OBJECT HALLUCINATIONS ARE WRITTEN BY LATE FEED-FORWARD LAYERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models often describe objects that are not in the image. Training-free decoding methods can reduce these hallucinations, but the most effective ones also make the model mention fewer of the objects that are present, and layer-level methods such as DoLa and DeCo help some models while hurting others. To understand why, we decompose the output logit of each generated object into the direct contributions of the attention and feed-forward network (FFN) sublayers at every layer. In four MLLMs, the two sublayers of the last few layers work against each other: attention favors objects present in the image, while the FFN favors hallucinated objects by a larger margin. We call this pattern FFN overcommitment. When the caption is held fixed and the image is progressively occluded, the attention advantage nearly vanishes but the FFN contribution to hallucinated objects stays the same, so the pull toward absent objects does not come from the visual input. Layer-level methods rescale both contributions together, so they weaken the support for present objects whenever they weaken the push toward absent ones. We therefore propose Component-wise Contrastive Decoding (CCD), which subtracts only the positive FFN contributions of the last few layers from the output logits and requires neither training nor an extra forward pass. Across four models, CCD reduces hallucination on CHAIR and AMBER and raises POPE accuracy on every split while preserving recall on CHAIR and POPE, whereas every compared method that reaches a lower caption hallucination rate does so by mentioning fewer of the objects that are present.

1 INTRODUCTION

Multimodal large language models (MLLMs) often describe objects that are not in the image (Rohrbach et al., 2018; Li et al., 2023c; Lovenia et al., 2024). These errors are fluent and confidently stated, which makes them hard for users to catch and risky in applications such as medical imaging and autonomous driving (Wu et al., 2024; Zhou et al., 2023a). They are commonly attributed to the language prior overriding the visual evidence (Zhou et al., 2023b; Leng et al., 2024): once a caption has mentioned a kitchen, "bowl" becomes a likely next word whether or not a bowl is in view. Training-free decoding methods build on this account. Contrastive methods compare the output with one computed from a corrupted image or instruction (Leng et al., 2024; Wang et al., 2024c) or from an intermediate layer (Chuang et al., 2023; Wang et al., 2024a), and attention-based methods increase the weight on image tokens or switch off heads that attend too much to the generated text (Huang et al., 2024; Liu et al., 2024b; An et al., 2025; Yang et al., 2025b). If the account were complete, suppressing the prior or reinforcing the image would remove hallucinated objects and leave the correct ones untouched.

That is not what happens. The methods that remove the most hallucinated objects from captions also mention fewer of the objects that are actually present, and recent methods from 2026 behave the same way (Jo et al., 2026; Ha et al., 2026). Whatever these methods suppress must carry some of the evidence for present objects as well. Layer-level methods raise a second puzzle. DoLa (Chuang et al., 2023) contrasts the final prediction with that of an intermediate layer, which amplifies what the later layers add, while DeCo (Wang et al., 2024a) mixes the intermediate prediction back into the final one, which damps it. If the late update simply pushed toward hallucination, damping it would help every model and amplifying it would hurt every model. Instead, each method lowers hallucination on

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.

