MARKOV PROPOSALS: ITERATING VISUAL ATTENTION FOR VISION–LANGUAGE ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite rapid progress, Large Vision-Language Models (LVLMs) still generate descriptions unsupported by the visual input. One possible contributor is insufficient use of visual evidence during generation, involving both the amount of attention allocated to image tokens and its distribution across image regions. We propose a training-free method that guides the language decoder's attention using a spatial prior derived from the vision encoder. We interpret the encoder's self-attention as a weighted directed graph over image patches as nodes and propagate attention through this graph as a discrete-time Markov chain. The resulting *Markov proposal* incorporates indirect relationships among nodes to expose spatial structure beyond direct attention connections. We blend this image-dependent prior into the decoder's image-token attention at selected layers and normalize the full attention row. This intervention increases total image attention while guiding its spatial allocation through a combination of the shared proposal and the decoder's token-dependent attention. Our empirical analysis shows that a frozen LVLM can benefit from increased attention to annotated object regions. Controlled comparisons further show that Markov steering improves supported-token margins beyond proportional image-attention scaling, supporting a benefit from proposal-based redistribution of existing visual information. The method introduces no additional learned parameters and requires no auxiliary data or training. Experiments on LLaVA-1.5-7B and MiniGPT-4 demonstrate improvements across CHAIR, SPICE, BLEU, and POPE.

1 INTRODUCTION

Large Vision-Language Models (LVLMs) (Liu et al., 2024a; 2023; Zhu et al., 2023) perform strongly on image captioning and visual question answering, yet still generate objects, attributes, and relations unsupported by the image (Li et al., 2023b; Rohrbach et al., 2018). Recent analyses associate such hallucinations with insufficient visual attention and changes in visual representations across decoder layers (Yang et al., 2025; Jiang et al., 2025). These findings motivate making better use of visual evidence already available within the model.

Existing approaches mitigate hallucination through modified decoding (Chen et al., 2024; Huang et al., 2024; Leng et al., 2024), latent-feature steering (Liu et al., 2024b), or additional grounding and feature-alignment training (Shao et al., 2024; Wang et al., 2025; Yoon et al., 2025). We investigate a complementary approach: directly guiding decoder attention with an image-dependent spatial prior derived from the vision encoder.

In this paper, we guide the language decoder's attention with an image-dependent spatial prior derived from the vision encoder. Our premise is that pretrained LVLMs already contain useful correspondences between visual representations and vocabulary tokens, but generation may not fully exploit them. We represent the encoder's patch attention as a directed graph and iteratively propagate a distribution over its image-patch nodes as a discrete-time Markov chain. We call the resulting spatial prior the *Markov proposal*. This propagation aggregates indirect patch relationships, whose ability to reveal spatial grouping has been demonstrated in prior work on multi-bounce attention (Erel et al., 2025).

The proposal is computed once per image and blended into the decoder's token-dependent image attention, steering both total image attention and its spatial allocation. Controlled experiments show

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.

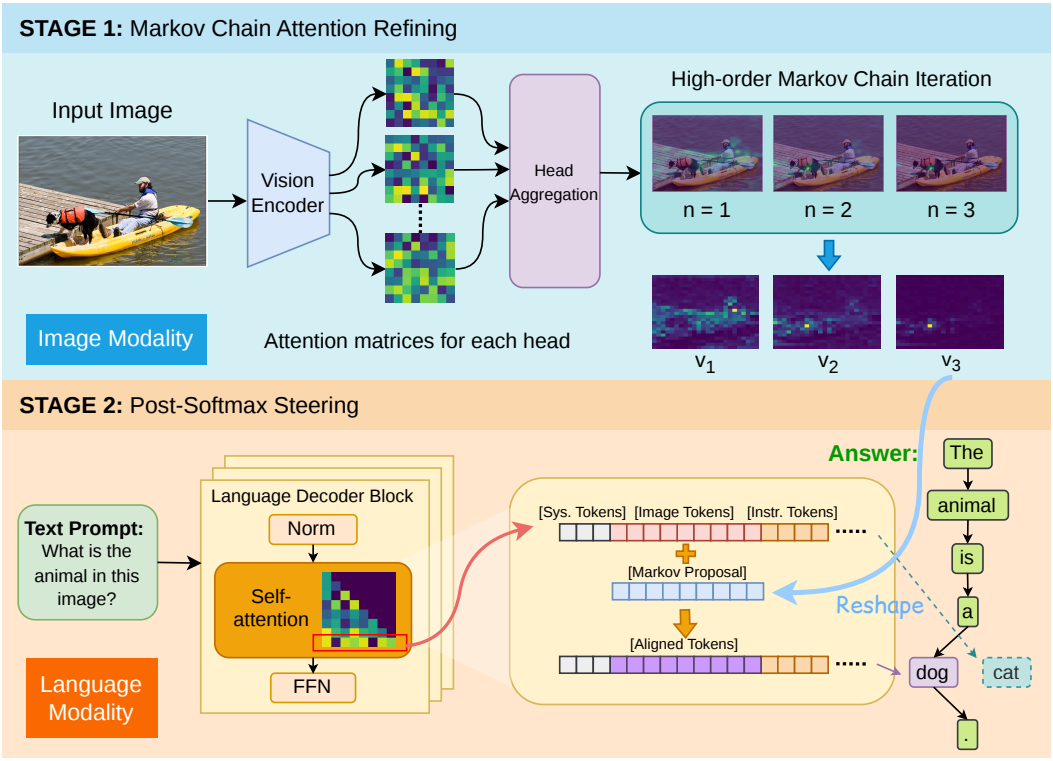


Figure 1: Overview of the two-stage framework. Stage 1 (**Markov Proposal Construction**) aggregates vision-encoder attention and propagates a distribution over image patches to obtain an image-dependent spatial prior. Stage 2 (**Post-Softmax Steering**) blends the proposal into the decoder's image-token attention and normalizes the full attention row. Section 3 describes the two operations.

that directing attention toward annotated object regions improves supported-word preference on average, and that Markov steering yields additional supported-token margin gains beyond proportional image-attention scaling. The method requires no additional training, learned parameters, or auxiliary data.

Our contributions are threefold:

- We introduce *Markov Proposal*, a training-free attention-steering method that propagates the vision encoder's self-attention as a discrete-time Markov chain to construct an image-dependent prior, which is then blended into decoder attention during generation.
- We provide controlled empirical evidence that pretrained visual representations can support improved object-word preference, and that Markov steering yields additional margin gains beyond proportionally increasing total image attention.
- We demonstrate reduced object hallucination on LLaVA-1.5-7B (Liu et al., 2024a) and MiniGPT-4 (Zhu et al., 2023), with additional improvements in SPICE, BLEU, and POPE evaluations.

2 Related Works

Hallucination in LVLMs. Object hallucination has been addressed through decoding-time corrections such as contrastive decoding and beam-level interventions (Chen et al., 2024; Huang et al., 2024; Leng et al., 2024), post-hoc revision (Zhou et al., 2024), and mechanistic analyses that locate the primary drivers of hallucination in attention heads at middle-to-deep layers (Jiang et al., 2025; Kang et al., 2025; Yang et al., 2025). These analyses motivate the layer-localized form of our intervention. The closest prior work to our regime is VTI (Liu et al., 2024b), which applies PCA-derived feature shifts from $\sim$50 paired hallucinated/non-hallucinated captions uniformly across layers. Our method differs in that its signal is derived from the input image alone via the vision encoder's own attention

dynamics, it operates on attention distributions rather than feature representations, and it is applied at selected decoder layers.

