CHAIR_I among all compared methods, reducing the regular scores from 34.40 to 17.40 and from 7.75 to 3.09, respectively. These improvements correspond to relative reductions of 49.4% and 60.1%. POTION further outperforms the strongest competing result for each metric by 3.2 percentage points in CHAIR_S over OPERA and 1.19 percentage points in CHAIR_I over MaskCD. These gains are achieved with slightly longer responses than either competitor: 172.45 tokens versus 167.43 for OPERA and 167.44 for MaskCD.

On Qwen, POTION achieves the joint-lowest CHAIR_I of 5.6, matching CMAC, and the second-lowest CHAIR_S of 31.4. Relative to regular, these results represent reductions of 49.1% in CHAIR_I and 40.3% in CHAIR_S, while Recall increases from 66.8 to 68.16. Together, these results demonstrate substantial hallucination reductions over Regular across both models, with POTION leading both CHAIR metrics on Llama.

Results on AMBER. POTION achieves the lowest CHAIR on both evaluated models (Table 2). On Qwen, it reduces CHAIR from 8.0 to 5.5 relative to regular, a 31.3% relative reduction. It also lowers Hal from 49.0 to 38.3 and Cog from 5.3 to 3.1, ranking second on both metrics. Compared

Table 1: **MSCOCO-500 in the 512-token evaluation setting.** C_S and C_I denote CHAIR_S and CHAIR_I; quality scores are percentages. Avg. Tok. is realized output length. Bold and underline mark the best and second-best quality values within one model.

	Qwen2.5-VL-3B					Llama-3.2-Vision-11B				
Method	$C_S\downarrow$	$C_I\downarrow$	Recall↑	Avg. Tok.	Tok/s↑	$C_S\downarrow$	$C_I\downarrow$	Recall↑	Avg. Tok.	Tok/s↑
Regular	52.6	11.0	66.8	409.5	32.38	34.40	7.75	67.46	199.54	31.69
VCD	48.4	8.1	**73.0**	453.4	15.24	28.60	6.27	67.52	144.48	15.11
OPERA	33.8	<u>6.0</u>	<u>70.9</u>	321.7	9.54	<u>20.60</u>	4.47	67.90	167.43	14.59
HALC	34.0	7.6	67.5	329.1	14.9	27.60	5.35	**70.80**	209.13	4.0
CMAC	**28.4**	**5.6**	70.7	194.0	7.96	24.20	5.30	<u>69.37</u>	144.91	12.98
ICT	51.2	11.1	67.3	412.5	27.54	34.80	8.06	67.07	202.87	26.93
MaskCD	39.8	7.2	69.2	377.2	14.2	21.00	<u>4.28</u>	69.05	167.44	13.24
POTION	<u>31.4</u>	**5.6**	68.16	355.1	23.55	**17.40**	**3.09**	66.37	172.45	23.12

Table 2: **Generative evaluation on AMBER-1,004.** The quality comparison uses the 512-token evaluation setting. POTION has the lowest CHAIR in both model blocks. Quality scores are percentages; highlighting follows Table 1.

	Qwen2.5-VL-3B						Llama-3.2-Vision-11B					
Method	CHAIR↓	Cover↑	Hal↓	Cog↓	Avg. Tok.	Tok/s↑	CHAIR↓	Cover↑	Hal↓	Cog↓	Avg. Tok.	Tok/s↑
Regular	8.0	<u>69.8</u>	49.0	5.3	213.93	30.47	<u>5.7</u>	67.2	44.9	3.4	186.58	31.8
VCD	11.6	**74.4**	75.0	9.6	394.63	5.12	8.3	<u>67.9</u>	53.0	<u>2.9</u>	154.27	15.45
OPERA	7.2	67.8	46.5	4.9	207.30	6.72	5.9	66.1	**39.3**	**2.8**	163.12	14.19
HALC	9.4	61.9	48.1	3.8	234.45	5.74	6.7	67.6	49.9	3.3	193.95	4.14
CMAC	<u>5.8</u>	61.8	**32.2**	**2.0**	98.36	7.32	7.3	**68.4**	51.7	3.1	147.11	13.23
ICT	8.0	69.0	50.2	5.2	214.53	26.05	5.9	67.3	48.9	3.6	189.03	27.58
MaskCD	7.3	67.9	49.6	4.5	206.19	12.42	6.4	67.8	49.7	3.8	178.46	13.03
POTION	**5.5**	64.2	<u>38.3</u>	<u>3.1</u>	181.77	21.81	**5.1**	64.2	<u>40.1</u>	3.0	185.75	21.47

Table 3: **Component ablation on Qwen2.5-VL-3B using the AMBER benchmark (maximum new tokens = 1024).** POT attains the lowest hallucination scores but at the cost of reduced coverage. The integration of KL successfully recovers this coverage deficit.