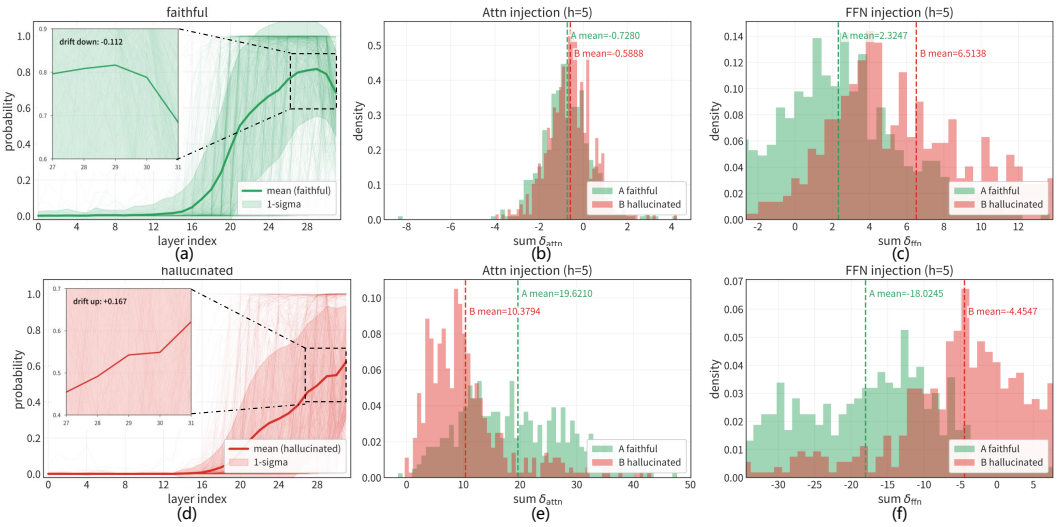


Figure 1: Direct logit contributions to 908 faithful and 908 hallucinated object tokens generated by LLaVA-1.5-7B, summed over the last five layers. (a, d) Layer-wise probability of the generated token: faithful tokens lose probability in the last layers, while hallucinated tokens gain it. (b, c) Before the final normalization, attention writes almost the same logit to both sets, whereas the FFN writes a much larger one to hallucinated tokens. (e, f) On the output logit, attention favors faithful tokens and the FFN favors hallucinated tokens by a larger margin.

some models and raises it on others (Section 4.2). Both puzzles suggest that these methods act on a unit that bundles the evidence for present objects together with whatever drives the error.

The error is made late in the network. Intermediate layers of an MLLM often already rank the correct object highly, and the final layers replace it with another (Wang et al., 2024a). On average we observe the same shift. Over the last few layers, objects that are present in the image, which we call faithful, lose probability, while hallucinated objects gain it (Figure 1a,d). A late layer, however, is not one computation but two. Its attention sublayer gathers information from other positions, including the image tokens, and its feed-forward network (FFN) transforms the representation at the current position. FFNs are known to store learned associations and to promote words directly in the output vocabulary (Geva et al., 2021; 2022; Dai et al., 2022; Meng et al., 2022), and causal tracing of MLLMs implicates deep-layer FFNs in how objects are represented (Li et al., 2026). Both sublayers add their outputs to the same residual stream, and a method that contrasts or mixes whole layers cannot tell them apart. We therefore ask a narrower question. When a model names an object that is not in the image, which of the two sublayers writes that choice into the output?

Because each sublayer adds its output, which we call its write, to the residual stream, this question has an exact answer. The logit of any word is the sum of contributions from the input embedding and from every attention and FFN sublayer, each obtained by projecting one write onto the output direction of that word, a technique known as direct logit attribution (Elhage et al., 2021; nostalgebraist, 2020; Belrose et al., 2023; Dar et al., 2023). We apply it to balanced sets of faithful and hallucinated object words generated by LLaVA-1.5 (Liu et al., 2023b), InstructBLIP (Dai et al., 2023) and two Qwen3-VL models (Bai et al., 2025). In the last few layers, the two sublayers work against each other. On the output logit, attention favors faithful objects, while the FFN favors hallucinated objects by a larger margin (Figure 1e,f). The pattern holds in all four models and is statistically significant in each (Appendix E), and we call it *FFN overcommitment*. The late layer that layer-level methods treat as a single unit therefore holds two opposing writes, and the stronger one favors the hallucinated object.

Only the attention side of this opposition depends on the image. On LLaVA-1.5, we hold each caption fixed and progressively degrade the image, from added noise to full occlusion (Figure 3a). As the image disappears, the advantage that attention gives faithful objects nearly vanishes, while the positive FFN write to hallucinated objects changes by less than one percent. Visual evidence reaches the output through attention, whereas the push toward a plausible but absent object comes through the late FFN and does not depend on the image. This is where the language prior of the standard account enters the output. In this sense, object hallucinations are written by late feed-forward layers.