Attention as a Markov chain and graph propagation. Several works interpret post-softmax attention as a stochastic transition operator and exploit iterative propagation for spatial grouping, formalizing this view for vision transformers (Erel et al., 2025), applying it to diffusion self-attention for unsupervised segmentation (Karmann & Urfalioglu, 2025), and proving that self-self attention with sub-unit-Lipschitz projections acts as a contractive clustering operation (Bousselham et al., 2024). Related precedents include random-walk spatial proposals over CNN features (Zhu et al., 2017), cross-layer attention composition (Abnar & Zuidema, 2020), and a Markov-chain view of autoregressive generation over vocabulary tokens (Ildiz et al., 2024). To our knowledge, ours is the first work to route a propagated attention distribution into the language decoder of an LVLM as an attention prior, rather than reading it out for segmentation.

Visual instruction tuning and grounded generation. Stronger visual grounding has also been pursued through grounded chain-of-thought training data (Wang et al., 2025) or alignment losses against frozen vision features (Yoon et al., 2025). These approaches require additional data and weight updates. Work extracting localization signals from CLIP self-attention (Bousselham et al., 2024; Zhou et al., 2022) motivates our use of an attention-derived spatial prior. We investigate its use for decoder attention steering during generation.

Attention intervention. Inference-time modification of attention weights has been used to steer language models (Li et al., 2023a; Zou et al., 2023) and to mitigate LVLM hallucination by suppressing offline-identified "hallucination heads" (Yang et al., 2025). We share the same architectural locus but differ in signal source. We derive a positive, per-image attention prior from a Markov process on the model's own visual attention, rather than relying on a fixed set of heads identified from labeled counterfactual data.

3 METHOD

Our method consists of two conceptual steps: (1) constructing a spatial prior over image patches as nodes from the vision encoder, and (2) injecting this prior into the decoder's attention during generation. Section 3.1 reviews the background knowledge. Section 3.2 casts self-attention as a graph and identifies its spatial structure. Section 3.3 designs an iterative propagation on this graph. Section 3.4 defines the Markov proposal and the steering operation that injects it into the language decoder. Algorithm 1 and Algorithm 2 showcase the two stages of our method.

Algorithm 1: Markov Proposal Generation

Input: Image $\mathbf{x}$, bounces n, CLIP layer $\ell^\star$
Extract CLIP attention $\{A_h^{(\ell^\star)}\}_{h=1}^{H}$ on $\mathbf{x}$;
$\bar{A}^{(\ell^\star)} \leftarrow \frac{1}{H}\sum_h A_h^{(\ell^\star)}$;
// uniform head aggregation
$j^\star \leftarrow \arg\max_j \sum_i \bar{A}^{(\ell^\star)}[i,j]$
// largest incoming attention
$\mathbf{v}_0 \leftarrow \mathbf{e}_{j^\star}$; // one-hot seed
$\mathbf{m} \leftarrow \mathbf{v}_0^\top (\bar{A}^{(\ell^\star)})^n$;
// propagate via Markov chain
Normalize $\mathbf{m}$;
Output: Markov Proposal $\mathbf{m}$

Algorithm 2: Attention Steering

Input: Markov Proposal $\mathbf{m}$, image-token index set $\mathcal{I}$, blend α
for *decoding step* t **do**
  for *head* h *in layer* ℓ **do**
    $A_{t,\mathcal{I}}^{(\ell,h)} \leftarrow (1-\alpha)A_{t,\mathcal{I}}^{(\ell,h)} + \alpha\,\mathbf{m}$;
    // post-softmax blend
    $A_{t,:}^{(\ell,h)} \leftarrow A_{t,:}^{(\ell,h)} / \sum_j A_{t,j}^{(\ell,h)}$;
    // renormalize the full row
  end
  Self-attention calculation;
end

3.1 PRELIMINARIES

Multi-head attention. Given an input $\mathbf{X} \in \mathbb{R}^{N\times d}$ of N tokens with d feature dimension, a single attention head with parameters $\mathbf{W}_q^h, \mathbf{W}_k^h \in \mathbb{R}^{d\times d}$ and $\mathbf{W}_v^h \in \mathbb{R}^{d\times d}$ computes

$$\mathbf{A}^h = \mathrm{softmax}\left(\frac{\mathbf{X}\mathbf{W}_q^h(\mathbf{X}\mathbf{W}_k^h)^\top}{\sqrt{d}}\right) \in \mathbb{R}^{N\times N}, \qquad \mathbf{O}^h = \mathbf{A}^h\mathbf{X}\mathbf{W}_v^h. \tag{1}$$

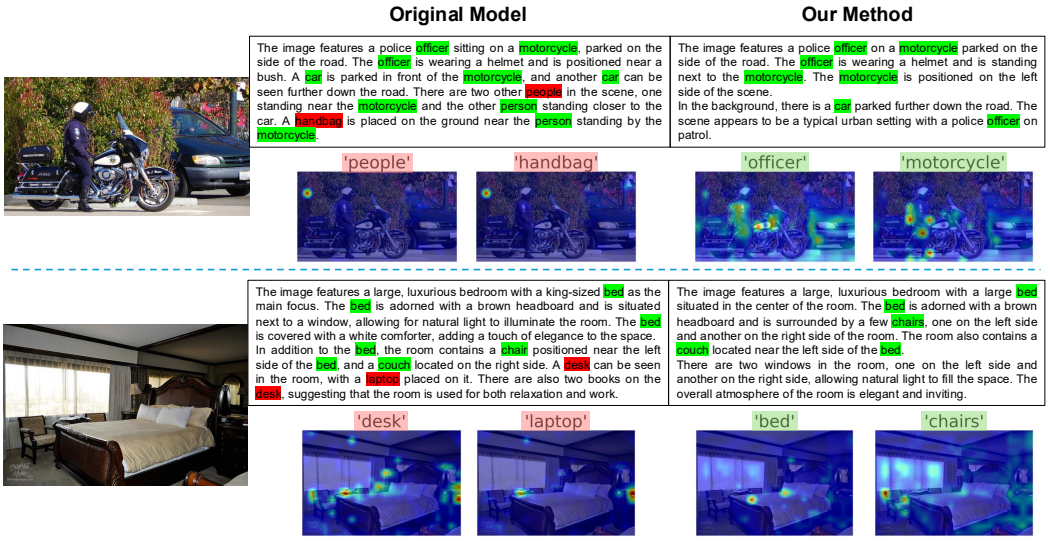


Figure 2: Qualitative comparison between our method and the original LLaVA-1.5. Words highlighted in green are the ground-truth words, and the ones in red are hallucinations. The smaller images below each caption show the attention paid by the hallucinated words and ground-truth words.