Variant	CHAIR↓	Cover↑	Hal↓	Cog↓	Tok/s↑
Base	8.0	**69.8**	49.0	5.4	61.56
Base + POT	**2.4**	47.2	**12.0**	**0.7**	44.02
Base + POT + KL	5.8	64.5	38.6	3.3	44.65

with CMAC, the next-best method in CHAIR, POTION achieves lower CHAIR (5.5 versus 5.8) together with higher Cover (64.2 versus 61.8).

On Llama, POTION reduces CHAIR from 5.7 to 5.1 and Hal from 44.9 to 40.1 relative to regular, corresponding to relative reductions of 10.5% and 10.7%, respectively. Notably, these improvements occur at nearly unchanged average output length: 185.75 tokens for POTION versus 186.58 for regular. POTION also obtains the second-lowest Hal on this model. Its leading CHAIR scores and second-best Hal scores across both models provide complementary evidence of its effectiveness in mitigating object hallucination.

5.3 Component Ablation

Larger candidate sets yield modest coverage gains at lower observed throughput. Increasing K from 5 to 20 raises Cover from 64.5 to 65.5, while the recorded rate decreases from 44.61 to 35.75 Tok/s, a 19.9% reduction (Table 4). Hallucination does not improve monotonically: CHAIR is lowest at $K = 15$, Hal at $K = 10$, and Cog is tied at $K = 5$ and $K = 15$. Thus this series

supports $K = 5$ as a lower-cost operating point with only a small coverage difference from larger candidate sets; it does not establish $K = 5$ as the quality optimum.

Table 4: **Candidate-set size: coverage and observed throughput.** All quality metrics are percentages. Rates are observations from this series, not a cross-series speed comparison.

K	CHAIR↓	Cover↑	Hal↓	Cog↓	Observed Tok/s↑
5	5.8	64.5	38.6	**3.3**	44.61
10	5.7	64.9	**37.6**	3.4	40.36
15	**5.4**	65.1	38.9	**3.3**	38.22
20	5.7	**65.5**	40.2	3.5	35.75

KL moderates hallucination suppression. In the ablation (Table 3), POT-only lowers CHAIR from 8.0 to 2.4 and Hal from 49.0 to 12.0, but Cover drops by 22.6 points, from 69.8 to 47.2. Adding KL increases Cover to 64.5, recovering 17.3 points, or 76.5% of the lost coverage (Figure 4b). CHAIR rises to 5.8 but remains below the Base value of 8.0. Thus, KL moderates aggressive correction and partially recovers valid visual content; Cover still trails Base by 5.3 points.

KL controls the trade-off. Figure 4a also shows the KL-sensitivity results from Table 5. Increasing δ from 0.01 to 0.50 lowers CHAIR from 6.3 to 4.6, while Cover decreases from 66.8 to 60.7. The highlighted $\delta = 0.10$ setting yields CHAIR of 5.8 and Cover of 64.5: it suppresses hallucination relative to Base while retaining substantially more coverage than POT-only. These results illustrate the complementary roles of transport rescoring and KL control, rather than a single setting that maximizes both suppression and coverage.

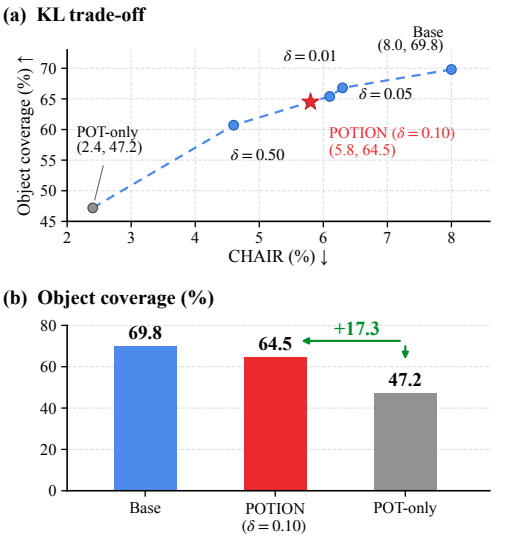


Figure 4: **KL balances hallucination and coverage.** (a) CHAIR–coverage trade-off across KL budgets. (b) Adding KL at $\delta = 0.10$ recovers 17.3 percentage points of coverage over POT-only. Values are from the component-ablation and KL-sensitivity series.

Takeaway. The ablations show that POTION's two stages address complementary failure modes: partial transport lets unsupported candidates abstain from forced visual alignment, while the KL trust region prevents over-correction of valid content. KL recovers 76.5% of the coverage lost by unrestricted rescoring, and its radius exposes a controllable hallucination–coverage operating point. This supports a broader inference-time principle: separate visual-support estimation from control of how strongly it alters generation.