For layer-level methods, this opposition ties hallucination reduction to a loss of support for present objects. The late update that DoLa amplifies and DeCo damps contains the attention support for faithful objects together with the FFN push toward hallucinated ones, and rescaling the update rescales both (Section 3.4). Where these methods reduce hallucination most, as DeCo does on LLaVA-1.5 and InstructBLIP, they lose recall, and where they keep recall, as on Qwen3-VL, they barely reduce hallucination or raise it. No reweighting of whole layers can keep one write and drop the other.

If the error enters through one write, the correction should remove that write and nothing else. We propose Component-wise Contrastive Decoding (CCD), which does so by subtracting a fraction of the positive FFN writes of the last few layers from the output logits at each decoding step. The attention writes, which carry the visual evidence, stay untouched, and so do the negative FFN writes, which already push wrong candidates down. Everything CCD uses is already computed in the ordinary forward pass, so it needs neither a second forward pass nor training. It also needs no detector for hallucinated tokens, because hallucinated objects receive larger FFN writes to begin with (Figure 1c) and the same correction therefore lowers them more. In practice, it changes the prediction six times as often when the original choice is hallucinated as when it is faithful. Since the correction applies from the first generated token, it acts on yes-or-no answers as well as on long captions.

Across four models, CCD reduces hallucination on CHAIR and AMBER, raises POPE accuracy on every split while preserving recall on CHAIR and POPE, and achieves the highest AMBER and MME total scores on every model. Every competing method that reaches a lower sentence-level hallucination rate on CHAIR does so by mentioning fewer present objects (Yang et al., 2025b; Jo et al., 2026; Ha et al., 2026).

Our contributions are as follows.

- We find that in the last few layers of four MLLMs, attention and the FFN write opposite preferences into the output logit, with the FFN favoring hallucinated objects by the larger margin. On LLaVA-1.5, removing the image nearly erases the attention advantage but leaves the FFN write to hallucinated objects unchanged.
- We show that layer-level contrastive decoding can only rescale the two writes together, which ties its reduction of hallucination to a loss of support for present objects.
- We propose CCD, a training-free, single-pass decoding rule that removes only the positive late-layer FFN writes, reducing hallucination on four models while preserving recall.

2 Related Work

Object hallucination in MLLMs.MLLMs (Alayrac et al., 2022; Liu et al., 2023c;b; Dai et al., 2023; Li et al., 2023a; Zhu et al., 2023; Bai et al., 2023; 2025; Achiam et al., 2023) frequently mention objects absent from the image (Rohrbach et al., 2018; Li et al., 2023c; Lovenia et al., 2024; Lin et al., 2024; Tong et al., 2024). The errors have been attributed to language priors (Zhou et al., 2023b; Leng et al., 2024), co-occurrence statistics (Li et al., 2023c) and visual grounding that weakens during generation (Huang et al., 2024; An et al., 2025), and are measured by CHAIR (Rohrbach et al., 2018), POPE (Li et al., 2023c), AMBER (Wang et al., 2023) and MME (Fu et al., 2023). We locate where in the forward pass this bias enters the output.