The row-wise softmax makes each row of $\mathbf{A}^h$ a probability distribution over the N tokens, so $\mathbf{A}^h$ is row-stochastic:

$$\sum_{j=1}^{N} \mathbf{A}^h[i,j] = 1, \qquad \mathbf{A}^h[i,j] \geq 0. \tag{2}$$

Discrete-time Markov chains. A discrete-time Markov chain (DTMC) over a finite state space $\mathcal{S} = \{1, \ldots, m\}$ is specified by a transition matrix $\mathbf{A} \in \mathbb{R}^{m \times m}$ with

$$\mathbf{A}[i,j] = \Pr(s_{t+1} = j \mid s_t = i), \qquad \sum_{j=1}^{m} \mathbf{A}[i,j] = 1, \qquad \mathbf{A}[i,j] \geq 0. \tag{3}$$

A distribution over states at time t is a row vector $\mathbf{v}_t^\top \in \Delta^{m-1}$, advanced by

$$\mathbf{v}_{t+1}^\top = \mathbf{v}_t^\top \mathbf{A}, \qquad \mathbf{v}_n^\top = \mathbf{v}_0^\top \mathbf{A}^n. \tag{4}$$

Under irreducibility and aperiodicity, $\mathbf{v}_n$ converges as $n \to \infty$ to a unique stationary distribution $\boldsymbol{\pi}^\top$ satisfying $\boldsymbol{\pi}^\top = \boldsymbol{\pi}^\top \mathbf{A}$. Any row-stochastic matrix admits this DTMC interpretation.

3.2 Self-Attention as a Graph over Tokens

We view a self-attention matrix $\mathbf{A} \in \mathbb{R}^{N \times N}$ as a weighted directed graph $\mathcal{G} = (V, E, \mathbf{A})$ over tokens:

$$V = \{1, \ldots, N\}, \qquad V_n \leftrightarrow \mathbf{X}[n,:], \qquad E[i,j] = \mathbf{A}[i,j]. \tag{5}$$

Each node represents an image token, and the edge weight $\mathbf{A}[i,j]$ specifies a Markov transition from node i to node j, corresponding to how strongly token i attends to token j.

We specialize to the self-attention of a ViT vision encoder, where the input image is divided into m non-overlapping patches and each token is the embedding of one patch. In this view, each node of $\mathcal{G}$ corresponds to an image patch, and each edge measures how strongly one region attends to another.

3.3 Iterative Propagation on the Attention Graph

To make the notation clearer, we note $\mathbf{P}$ as the row-normalized patch-attention matrix in the vision encoder and $\mathbf{A}$ as the attention map in the language decoder. Starting from a distribution $\mathbf{v}_0^\top$, we define the iterative Markov propagation by

$$\mathbf{v}_{n+1}^\top = \mathbf{v}_n^\top \mathbf{P}, \qquad \mathbf{v}_n^\top = \mathbf{v}_0^\top \mathbf{P}^n. \tag{6}$$

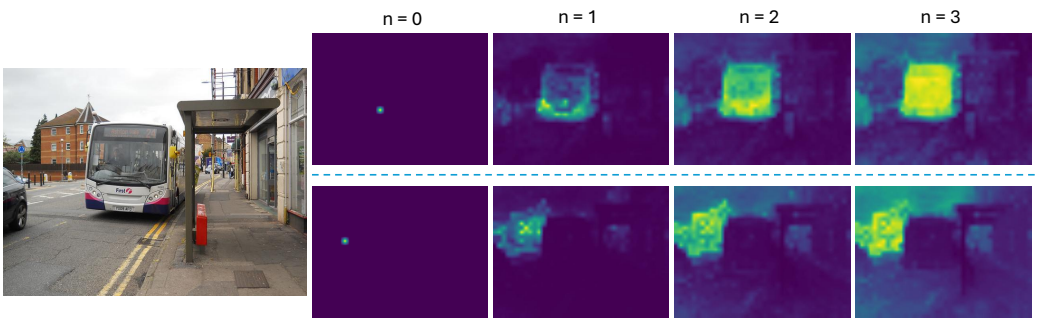


Figure 3: The visualization of how the Markov Proposals are formed with comparisons between different iterations.

Each step incorporates an additional level of indirect attention relationships. A node can therefore receive probability through intermediate nodes even when its direct connection to the seed is weak. This provides a mechanism for extending a local response across a connected image region.

Choice of initial distribution. We initialize propagation at the image token receiving the largest total incoming attention, where P_{ij} denotes the transition probability from patch i to patch j:

$$j^\star = \arg\max_j \sum_i P_{ij}, \qquad \mathbf{v}_0 = \mathbf{e}_{j^\star}. \tag{7}$$

where $\mathbf{e}_{j^\star}$ denotes the one-hot vector whose $j^\star$-th entry is 1 and all other entries are 0. Thus, propagation starts with all probability mass concentrated at the selected image patch.

3.4 The Markov Proposal and Decoder Attention Steering

The Markov proposal. We define the proposal as

$$\mathbf{p}_{\text{prop}}^\top = \mathbf{v}_0^\top \mathbf{P}^n. \tag{8}$$

Because $\mathbf{P}$ is row-stochastic, $\mathbf{p}_{\text{prop}}$ is a probability distribution over the image patches as nodes. The proposal supplies a spatial prior derived from the encoder's attention graph. It contains no text input and is not independently conditioned on the question or the next generated token.

Steering the decoder's attention. Let $\mathbf{A}[i,:]^{(\ell,h)}$ denote the decoder's post-softmax attention row at position i, layer ℓ, and head h, and let $\mathcal{I}$ index image-token columns. For each steered row, we first modify the image block and then normalize over all unmasked key positions:

$$\widehat{\mathbf{A}}[i,j]^{(\ell,h)} = \begin{cases} (1-\alpha)\mathbf{A}[i,j]^{(\ell,h)} + \alpha \mathbf{p}_{\text{prop},j}, & j \in \mathcal{I}, \\ \mathbf{A}[i,j]^{(\ell,h)}, & j \notin \mathcal{I}, \end{cases} \qquad \widetilde{\mathbf{A}}[i,j]^{(\ell,h)} = \frac{\widehat{\mathbf{A}}[i,j]^{(\ell,h)}}{\sum_k \widehat{\mathbf{A}}[i,k]^{(\ell,h)}}, \tag{9}$$

where $\alpha \in [0,1]$ is the steering coefficient and proposal entries are indexed by their corresponding image-token positions. The intervention can change both total image attention and its distribution across image positions. Appendix 7.1.2 specifies these changes and the corresponding controls, while Section 4 evaluates their effects on supported-token preference. Figure 3 illustrates two different proposals at different propagation steps.

4 Analysis

We investigate whether Markov steering improves the decoder's use of existing visual information. Our analysis tests an empirical proposition about supported-word preference, motivated by the premise that a pretrained decoder can benefit from stronger use of object-associated visual representations.

Comparison	Mean margin advantage	95% CI
Annotated region vs. baseline	0.1057	[0.0845, 0.1277]
Annotated region vs. control region	0.0782	[0.0561, 0.1011]
Markov vs. baseline	0.3791	[0.3113, 0.4472]
Markov vs. mass-only	0.1773	[0.1353, 0.2199]

Table 1: Paired margin comparisons on 300 COCO images with frozen LLaVA-1.5-7B. All comparisons use the full configured head/layer intervention at strength 0.3. Annotated-region and control-region steering preserve each edited row's image attention mass. Markov comparisons use the recorded normalized column-sum seed, blending, and full-row normalization.