6 Conclusion

We introduced POTION, a training-free decoding framework combining null-penalized partial optimal transport with KL-bounded distributional correction. Across two VLM families and two benchmarks, POTION consistently reduces object hallucination relative to regular decoding, achieving up to a 60.1% reduction in CHAIR metrics and the lowest AMBER CHAIR in all evaluated model–budget settings. Beyond these gains, POTION provides a reusable mechanism for faithful generation: allow unsupported candidate demand to remain unmatched, then bound its influence on the base decoder. Future work can extend this principle to open-vocabulary candidates and confidence-adaptive KL control.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript editing and submission-compliance checking. All AI-assisted suggestions were reviewed and verified by the authors. The authors take responsibility for the final content of this work, including its text, claims, experiments, and artifacts.

REPRODUCIBILITY STATEMENT

Section 5.1 specifies the evaluated models, datasets, baselines, generation protocol, and hyperparameters used in our experiments. Appendix A provides the complete POTION decoding procedure. Appendix B describes the evaluation metrics and aggregation protocols, while Appendices C–F report sensitivity analyses, generation-budget robustness, and runtime accounting. Additional qualitative and statistical evaluations are provided in Appendix G. Code is available in our repository.

REFERENCES

Yuki Arase, Han Bao, and Sho Yokoi. Unbalanced Optimal Transport for Unbalanced Word Alignment. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3966–3986, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.219. URL https://aclanthology.org/2023.acl-long.219/.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL Technical Report. arXiv preprint arXiv:2502.13923v1, 2025. URL https://arxiv.org/abs/2502.13923v1.

Luis Caffarelli and Robert J. McCann. Free boundaries in optimal transport and Monge-Ampère obstacle problems. *Annals of Mathematics*, 171(2):673–730, 2010. doi: 10.4007/annals.2010.171.673. URL https://annals.math.princeton.edu/2010/171-2/p02.

Zongsheng Cao, Yangfan He, Anran Liu, Jun Xie, Feng Chen, and Zepeng Wang. CoFi-Dec: Hallucination-Resistant Decoding via Coarse-to-Fine Generative Feedback in Large Vision-Language Models. arXiv preprint arXiv:2512.23453v1, 2025. URL https://arxiv.org/abs/2512.23453v1.

Laetitia Chapel, Mokhtar Z. Alaya, and Gilles Gasso. Partial Optimal Transport with Applications on Positive-Unlabeled Learning. arXiv preprint arXiv:2002.08276v2, 2020. URL https://arxiv.org/abs/2002.08276v2.

Guangyi Chen, Weiran Yao, Xiangchen Song, Xinyue Li, Yongming Rao, and Kun Zhang. PLOT: Prompt Learning with Optimal Transport for Vision-Language Models. arXiv preprint arXiv:2210.01253v2, 2023. URL https://arxiv.org/abs/2210.01253v2.

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. arXiv preprint arXiv:2411.15268v1, 2024a. URL https://arxiv.org/abs/2411.15268v1.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: UNiversal Image-TExt Representation Learning. arXiv preprint arXiv:1909.11740v3, 2020. URL https://arxiv.org/abs/1909.11740v3.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. HALC: Object Hallucination Reduction via Adaptive Focal-Contrast Decoding. arXiv preprint arXiv:2403.00425v2, 2024b. URL https://arxiv.org/abs/2403.00425v2.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. arXiv preprint arXiv:2305.06500v2, 2023. URL https://arxiv.org/abs/2305.06500v2.

Jingyuan Deng and Yujiu Yang. MaskCD: Mitigating LVLM Hallucinations by Image Head Masked Contrastive Decoding. arXiv preprint arXiv:2510.02790v1, 2025. URL https://arxiv.org/abs/2510.02790v1.

Alessio Figalli. The Optimal Partial Transport Problem. *Archive for Rational Mechanics and Analysis*, 195(2):533–560, 2010. doi: 10.1007/s00205-008-0212-7. URL https://doi.org/10.1007/s00205-008-0212-7.

Nuno M. Guerreiro, Pierre Colombo, Pablo Piantanida, and André F. T. Martins. Optimal Transport for Unsupervised Hallucination Detection in Neural Machine Translation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13766–13784, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.770. URL https://aclanthology.org/2023.acl-long.770/.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: Alleviating Hallucination in Multi-Modal Large Language Models via Over-Trust Penalty and Retrospection-Allocation. arXiv preprint arXiv:2311.17911v3, 2024. URL https://arxiv.org/abs/2311.17911v3.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding. arXiv preprint arXiv:2311.16922v1, 2023. URL https://arxiv.org/abs/2311.16922v1.

Jiaming Li, Jiacheng Zhang, Zequn Jie, Lin Ma, Ming Li, Xiaonan Luo, and Guanbin Li. Cross-Modal Attention Calibration for LVLM Hallucination Mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 40186–40196, June 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Li_Cross-Modal_Attention_Calibration_for_LVLM_Hallucination_Mitigation_CVPR_2026_paper.html.