Prediction margin. Fix an image and a text prefix. Let w^+ be a visually supported candidate token and w^- an alternative candidate. For intervention condition s, let z_w^s denote the final vocabulary logit of token w, and define

$$D^s = z^s_{w^+} - z^s_{w^-}, \qquad \Delta D^s = D^s - D^0, \tag{10}$$

where $s = 0$ denotes the unmodified decoder. A positive ΔD^s indicates increased preference for the supported token relative to the alternative. Let $s = \mathrm{M}$ denote Markov steering and $s = \mathrm{mass}$ a control that proportionally scales image attention using the same local image-mass increase rule, preserving relative weights within the image block. All comparisons hold the image, prefix, candidate tokens, and model parameters fixed. Expectations below are over image–candidate pairs under the evaluation protocol.

Proposition 4.1 (Empirical benefit of Markov steering)**.** *In the evaluated setting, Markov steering improves the average supported-token margin over both the unmodified decoder and proportional image-attention scaling:*

$$\mathbb{E}[\Delta D^{\mathrm{M}}] > 0, \qquad \mathbb{E}[D^{\mathrm{M}} - D^{\mathrm{mass}}] > 0. \tag{11}$$

Assumption 4.2 (usable visual grounding)**.** *The decoder must be able to translate object-associated visual information into a preference for the corresponding word.*

We test this assumption by redistributing attention toward an annotated object region R or an equal-area control region C outside that object, preserving each edited row's total image attention. Let D^R and D^C denote the resulting margins. The premise predicts

$$\mathbb{E}[D^R - D^0] > 0, \qquad \mathbb{E}[D^R - D^C] > 0. \tag{12}$$

These comparisons test whether the frozen decoder can use the object's visual information when its contribution is increased. The annotation supplies a diagnostic intervention, and the Markov method itself uses only the image-derived proposal.

When steering favors a visually supported continuation. We evaluate the full configured intervention on 300 COCO images using frozen LLaVA-1.5-7B and fixed prefixes, with intervention strength 0.3. The ground-truth annotated-region steering increases the mean margin by 0.1057 over baseline and by 0.0782 over the outside-region control, supporting the visual-grounding premise. Markov steering yields a mean ΔD^{M} of 0.3791. Its mean advantage over proportional image-attention scaling is 0.1773, supporting a benefit from proposal-based redistribution beyond the tested image-mass increase. The paired confidence intervals for all four contrasts are positive, providing empirical support for Proposition 4.1. The results are shown in Table 6. Appendix 7.1 gives the detailed configuration, control definitions, and numerical results. These findings support the decoder's ability to use existing visual information and the usefulness of the tested Markov intervention.

5 Experiments

5.1 Experiment Setup

Models. We evaluate on two LVLMs: LLaVA-1.5-7B (Liu et al., 2024a) and MiniGPT-4 (Zhu et al., 2023). The two models differ in both vision encoder and visual-language connector design, allowing us to test whether the Markov proposal generalizes beyond the LLaVA architecture from which it is most directly motivated.

Table 2: Hallucination rates in terms of CHAIR_S and CHAIR_I on COCO and Nocaps (Out-of-Domain) image captioning tasks

Methods	COCO				Nocaps (Out-of-Domain)			
	LLaVA-7B		MiniGPT-4		LLaVA-7B		MiniGPT-4	
	$\text{CHAIR}_{\text{S}\downarrow}$	$\text{CHAIR}_{\text{I}\downarrow}$	$\text{CHAIR}_{\text{S}\downarrow}$	$\text{CHAIR}_{\text{I}\downarrow}$	$\text{CHAIR}_{\text{S}\downarrow}$	$\text{CHAIR}_{\text{I}\downarrow}$	$\text{CHAIR}_{\text{S}\downarrow}$	$\text{CHAIR}_{\text{I}\downarrow}$
Greedy	51.8	13.3	40.6	13.7	43.2	14.3	57.4	20.0
DoLA (Chuang et al., 2024)	53.8	13.9	41.0	13.8	42.0	13.7	57.2	20.4
OPERA (Huang et al., 2024)	50.2	14.5	<u>35.2</u>	12.8	44.2	14.4	<u>46.2</u>	<u>16.2</u>
VCD (Leng et al., 2024)	55.4	15.7	38.8	14.8	43.6	14.4	48.2	17.5
LURE (Zhou et al., 2024)	51.2	13.4	46.4	14.2	41.8	14.4	55.8	19.6
HALC (Chen et al., 2024)	50.2	12.4	36.4	11.8	40.2	12.2	53.0	18.0
AD-HH (Yang et al., 2025)	<u>29.6</u>	<u>8.0</u>	<u>35.2</u>	<u>11.7</u>	<u>35.6</u>	**9.4**	46.8	<u>16.2</u>
Ours	**20.2**	**5.3**	**28.8**	**11.4**	**28.2**	<u>9.7</u>	**42.4**	**16.1**

Datasets. For object hallucination, we evaluate on CHAIR (Rohrbach et al., 2018) using two image sources: a 500-image subset of MSCOCO (Lin et al., 2014) (in-domain) and a 500-image subset of the Nocaps out-of-domain split (Agrawal et al., 2019), which contains object categories underrepresented in the COCO training distribution and serves as a test of generalization. For each image, we generate a caption with the prompt "Please describe this image in detail." We use standard CHAIR for COCO and a Nocaps-adapted version with dataset-specific vocabulary and ground truth from labels and reference captions.

Evaluation metrics. CHAIR measures object hallucination at two granularities,

$$\text{CHAIR}_S = \frac{|\text{captions with hallucinated objects}|}{|\text{all captions}|}, \qquad \text{CHAIR}_I = \frac{|\text{hallucinated objects}|}{|\text{all mentioned objects}|}, \quad (13)$$

with lower being better for both. CHAIR_S captures sentence-level hallucination (does the caption contain any hallucinated object?), while CHAIR_I captures instance-level hallucination (what fraction of mentioned objects are hallucinated?). The two metrics together characterize both the prevalence and the density of hallucination.

Baselines. We compare against the unmodified models under greedy decoding (Greedy) and six baseline methods: DoLA (Chuang et al., 2024), which contrasts logits across decoder layers; VCD (Leng et al., 2024), which contrasts logits between the original and a noise-corrupted image; OPERA (Huang et al., 2024), which penalizes over-trust on recent tokens during beam search; HALC (Chen et al., 2024), which uses focal-contrast decoding over object regions; LURE (Zhou et al., 2024), which post-hoc rewrites generated captions via a separately trained revisor; and AD-HH (Yang et al., 2025), which suppresses identified hallucination heads in the decoder. All baselines are run with the hyperparameters reported in their original papers on the same image subsets. More implementation details can be found in Section 7.2.

5.2 Main Results

Table 2 reports CHAIR_S and CHAIR_I on COCO (in-domain) and Nocaps (out-of-domain) for both LLaVA-1.5-7B and MiniGPT-4.

Markov steering reduces hallucination across models and domains. Table 2 shows that our method achieves the lowest hallucination rate in seven of eight model–dataset–metric combinations. On LLaVA-1.5-7B with COCO, CHAIR_S decreases from 51.8% to 20.2% and CHAIR_I from 13.3% to 5.3%, reductions of approximately 61% and 60% relative to greedy decoding. Improvements over greedy decoding also hold for MiniGPT-4 and the Nocaps out-of-domain split. AD-HH achieves a slightly lower CHAIR_I on LLaVA-1.5-7B with Nocaps (9.4% versus 9.7%), indicating that the advantage varies across evaluation settings. These results establish the method's effectiveness on the evaluated captioning tasks. Section 4 separately examines its effects on fixed-prefix word preference, and Figure 2 provides qualitative examples.

5.3 Caption Quality and Discriminative Evaluation

CHAIR alone does not establish whether hallucination reduction preserves descriptive content. We therefore complement it with object-presence and caption-quality evaluations.

Table 3: POPE on COCO (LLaVA-1.5-7B, 3000 questions per split).

Split	Greedy		+ Markov Steering	
	Acc.↑	F1↑	Acc.↑	F1↑
Random	88.27	87.06	**88.90**	**87.84**
Popular	87.10	85.95	**87.70**	**86.70**
Adversarial	**84.83**	83.88	84.77	**84.02**
Mean	86.73	85.63	**87.12**	**86.19**

Table 4: Caption quality on the 500-image COCO subset. SPICE and BLEU use the detailed-caption protocol of Section 5.1; CIDEr uses the matched-length protocol described in the text.

	SPICE↑	BLEU↑	CIDEr↑
Greedy	19.27	17.9	1.18
AD-HH (Yang et al., 2025)	20.23	17.8	1.13
Ours	**21.88**	**23.1**	**1.22**

Table 5: SPICE subcategory breakdown on the 500-image COCO subset (LLaVA-1.5-7B).

	Object↑	Attribute↑	Relation↑	Color↑	Cardinality↑	Size↑
Greedy	35.5	12.8	3.3	26.3	**31.5**	17.4
AD-HH (Yang et al., 2025)	38.1	13.2	4.3	26.1	29.2	**17.7**
Ours	**39.6**	**15.4**	**4.6**	**28.6**	28.1	15.4

Discriminative question answering. POPE (Li et al., 2023b) poses balanced Yes/No questions about object presence. Table 3 reports accuracy and F1 on all three COCO splits under the configuration of Section 5.1. Steering improves the mean by $+0.39$ accuracy and $+0.56$ F1. The gains are modest and we do not present them as a headline result, but their direction is informative. On the random split, recall rises from 78.93 to 80.20 at flat precision, and the yes-rate rises from 40.7% to 41.3%. These results provide complementary evidence of improved object-presence prediction. The saved POPE responses are single-word Yes/No answers. Steering at the final prefill position influences the first answer token and remains enabled during subsequent decoding.

Caption quality. Table 4 reports SPICE, BLEU, and CIDEr against COCO references. SPICE rises from 19.27 to 21.88 and BLEU from 17.9 to 23.1 over greedy decoding, both under the detailed-caption protocol of Section 5.1. CIDEr requires a separate protocol, which we state rather than fold silently into the same table. COCO references average roughly eleven words, and CIDEr-D's length penalty drives the $\sim$58-word detailed captions to near-zero for every method including greedy, making that comparison uninformative rather than favourable to any particular method. We therefore regenerate captions for the same 500 images with the prompt "Provide a one-sentence caption for the provided image." and score CIDEr in that matched-length regime. Generation is otherwise identical, with the same images, seed, greedy decoding, and steering configuration.

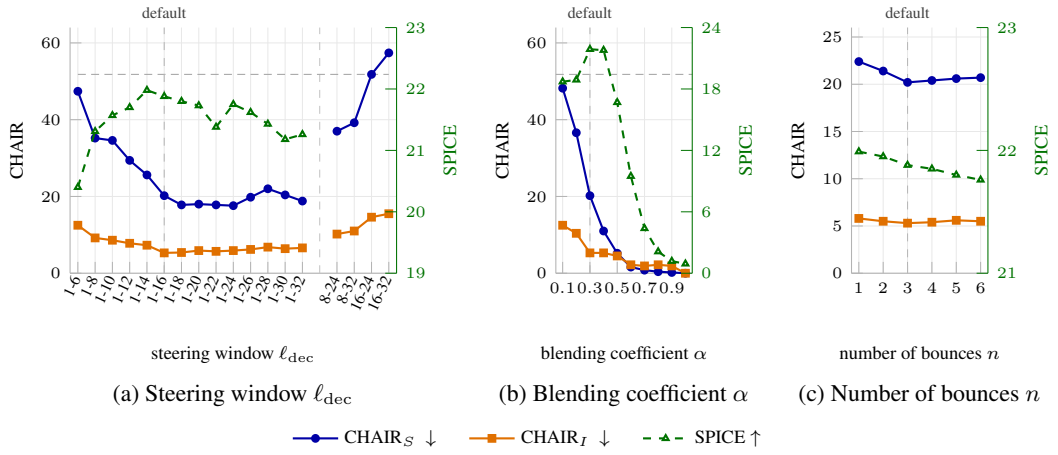
Beyond object mentions. Table 5 shows improvements in Object, Attribute, Relation, and Color scores over both greedy decoding and AD-HH. Relative to greedy decoding, Attribute increases from 12.8 to 15.4, Color from 26.3 to 28.6, and Relation from 3.3 to 4.6. However, Cardinality decreases from 31.5 to 28.1 and Size from 17.4 to 15.4. The gains therefore extend beyond object naming but are not uniform across semantic categories.

Coverage. Steering does reduce caption length, and we state the scope of our claim accordingly: the method preserves question-answering accuracy while trading descriptive coverage for precision. POPE shows that the reduction is not a loss of visual-question competence, and the SPICE subcategories show that it is not a uniform withdrawal from description, but it is a real trade and we do not claim otherwise.

5.4 Ablation Studies

We ablate the three main hyperparameters on LLaVA-1.5-7B using the COCO captioning setup of Section 5.1. Alongside CHAIR_S and CHAIR_I, we report SPICE to assess semantic caption quality. This joint evaluation helps identify settings that reduce hallucination without substantially sacrificing descriptive content. Unless otherwise stated, the remaining hyperparameters are fixed at their defaults ($\alpha = 0.3$, $n = 3$, steering range $\ell_{\text{dec}} \in \{1, \ldots, 16\}$, CLIP layer $\ell^\star = 18$, and uniform head aggregation).

Steering layer range. Figure 4a compares steering windows across the 32-layer Vicuna decoder. Extending a window anchored at layer 1 into the middle layers substantially improves CHAIR, while



(a) Steering window ℓ_{dec} (b) Blending coefficient α (c) Number of bounces n

Figure 4: Ablations on LLaVA-1.5-7B on COCO. Unless varied, the defaults are $\alpha = 0.3$, $n = 3$, $\ell_{\text{dec}} \in \{1, \ldots, 16\}$, CLIP layer $\ell^\star = 18$, and uniform head aggregation. CHAIR$_S$ and CHAIR$_I$ use the left axis (lower is better), and SPICE uses the green right axis (higher is better). Axis ranges differ across panels. Horizontal dashed lines in (a,b) mark the unmodified model's CHAIR$_S = 51.8$, and vertical default markers indicate the selected settings. (a) The categorical axis separates 14 windows starting at layer 1 from four later-starting windows. SPICE is shown only where available. (c) Across the bounce sweep, CHAIR$_S$, CHAIR$_I$, and SPICE span 2.2, 0.5, and 0.23 points, respectively.

windows excluding early layers perform less favorably on these metrics. Layers 1–16 achieve the lowest observed CHAIR$_I$ of 5.3 and a SPICE score of 21.88, close to the peak of 21.98 at layers 1–14. Extending to layers 1–18 lowers CHAIR$_S$ from 20.2 to 17.8, but yields CHAIR$_I = 5.4$ and SPICE $= 21.80$. We retain layers 1–16 as a default that prioritizes instance-level hallucination reduction while maintaining strong semantic caption quality.