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft COCO: Common Objects in Context. arXiv preprint arXiv:1405.0312v3, 2015. URL https://arxiv.org/abs/1405.0312v3.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A Survey on Hallucination in Large Vision-Language Models. *arXiv preprint arXiv:2402.00253v2*, 2024. URL https://arxiv.org/abs/2402.00253v2.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. arXiv preprint arXiv:2304.08485v2, 2023. URL https://arxiv.org/abs/2304.08485v2.

Meta AI. Llama 3.2: Revolutionizing Edge AI and Vision with Open, Customizable Models. Meta AI blog, September 2024. URL https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/. Published September 25, 2024. Accessed September 25, 2026.

Kyungmin Min, Minbeom Kim, Kang-il Lee, Dongryeol Lee, and Kyomin Jung. Mitigating hallucinations in large vision-language models via summary-guided decoding. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 4183–4198, 2025. doi: 10.18653/v1/2025.findings-naacl.235. URL https://aclanthology.org/2025.findings-naacl.235/.

Mariia Onyshchuk, Maksym-Vasyl Tarnavskyi, and Marta Sumyk. Layer-Resolved Optimal Transport for Hallucination Detection in NMT and Abstractive Summarization. arXiv preprint arXiv:2606.13216v1, 2026. URL https://arxiv.org/abs/2606.13216v1.

Gabriel Peyré and Marco Cuturi. Computational Optimal Transport. arXiv preprint arXiv:1803.00567v4, 2020. URL https://arxiv.org/abs/1803.00567v4.

Shraman Pramanick, Li Jing, Sayan Nag, Jiachen Zhu, Hardik Shah, Yann LeCun, and Rama Chellappa. VoLTA: Vision-Language Transformer with Weakly-Supervised Local-Feature Alignment. arXiv preprint arXiv:2210.04135v3, 2023. URL https://arxiv.org/abs/2210.04135v3.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object Hallucination in Image Captioning. arXiv preprint arXiv:1809.02156v2, 2019. URL https://arxiv.org/abs/1809.02156v2.

Ziqiang Shi, Rujie Liu, Shanshan Yu, Satoshi Munakata, and Koichi Shirahata. Scalpel: Fine-Grained Alignment of Attention Activation Manifolds via Mixture Gaussian Bridges to Mitigate Multimodal Hallucination. arXiv preprint arXiv:2602.09541v1, 2026. URL https://arxiv.org/abs/2602.09541v1.

Cédric Villani. *Optimal transport: old and new*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009. doi: 10.1007/978-3-540-71050-9. URL https://link.springer.com/book/10.1007/978-3-540-71050-9.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. AMBER: An LLM-free Multi-dimensional Benchmark for MLLMs Hallucination Evaluation. arXiv preprint arXiv:2311.07397v2, 2024. URL https://arxiv.org/abs/2311.07397v2.

Xingyu Zhu, Kesen Zhao, Liang Yi, Shuo Wang, Zhicai Wang, Beier Zhu, Hanwang Zhang, and Xiangnan He. Look Carefully: Adaptive Visual Reinforcements in Multimodal Large Language Models for Hallucination Mitigation. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2602.24041v1.

A POTION ALGORITHM AND KL PROJECTION

Algorithm 1 summarizes one decoding step. Cache visual atoms once per response and repeat the step until the model's stopping criterion is met.

Algorithm 1 POTION: one decoding step

Require: Frozen VLM; context x_t; cached atoms $\{v_i\}_{i=1}^{n}$;
object matcher; $K, \tau > 0$, $\kappa, \eta, \lambda, \delta \geq 0$; bisection tolerance $\xi \in (0, 1)$.
Ensure: Next token y_t.

```
Compute base logits z_t and K_t ← TopK(z_t).
p_t ← softmax(z_t|_{K_t}/τ); A_t ← ObjectMatch(x_t, K_t).
Set d_t(w) ← 0 for all w ∈ K_t.
if A_t ≠ ∅ then
    Form C, a, b by Eq. equation 4.
    Obtain (P*, s*) by solving Eq. equation 5.
    Evaluate h_j = H^(ℓ)(x_t, c_{t,j})_last in batched.
    Compute m_j, u_j, r_j by Eq. equation 6 and d_t by Eq. equation 9.
end if
Select α_t by Eq. equation 12: retain 1 if feasible; otherwise bisect.
p*_t ← p_t^(α_t) using Eq. equation 10.
Decode y_t from p*_t and advance only its unmodified branch.
```

B ADDITIONAL EXPERIMENTAL DETAILS

B.1 METRICS AND AGGREGATION

MSCOCO. CHAIR_I is the fraction of evaluated object mentions labeled hallucinated; CHAIR_S is the fraction of evaluated captions containing at least one hallucinated object (Rohrbach et al., 2019).

AMBER. CHAIR and Hal measure object-mention and response-level hallucination, respectively; Cover measures the fraction of annotated objects mentioned. Cog is a hallucination measure (Wang et al., 2024). We use ↓ for CHAIR, Hal, and Cog, and ↑ for Cover. The supplied cap table used Cog↑ with all numbers retained.

For the public AMBER evaluator, let h_i and m_i denote its hallucinated and total evaluated mention counts, c_i and a_i its covered and annotated safe-object counts, and g_i and b_i its matched and total target hallucination-object counts.

$$\text{CHAIR} = 100\frac{\sum_i h_i}{\sum_i m_i}, \qquad \text{Cover} = 100\frac{\sum_i c_i}{\sum_i a_i}, \tag{14}$$

$$\text{Hal} = \frac{100}{N}\sum_i \mathbf{1}[h_i > 0], \qquad \text{Cog} = 100\frac{\sum_i g_i}{\sum_i b_i}. \tag{15}$$

Length and stopping. Avg. Tok. is the mean number of generated response tokens under the run's tokenizer. A cap hit records reaching the configured output limit; whether an EOS-at-cap example counts as a hit must be specified. Cap hits indicate possible output censoring. Token counts are compared within a model.

C COMPONENT ABLATION AND SENSITIVITY ANALYSES

The component comparison tests the role of limiting an existing correction. The three sensitivity analyses then examine its KL control, candidate-set size, and representation depth. They support conclusions about the reported operating points.

C.1 SENSITIVITY TO KL CONTROL

Stronger recorded settings shift the operating point. Across the recorded setting range 0.01–0.50, CHAIR decreases from 6.3 to 4.6 and Hal from 44.3 to 29.4, while Cover decreases from 66.8

Table 5: **Sensitivity to the recorded KL-control setting.** All quality metrics are percentages. Rates are observations from this series.

δ	CHAIR↓	Cover↑	Hal↓	Cog↓	Observed Tok/s↑
0.01	6.3	**66.8**	44.3	3.8	42.33
0.05	6.1	65.4	40.6	3.6	44.06
0.10	5.8	64.5	38.6	3.3	45.91
0.50	**4.6**	60.7	**29.4**	**2.2**	45.92

to 60.7 (Table 5). The setting 0.10 lies between these endpoints, with CHAIR 5.8 and Cover 64.5. The result supports adjustable hallucination–coverage behavior.

C.2 Representation Depth

Table 6: **Representation-depth sensitivity on Qwen2.5-VL-3B.** All quality metrics are percentages. Rates are observations from this series.

Depth	CHAIR↓	Cover↑	Hal↓	Cog↓	Observed Tok/s↑
$L/4$	6.1	64.2	39.6	3.1	44.29
$L/2$	6.1	64.2	40.4	3.2	43.95
$3L/4$	**5.8**	64.3	39.5	**3.0**	43.22
Final	**5.8**	**64.5**	**38.6**	3.3	43.15

The final layer improves Hal. On the reported Qwen depth series, the final layer attains the lowest Hal (38.6), highest Cover (64.5), and a tied-lowest CHAIR (5.8) with $3L/4$ (Table 6). However, $3L/4$ has the lowest Cog (3.0 versus 3.3 at the final layer). Cover varies by only 0.3 points across tested depths, and recorded throughput ranges from 43.15 to 44.29 Tok/s. These observations support the final layer for the chosen Hal/coverage operating point.

D Generation Budget and Response Length

Table 7: **Generation-cap analysis.** Quality values, mean lengths, and cap-hit counts is reported. $\widehat{v}$ is an approximate rate derived from the reported means. Cog uses the lower-is-better AMBER convention.

Cap	CHAIR↓	Cover↑	Hal↓	Cog↓	Avg. Tok.	Cap hits	Derived $\widehat{v}$ (Tok/s)
128	4.7	56.1	28.7	1.8	114.15	547	20.35
256	5.6	63.8	37.3	3.1	160.64	271	47.67
512	5.5	64.2	38.3	3.1	181.77	28	21.16
1024	5.8	64.5	38.6	3.3	188.49	0	45.86

A shorter cap can improve hallucination scores while losing coverage. At cap 128, CHAIR and Hal are 4.7 and 28.7, compared with 5.8 and 38.6 at cap 1,024 (Table 7). However, Cover is 56.1 rather than 64.5, and 547 responses reach the shorter cap.

Coverage changes little beyond 256 in this series. From cap 256 to 512, mean length increases by 21.13 tokens, from 160.64 to 181.77, but Cover increases by only 0.4 points, from 63.8 to 64.2. CHAIR changes from 5.6 to 5.5, Hal from 37.3 to 38.3, and Cog remains 3.1. Increasing the cap further to 1,024 adds 6.72 tokens on average and 0.3 points of Cover, with slightly higher CHAIR, Hal, and Cog. The cap-hit count falls from 271 to 28 to zero across these three settings.