Blending coefficient α. Figure 4b illustrates why CHAIR alone is insufficient for selecting steering strength. Although CHAIR$_S$ decreases as α increases, SPICE peaks at 21.88 for $\alpha = 0.3$ and drops sharply at larger coefficients, reaching 16.69 at $\alpha = 0.5$ and 0.92 at $\alpha = 1.0$. Thus, very low hallucination scores can coincide with poor semantic caption quality. We retain $\alpha = 0.3$, which achieves the highest observed SPICE together with CHAIR$_S = 20.2$ and CHAIR$_I = 5.3$. The neighboring setting $\alpha = 0.4$ offers lower CHAIR$_S$ (11.0) and the same CHAIR$_I$, with slightly lower SPICE (21.79), reflecting a trade-off between the metrics. The caption-quality and POPE evaluations in Section 5.3 provide complementary evidence for the default configuration.

Number of bounces n. Figure 4c compares one to six propagation steps with the remaining settings fixed. Increasing from one to three steps reduces CHAIR$_S$ from 22.4 to 20.2 and CHAIR$_I$ from 5.8 to 5.3, while SPICE decreases modestly from 21.99 to 21.88. Three steps achieve the lowest values of both CHAIR metrics. Further propagation provides no additional hallucination reduction and gradually lowers SPICE to 21.76 at six steps. These results support $n = 3$ as an empirical trade-off between hallucination reduction and semantic caption quality.

6 Conclusion

We presented a training-free method that constructs a Markov proposal from the vision encoder's attention graph and uses it to steer decoder attention during generation. Controlled diagnostics show that the frozen decoder can benefit from object-associated visual information and that the tested Markov intervention improves supported-token margins beyond proportional image-attention scaling. Captioning evaluations on LLaVA-1.5-7B and MiniGPT-4 show reduced object hallucination across COCO and Nocaps, with the best CHAIR scores in seven of eight model–dataset–metric combinations. Additional caption-quality and POPE results provide complementary evidence of the method's usefulness. The method introduces no learned parameters and requires no additional training or auxiliary data.

AI Disclosure Statement In this work, we used generative AI tools to check grammar, improve paper writing, and search for related papers. We have not used generative AI tools for generating synthetic data sets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, assisting with translation, cleaning and reformatting datasets, supporting qualitative and thematic data analysis, and interpreting results. All AI-assisted writing was reviewed and adjusted by the authors, and all referenced papers were checked for authenticity. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, 2020.

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2019.

Walid Bousselham, Felix Petersen, Vittorio Ferrari, and Hilde Kuehne. Grounding everything: Emerging localization properties in vision-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. Halc: Object hallucination reduction via adaptive focal-contrast decoding. In *Forty-first International Conference on Machine Learning*, 2024.

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 *The Twelfth International Conference on Learning Representations*, 2024.

Yotam Erel, Olaf Dünkel, Rishabh Dabral, Vladislav Golyanik, Christian Theobalt, and Amit Haim Bermano. Attention (as discrete-time markov) chains. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Muhammed Emrullah Ildiz, Yixiao Huang, Yingcong Li, Ankit Singh Rawat, and Samet Oymak. From self-attention to markov models: Unveiling the dynamics of generative transformers. In *Forty-first International Conference on Machine Learning*, 2024.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. Your large vision-language model only needs a few attention heads for visual grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Markus Karmann and Onay Urfalioglu. Repurposing stable diffusion attention for training-free unsupervised interactive segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 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*, 2023b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, 2014.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2024a.

Sheng Liu, Haotian Ye, Lei Xing, and James Zou. Reducing hallucinations in vision-language models via latent space steering. *arXiv preprint arXiv:2410.15778*, 2024b.

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*, 2018.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual CoT: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Jiacong Wang, Zijian Kang, Haochen Wang, Haiyong Jiang, Jiawen Li, Bohong Wu, Ya Wang, Jiao Ran, Xiao Liang, Chao Feng, and Jun Xiao. VGR: Visual grounded reasoning. *arXiv preprint arXiv:2506.11991*, 2025.

Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. Understanding and mitigating hallucination in large vision-language models via modular attribution and intervention. In *The Thirteenth International Conference on Learning Representations*, 2025.

Heeji Yoon, Jaewoo Jung, Junwan Kim, Hyungyu Choi, Heeseong Shin, Sangbeom Lim, Honggyu An, Chaehyun Kim, Jisang Han, Donghyun Kim, Chanho Eom, Sunghwan Hong, and Seungryong Kim. Visual representation alignment for multimodal large language models. *arXiv preprint arXiv:2509.07979*, 2025.

Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from CLIP. In *European Conference on Computer Vision (ECCV)*, 2022.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *International Conference on Learning Representations (ICLR)*, 2024.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*, 2023.

Yi Zhu, Yanzhao Zhou, Qixiang Ye, Qiang Qiu, and Jianbin Jiao. Soft proposal networks for weakly supervised object localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

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

7 APPENDIX

7.1 ADDITIONAL ANALYSIS

This subsection describes the interventions and measurements used to assess Proposition 4.1. The evidence consists of paired forward-pass comparisons on a frozen model.

7.1.1 DIAGNOSTIC PROTOCOL

Examples and prediction task. We evaluate frozen LLaVA-1.5-7B on $N = 300$ COCO images using a common fixed generic prefix. Each image is paired with a single-token annotated object word w^+ and a single-token alternative w^- screened against instance annotations and reference captions. The model predicts at the position following the prefix, before either candidate is appended. All conditions use the same image, prefix, and candidates. Examples are selected without filtering on model outcomes: 213 have a positive baseline pairwise margin and 87 have a nonpositive margin. These counts describe preference between the two candidates, rather than vocabulary-wide prediction accuracy.

Intervention configuration. All reported steering conditions act on every head at zero-based decoder layer indices 1–16. Markov steering uses three propagation steps and blend coefficient $\alpha = 0.3$, and annotated-region and outside-region steering use redistribution strength $\lambda = 0.3$. The comparisons measure the complete forward pass with the configured interventions. They do not assume that isolated-head effects add across the decoder.

Recorded proposal initialization. Let $\mathbf{P}$ be the row-stochastic vision-encoder transition matrix over image patches. The reported diagnostics initialize propagation with the normalized incoming attention to each patch:

$$v_{0,j} = \frac{\sum_i P_{ij}}{\sum_{i,k} P_{ik}}, \qquad \mathbf{m}^\top = \mathbf{v}_0^\top \mathbf{P}^3, \tag{14}$$

where $\mathbf{v}_0$ is the initial patch distribution and $\mathbf{m}$ is the normalized proposal. Proposal entries are aligned with their corresponding decoder image-token positions.