Timing conversion. For each row we compute

$$\widehat{v} = \frac{\overline{L}}{\overline{t}}. \tag{16}$$

This equals $\sum_i L_i / \sum_i t_i$ only when both means describe the same samples and timing interval. The displayed rates are approximate because the inputs were rounded.

E Observed Throughput and Runtime Accounting

Table 8: **Response length and observed throughput on MSCOCO-500.** Lengths come from the latest 512-setting quality table.

	Qwen2.5-VL-3B		Llama-3.2-Vision-11B	
Method	Avg. Tok.	Observed Tok/s↑	Avg. Tok.	Observed Tok/s↑
Regular	409.5	32.38	199.54	31.69
VCD	453.4	15.24	144.48	15.11
OPERA	321.7	9.54	167.43	14.59
HALC	329.1	14.9	209.13	4.0
CMAC	194.0	7.96	144.91	12.98
ICT	412.5	27.54	202.87	26.93
MaskCD	377.2	14.2	167.44	13.24
POTION	355.1	23.55	172.45	23.12

Table 9: **Response length and observed throughput on AMBER-1,004.** Lengths come from the latest 512-setting quality table.

	Qwen2.5-VL-3B		Llama-3.2-Vision-11B	
Method	Avg. Tok.	Observed Tok/s↑	Avg. Tok.	Observed Tok/s↑
Regular	213.93	30.47	186.58	31.8
VCD	394.63	5.12	154.27	15.45
OPERA	207.30	6.72	163.12	14.19
HALC	234.45	5.74	193.95	4.14
CMAC	98.36	7.32	147.11	13.23
ICT	214.53	26.05	189.03	27.58
MaskCD	206.19	12.42	178.46	13.03
POTION	181.77	21.81	185.75	21.47

Reported rates retain their measurement provenance. Tables 8 and 9 retain all supplied rates, including missing values. Most baseline rates originate from 512-cap source runs with $K = 5$; the Llama POTION rates are 23.12 Tok/s on MSCOCO and 21.47 Tok/s on AMBER.

The observed Qwen POTION rates are 23.55 Tok/s on MSCOCO and 21.81 on AMBER, below the corresponding regular and ICT rates.

Timing measurement. For a fixed model and protocol, aggregate generation throughput is

$$v_{\text{gen}} = \frac{\sum_i L_i}{\sum_i t_i^{\text{gen}}}, \tag{17}$$

where L_i counts emitted response tokens, excluding discarded candidate branches.

F Generation-Budget Robustness with CMAC

Motivation Hallucination metrics can depend on response length: shorter outputs have fewer opportunities to introduce unsupported object mentions. We therefore compare regular, CMAC, and POTION under common maximum generation budgets of 256 and 512 tokens to examine whether the relative behavior of POTION persists across generation caps. CMAC describes a separate GPC+EOS variant; its stopping behavior should not be attributed to every CMAC implementation (Li et al., 2026).

Protocol. For each model–dataset pair, all methods use the same images, prompt, answer-parsing procedure, evaluator revision, and maximum generation cap. Method-specific hyperparameters are kept fixed across the two caps. We report the realized average response length together with the quality metrics so that remaining differences in generation length are explicit.

Table 10: **Generation-budget robustness on MSCOCO-500.** Cap denotes the maximum number of generated tokens. Avg. Tok. is the realized average response length.

Cap	Method	$\text{CHAIR}_S \downarrow$	$\text{CHAIR}_I \downarrow$	Recall ↑	Avg. Tok.
Qwen2.5-VL-3B					
256	Regular	37.00	8.80	63.43	247.36
	CMAC	27.40	**5.44**	**68.09**	179.35
	POTION	**26.40**	6.10	63.43	238.12
512	Regular	52.60	11.00	66.80	409.50
	CMAC	**28.40**	**5.60**	**70.70**	194.00
	POTION	31.40	**5.60**	68.16	355.10
Llama-3.2-Vision-11B					
256	Regular	30.80	6.86	67.26	182.50
	CMAC	25.40	5.77	**69.31**	142.70
	POTION	**21.60**	**4.85**	65.15	161.38
512	Regular	34.40	7.75	67.46	199.54
	CMAC	24.20	5.30	**69.37**	144.91
	POTION	**17.40**	**3.09**	66.37	172.45

Generation-budget robustness on MSCOCO. Table 10 shows that the effect of POTION is architecture-dependent. On Llama-3.2-Vision-11B, POTION achieves the lowest CHAIR_S and CHAIR_I at both budgets: $21.6/4.85$ at 256 tokens and $17.4/3.09$ at 512 tokens. CMAC retains higher object recall at both operating points, indicating that POTION's stronger hallucination suppression comes with a small recall trade-off. On Qwen2.5-VL-3B, the comparison is more mixed: POTION achieves lower CHAIR_S than CMAC at 256 tokens (26.4 vs. 27.4), whereas CMAC performs better at 512 tokens (28.4 vs. 31.4); the two methods tie on CHAIR_I at the 512-token setting. Thus, POTION's advantage is consistent across generation budgets on Llama, while the Qwen results exhibit a stronger trade-off.