7.1.2 ATTENTION INTERVENTIONS AND CONTROLS

For one edited attention row, let a_j be the original post-softmax weight at available key position j, with $a_j \geq 0$ and $\sum_j a_j = 1$. a_j is the same as $A[i,j]^{(l,h)}$ in Section 3.4. We use a_j here for simplicity reasons. Let $\mathcal{I}$ index image-token positions and $S = \sum_{j\in\mathcal{I}} a_j$ denote their total attention. Each condition applies its update to its own current attention row at every selected head and layer.

Annotated-region and outside-region steering. Let $R \subseteq \mathcal{I}$ be the annotated object's patch region and $C \subseteq \mathcal{I} \setminus R$ an equal-area control region. Their normalized masks, $\mathbf{h}_R$ and $\mathbf{h}_C$, have nonnegative entries summing to one over image positions and are supported on their respective regions. For the annotated region, set

$$a_j^R = \begin{cases} (1-\lambda)a_j + \lambda S h_{R,j}, & j \in \mathcal{I}, \\ a_j, & j \notin \mathcal{I}. \end{cases} \tag{15}$$

The control replaces $\mathbf{h}_R$ with $\mathbf{h}_C$. Both updates preserve image attention mass because $\sum_{j\in\mathcal{I}} a_j^R = (1-\lambda)S + \lambda S = S$, and both leave non-image entries unchanged. Thus, their local intervention differs in where image attention is redistributed. Region-control replicates are averaged within each image. Annotations and region masks are used only for these diagnostics. This comparison tests whether directing attention toward the annotated object region improves supported-word preference more than directing it toward an equal-area region outside the object.

Markov steering. For proposal entries $m_j \geq 0$ satisfying $\sum_{j\in\mathcal{I}} m_j = 1$, blending the image block and normalizing the full row gives

$$a_j^{\mathrm{M}} = \begin{cases} \big[(1-\alpha)a_j + \alpha m_j\big]/Z, & j \in \mathcal{I}, \\ a_j/Z, & j \notin \mathcal{I}, \end{cases} \qquad Z = 1 + \alpha(1-S). \tag{16}$$

Comparison	Mean margin advantage	95% CI
Annotated region vs. baseline	0.1057	[0.0845, 0.1277]
Annotated region vs. control region	0.0782	[0.0561, 0.1011]
Markov vs. baseline	0.3791	[0.3113, 0.4472]
Markov vs. mass-only	0.1773	[0.1353, 0.2199]

Table 6: Paired margin comparisons on 300 COCO images with frozen LLaVA-1.5-7B. All comparisons use the full configured head/layer intervention at strength 0.3. Annotated-region and control-region steering preserve each edited row's image attention mass. Markov comparisons use the recorded normalized column-sum seed, blending, and full-row normalization.

Here m_j can also be written as $\mathbf{p}_{\mathrm{prop,j}}$ to align with Section 3.4, and Z is the full-row normalization constant. The resulting total image attention is

$$\widetilde{S} = \sum_{j \in \mathcal{I}} a_j^{\mathrm{M}} = \frac{(1-\alpha)S + \alpha}{1 + \alpha(1-S)}. \tag{17}$$

Markov steering therefore changes both total image attention and its distribution among image positions. This comparison tests whether the complete Markov intervention improves supported-word preference over the baseline.

Proportional image-attention scaling. The mass-only control uses the same image-mass update while preserving relative weights within each token group. For $0 < S < 1$, define

$$a_j^{\mathrm{mass}} = \begin{cases} (\widetilde{S}/S)a_j, & j \in \mathcal{I}, \\ \big[(1-\widetilde{S})/(1-S)\big]a_j, & j \notin \mathcal{I}. \end{cases} \tag{18}$$

For a common input row, the Markov and mass-only updates produce the same image and non-image attention totals, while the mass-only update preserves the original within-image proportions. Across multiple layers, the runs can develop different internal states. The comparison therefore matches the local mass-increase rule, since it does not impose equal internal image-attention totals throughout the two forward passes. This comparison tests whether proposal-based redistribution provides additional benefit beyond proportionally increasing total image attention mass.

7.1.3 Estimation and Results

For image i, let D_i^s be the margin under condition s. For any two conditions r and s, estimate their mean margin difference by

$$\widehat{\mu}_{r,s} = \frac{1}{N}\sum_{i=1}^{N}\big(D_i^r - D_i^s\big), \qquad N = 300. \tag{19}$$

When a region control has multiple replicates, D_i^C is its within-image mean. Confidence intervals are percentile intervals from 5,000 image-level bootstrap resamples, keeping conditions paired within each resampled image. The intervals are descriptive and are not adjusted for multiple comparisons.

Evidence for usable visual grounding. Annotated-region steering improves the mean margin over baseline by 0.1057 and over equal-area outside-region controls by 0.0782, with positive confidence intervals in both cases. Because the model is frozen and each local edit preserves total image attention, these results support the premise that existing object-associated representations can contribute to the corresponding word's relative score when their attention weight is increased.

Evidence for the empirical proposition. The paired estimates are $\widehat{\mu}_{\mathrm{M},0} = 0.3791$ and $\widehat{\mu}_{\mathrm{M,mass}} = 0.1773$, with confidence intervals strictly above zero. These contrasts support the two average-effect claims in (11). The Markov intervention has $D_i^{\mathrm{M}} - D_i^0 > 0$ on 71.7% of images, with interval $[66.3\%, 77.0\%]$. This is the fraction of images with an improved margin, rather than the fraction with a correct generated token.

Interpretation. Directing attention toward the annotated object region improves the mean supported-word margin by 0.1057 over baseline, while Markov steering produces a larger mean improvement of 0.3791. At the tested intervention strength, focusing attention on the annotated region alone therefore yields less benefit than the complete Markov intervention. These interventions also differ in total image attention mass, so their comparison alone cannot attribute the larger gain to the proposal's structural information. However, Markov steering further outperforms proportional mass-only scaling by 0.1773, indicating that increasing total image attention does not explain its full benefit. Together, these results support the effectiveness of proposal-based attention redistribution, while leaving the precise contribution of the proposal's spatial structure unresolved.

7.2 Implementation Details

LLaVA-1.5 details. The Markov proposal is constructed from layer 18 of the CLIP ViT-L/14 vision encoder, selected as one of the two layers (18 and 22) exhibiting the lowest column-sum entropy in a preliminary scan over 20 randomly sampled images. We use layer 18 throughout. Heads are aggregated uniformly ($w_h = 1/H$) and the chain is iterated $n = 3$ times. We remove the [CLS] row and column before uniformly averaging attention across heads. The seed is selected from the column sums of the resulting patch-attention matrix $\bar{A}^{(18)}$, which is then row-normalized to obtain $\mathbf{P}$ for propagation. We blend the proposal into the decoder's image-token attention with coefficient $\alpha = 0.3$ at decoder layers 1 through 16 (the first half of the 32-layer Vicuna decoder). The blend is applied at every generated position, with no lexical gating of any kind, and in every attention head of those layers. It is applied post-softmax and post-mask to the raw image-token columns of the attention row, after which the full row is renormalized over the entire key sequence, as in (9) and Algorithm 2. Because image tokens precede all generated positions, they are never masked out for any steered row. Inference uses greedy decoding with a maximum of 128 new tokens and seed 42. The intervention adds no learned parameters and requires no auxiliary data.

For POPE (Section 5.3), generation allows up to 64 new tokens. We apply steering to the final prefill row and subsequent decoding rows using the same layers, heads, and blending coefficient.

MiniGPT-4 details. We use the MiniGPT-4 Llama-2-7B variant with an EVA-CLIP ViT-g/14 encoder at 224×224 resolution, extracting attention from zero-based encoder layer index 18 (the 19th block). We uniformly average heads to obtain $\bar{A}$. Here, $\bar{A}$ denotes the row-stochastic attention matrix obtained by uniformly averaging post-softmax attention across heads over all 257 encoder tokens, including [CLS]. We initialize from the normalized column sums of $\bar{A}$ and propagate for $n = 3$ steps using the transition matrix. MiniGPT-4 differs from LLaVA by initializing from normalized column sums over all 257 encoder tokens, including [CLS], and propagating for $n = 3$ steps. After discarding [CLS] and normalizing, we map the 256-patch proposal to 64 decoder visual-token positions by averaging each group of four consecutive patches in raster order and renormalizing. This grouping matches the connector's feature concatenation and linear projection, which uses no Q-Former. The resulting weights align with the image-token span following the BOS token and tokenized prompt prefix. Steering uses $\alpha = 0.3$ in every head at zero-based decoder layers 1–16. The blend is applied post-mask and post-softmax to the final prefill row and each subsequent decoding row, followed by renormalization over the full key sequence.

Computational Resources We run the experiment on eight Nvidia A5000 GPUs with 24GB memory each. Caption generation takes about 2.5 to 3 seconds per image, making the whole inference with 500 images around 30 minutes.

7.3 Limitations

The proposal's usefulness depends on the spatial structure of the selected encoder attention graph and on the starting distribution. Strong incoming attention does not guarantee an object-relevant seed, and propagation can favor an unrelated region or concentrate on tokens carrying global rather than local information. A single seed can reach multiple regions, but may still provide uneven coverage of the image. Because the proposal is query-independent, it can be less useful for questions about small or less salient objects. Its influence is particularly strong in attention rows with low original image mass, where an unfavorable prior can substantially alter the decoder's spatial allocation.

Our analysis characterizes attention-row updates exactly and evaluates their downstream effects empirically. It does not establish that regional attention is sufficient for correct generation, or that the effects of interventions across multiple layers combine constructively. The encoder layer, propagation depth, and decoder steering range are selected empirically and may require adjustment for another architecture. Models without a direct correspondence between encoder patches and decoder image tokens additionally require an explicit mapping of the proposal into the decoder's visual-token space.

Finally, reduced hallucination must be considered together with descriptive coverage. Our method reduces caption length, and the existing evaluation does not establish that the omitted content is uninformative. The additional vision-encoder pass in our LLaVA implementation, together with attention extraction, propagation, and steering, introduces computational overhead.

7.4 More Caption Comparisons

Table 7: More image examples with caption comparisons

Input	
Instruction	Please describe this image in detail.
Greedy	The image features a picturesque canal in a city, with several boats floating along the waterway. There are at least nine boats visible in the scene, with some of them being small and others larger. Some of the boats are docked near the edge of the canal, while others are floating further away. In addition to the boats, there are two bicycles parked near the canal, one closer to the left side and the other near the center of the scene. A person can be seen standing near the center of the image, possibly enjoying the view or attending to one of the boats. The overall atmosphere
Ours	The image depicts a canal with several boats docked along the side. The boats are of various sizes and are parked in a row. The canal is surrounded by buildings, giving the scene a charming atmosphere. The boats are docked in a row, and the canal is calm and serene.
Input	
Instruction	Please describe this image in detail.
Greedy	The image captures a lively scene on a beach, where a man is riding a horse while another man is riding a donkey. The man on the horse is holding a stick, possibly a whip, as he and the donkey rider gallop across the sandy shore. There are several other people in the scene, some of them standing or walking around, possibly enjoying the beach atmosphere. The beach is also populated by a few birds, adding to the vibrant and dynamic nature of the scene.
Ours	In the image, a man is riding a horse on the beach, with another person riding a horse nearby. The horses are galloping on the sandy beach, creating an exciting scene. There are also two other people in the background, one of them is holding a whip, adding to the action-packed atmosphere. The beach is a popular spot for horseback riding, and the people are enjoying the experience.

Table 8: More image examples with caption comparisons

Input	
Instruction	Please describe this image in detail.
Greedy	The image features a box filled with a variety of donuts, each with different flavors and toppings. There are six donuts in total, with some of them being glazed and others having frosting. The donuts are arranged in a visually appealing manner, with some placed closer to the front of the box and others towards the back. The assortment of donuts creates a mouth-watering display, inviting people to indulge in these sweet treats.
Ours	The image features a box of assorted donuts, including glazed, chocolate, and frosted donuts. The donuts are arranged in a visually appealing manner, with some donuts placed on top of each other. The donuts are of various flavors, making the display more enticing. The donuts are placed in a cardboard box, which is filled with a variety of donuts.
Input	
Instruction	Please describe this image in detail.
Greedy	The image features a small bathroom with a white toilet positioned next to a sink. The toilet is situated on the left side of the bathroom, while the sink is located on the right side. The sink is accompanied by a mirror above it. In addition to the toilet and sink, there are several bottles scattered around the bathroom. Some of these bottles are placed near the sink, while others are located closer to the toilet. The bathroom also has a potted plant, which is placed near the sink, adding a touch of greenery to the space.
Ours	The image features a bathroom with a white toilet and a sink. The toilet is positioned next to the sink, creating a compact and functional space. The sink is located on the right side of the toilet, while the toilet is situated on the left side. The bathroom is well-maintained and clean, making it an inviting space for users.

Table 9: More image examples with caption comparisons

Input	
Instruction	Please describe this image in detail.
Greedy	The image features a cat sitting on a wooden desk, looking at the camera with a curious expression. The cat is positioned near the center of the desk, surrounded by various items. There are two books on the desk, one located towards the left side and the other on the right side. A remote control can be seen on the right side of the desk, and a cell phone is placed closer to the right edge. Additionally, there is a car visible in the background, possibly a toy or a decoration.
Ours	The image features a cat sitting on a desk, looking at the camera. The cat is sitting on a desk, and there are several books and papers scattered around the desk. The cat appears to be looking at the camera, possibly curious about its surroundings.
Input	
Instruction	Please describe this image in detail.
Greedy	The image features a dining table set up outdoors, with a variety of food and drinks arranged on it. There are several plates of food, including sandwiches, fries, and a bowl of salad. A wine glass is placed on the table, along with a bottle, suggesting that the meal is accompanied by wine. In addition to the food and drinks, there are multiple chairs surrounding the table, inviting guests to sit and enjoy the meal. The outdoor setting and the presence of a lake in the background create a pleasant atmosphere for a gathering or
Ours	The image features a dining table with a variety of food items on it. There are several plates of food, including sandwiches, fries, and a bowl of fruit. The table is set up for a picnic, and the food is arranged in an appetizing manner. The table is surrounded by chairs, and the scene is set in a park.