Generation-budget robustness on AMBER. Table 11 shows a more consistent pattern on AMBER: POTION obtains the lowest CHAIR score in all four model–budget settings. On Qwen2.5-VL-3B, POTION also maintains higher object coverage than CMAC at both caps, although CMAC achieves lower Hal and Cog. On Llama-3.2-Vision-11B, POTION yields lower CHAIR and Hal than both regular and CMAC at 256 and 512 tokens, while CMAC retains higher coverage. The Cog comparison is mixed, with CMAC slightly better at 256 tokens (2.9 vs. 3.0) and POTION slightly better at 512 tokens (3.0 vs. 3.1).

G Statistical and Qualitative Validation

Qualitative comparisons. Figures 5 and 6 show selected examples of hallucination suppression and coverage recovery. POT-only and POTION differ only in the KL constraint. Red underlining marks unsupported mentions, and blue bold text marks content recovered by POTION. Rows are verbatim excerpts.

H Limitations

POTION has two main limitations. First, intervention is restricted to object mentions identified among the top-K next-token candidates using a predefined object vocabulary and token-suffix matching. Mentions outside the top-K, absent from the vocabulary, or missed by this procedure may remain unaffected. Second, lower object hallucination does not uniformly improve all generation metrics. The hallucination–coverage trade-off depends on the model: stronger transport-based suppression can remove valid object mentions, and KL control mitigates but does not eliminate this coverage loss.

Table 11: **Generation-budget robustness on AMBER-1,004.** Cap denotes the maximum number of generated tokens. Avg. Tok. is the realized average response length.

Cap	Method	CHAIR ↓	Cover ↑	Hal ↓	Cog ↓	Avg. Tok.
Qwen2.5-VL-3B						
256	Regular	7.3	**68.2**	46.0	4.7	169.80
	CMAC	6.1	62.2	**32.7**	**2.4**	98.40
	POTION	**5.4**	63.5	37.2	2.9	158.67
512	Regular	8.0	**69.8**	49.0	5.3	213.93
	CMAC	5.8	61.8	**32.2**	**2.0**	98.36
	POTION	**5.5**	64.2	38.3	3.1	181.77
Llama-3.2-Vision-11B						
256	Regular	6.0	67.1	47.6	3.5	182.95
	CMAC	7.3	**68.6**	50.4	**2.9**	143.89
	POTION	**5.2**	63.9	**39.6**	3.0	183.78
512	Regular	5.7	67.2	44.9	3.4	186.58
	CMAC	7.3	**68.4**	51.7	3.1	147.11
	POTION	**5.1**	64.2	**40.1**	**3.0**	185.75

Table 12: **Robustness on the AMBER generative benchmark under stochastic decoding across three pre-declared seeds.** Values are mean ± sample standard deviation. Lower is better for CHAIR, Hal, and Cog; higher is better for Cover.

Method	CHAIR↓	Cover↑	Hal↓	Cog↓
Regular	8.0 ± 0.06	$\mathbf{69.8 \pm 0.69}$	49.0 ± 1.54	5.3 ± 0.15
POTION (top-K=5)	$\mathbf{5.5 \pm 0.26}$	64.2 ± 0.42	$\mathbf{38.3 \pm 0.97}$	$\mathbf{3.1 \pm 0.30}$

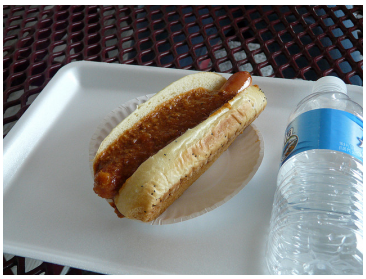
 **(a)** *no person is visible behind the tray.*	**Regular**	In the background, part of another person's legs can be seen, suggesting that the setting might be an outdoor event or gathering. ... The legs appear to be wearing dark pants and light-colored shoes, indicating casual attire suitable for a relaxed environment.	h/m: 3/19 c/a: 6/8 tok/cap: 226/N
	POT-only	The background of the image shows a metal mesh or table surface, which suggests that the setting might be an outdoor or casual dining environment, such as a picnic or a casual meal at a restaurant with an outdoor seating area.	h/m: 1/16 c/a: 4/8 tok/cap: 338/N
	POTION	The bun and chili are placed on a white paper plate, which is resting on a metal table with a grid pattern. ... The overall setting suggests a casual outdoor dining experience, possibly at a picnic or a casual gathering.	h/m: 1/12 c/a: 6/8 tok/cap: 171/N
 **(b)** *invented bicycle components.*	**Regular**	The bicycle has a chain and gear system, which is evident from the chainring on the front (crank) and the rear derailleur mechanism. ... The brakes are not clearly visible in the image, but given the context, they are likely to be hydraulic disc brakes, which are common on modern road bikes for their efficiency and effectiveness.	h/m: 8/29 c/a: 3/4 tok/cap: 512/Y
	POT-only	The frame of the bike is white, and it has a classic design with a straight handlebar and a single gear on the rear wheel. ... The bike has a chain and a gear system, indicating it is a manual transmission.	h/m: 0/12 c/a: 3/4 tok/cap: 207/N
	POTION	The frame of the bicycle is white, and it has a classic design with a straight handlebar and a drop handlebar. ... The bicycle has a chain and a gear system, indicating it is a standard road or city bike.	h/m: 3/20 c/a: 3/4 tok/cap: 365/N
 **(c)** *no sky is visible in the aerial crop.*	**Regular**	The image depicts a winding road cutting through a vast, snowy landscape. ... The sky above is clear, indicating that the photograph was likely taken during the day when visibility is good.	h/m: 5/16 c/a: 2/3 tok/cap: 318/N
	POT-only	The image depicts a winding, two-lane road cutting through a vast, snowy landscape. ... The overall scene is devoid of any human activity or structures, giving it a sense of isolation and tranquility.	h/m: 1/6 c/a: 2/3 tok/cap: 131/N
	POTION	The image depicts a winding road cutting through a vast, snowy landscape. ... The overall scene suggests a remote, possibly rural or wilderness area, where the road serves as the only visible connection to the outside world.	h/m: 1/9 c/a: 2/3 tok/cap: 134/N

Figure 5: Selected AMBER examples of hallucination suppression. Red underlining marks unsupported mentions; blue bold text marks content recovered by POTION. Metrics use full responses: h/m = hallucinated/evaluated mentions, c/a = covered/annotated objects, and tok/cap = generated tokens/512-token cap hit (Y/N).

 **(a)** *the chair is occupied, not empty.*	**Regular**	The foreground features a man sitting on a white metal chair with his back to the camera. ... In front of him on the grassy area, there is a small, fluffy dog that seems to be looking at the frisbee lying on the ground.	h/m: 0/27 c/a: 8/13 tok/cap: 434/N
	POT-only	The foreground features a grassy area with a white-framed metal lawn chair positioned on the left side. ... The back of the chair is visible, and it appears to be empty.	h/m: 0/19 c/a: 5/13 tok/cap: 355/N
	POTION	The foreground features a grassy area with **a man** sitting on a white metal chair. ... In front of the man, on the grass, there is a small, fluffy dog, which seems to be a Bichon Frise or a similar breed.	h/m: 0/36 c/a: 7/13 tok/cap: 401/N
 **(b)** *the camera is recovered.*	**Regular**	The image depicts a person lying on the grass, holding a camera in their hands.	h/m: 0/25 c/a: 4/4 tok/cap: 489/N
	POT-only	In the foreground, there is a young person lying on the grass, holding a black smartphone in their right hand.	h/m: 0/5 c/a: 3/4 tok/cap: 137/N
	POTION	In the foreground, there is a person lying on the grass, holding **a camera** up to their face, seemingly taking a photograph.	h/m: 0/18 c/a: 4/4 tok/cap: 309/N
 **(c)** *the wall and fence are recovered.*	**Regular**	In the background, there are buildings with brick facades. ... These buildings appear to be residential or commercial structures, given their architectural style and the presence of windows.	h/m: 1/26 c/a: 6/10 tok/cap: 396/N
	POT-only	The background of the image shows a residential area with several houses. ... The houses are two-story brick structures with pitched roofs, typical of urban residential areas.	h/m: 0/17 c/a: 4/10 tok/cap: 427/N
	POTION	The houses are separated from the street by a low **brick wall**, which has some greenery growing on it, including a small tree and some bushes. ... The wall is topped with a **black metal fence**, adding a touch of security to the residential area.	h/m: 0/31 c/a: 8/10 tok/cap: 456/N
 **(d)** *the visible saddle is recovered.*	**Regular**	The attire and equipment (such as the saddle and bridle) are indicative of this lifestyle.	h/m: 1/33 c/a: 9/10 tok/cap: 511/N
	POT-only	The horse is equipped with a bridle and a black leather reins, indicating it is being ridden.	h/m: 0/12 c/a: 5/10 tok/cap: 412/N
	POTION	He is mounted on the white horse, which is equipped with **a saddle**, reins, and a bridle.	h/m: 0/33 c/a: 9/10 tok/cap: 360/N

Figure 6: Selected AMBER examples of coverage recovery with the KL constraint. Red underlining marks unsupported mentions; blue bold text marks content recovered by POTION. Metrics use full responses: h/m = hallucinated/evaluated mentions, c/a = covered/annotated objects, and tok/cap = generated tokens/512-token cap hit (Y/N).