Mitigating hallucination. Training-based methods require additional data or retraining (Liu et al., 2023a; Yu et al., 2024a; Jiang et al., 2024; Sarkar et al., 2024; Chen et al., 2025; Yu et al., 2024b; Sun et al., 2024; Lyu et al., 2024; Yang et al., 2025c). Training-free methods contrast the output with one from a corrupted input or an intermediate layer (Leng et al., 2024; Zhang et al., 2025; Wang et al., 2024c; Chuang et al., 2023; Wang et al., 2024a; Huo et al., 2024; Chen et al., 2024; Wang et al., 2024b; Suo et al., 2025; Li et al., 2025), reweight or switch off attention (Huang et al., 2024; An et al., 2025; Liu et al., 2024b; Wan et al., 2025; Tang et al., 2025a;b; Jiang et al., 2025; Zhuang et al., 2025; Woo et al., 2025; Yang et al., 2025b; Jo et al., 2026), steer hidden representations (Zou et al., 2024; Yin et al., 2025; Yang et al., 2025a; Liu et al., 2024a; Dong et al., 2025; Duan et al., 2025; Shang et al., 2024; Wang et al., 2025; Li et al., 2023b), boost logits during decoding (Ha et al., 2026), or verify the output after generation (Yin et al., 2024; Gunjal et al., 2024; Ghosh et al., 2024). CCD is a contrastive method whose contrast comes from neither a second forward pass nor a whole-layer readout, but from the positive late-layer FFN writes already contained in the final logit. Appendix V discusses work on interpreting transformer components, including direct logit attribution and the hallucination heads of Yang et al. (2025b).

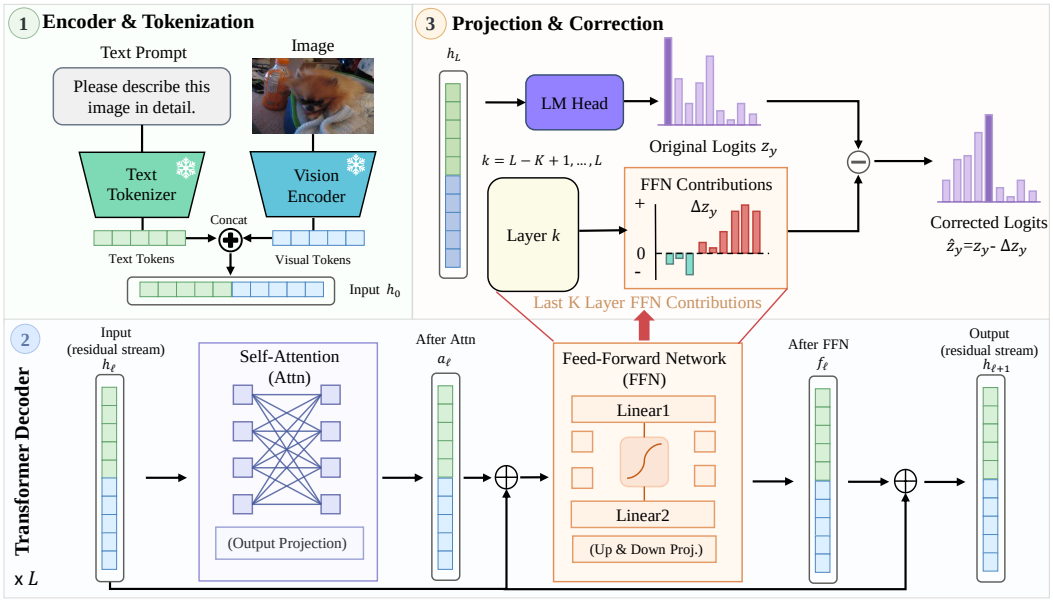


Figure 2: Component-wise Contrastive Decoding. (1) Prompt and image tokens form the decoder input $\mathbf{h}_0$. (2) Each layer adds an attention write $\mathbf{a}_\ell$ and an FFN write $\mathbf{f}_\ell$ to the residual stream; the input and output drawn as $\mathbf{h}_\ell$ and $\mathbf{h}_{\ell+1}$ are $\mathbf{h}_{\ell-1}$ and $\mathbf{h}_\ell$ in Eq. 2, and the squares inside the FFN are its hidden units. (3) For each of the last K layers, CCD projects the FFN write onto the vocabulary and keeps the positive part; the scaled sum $\Delta z_y = \alpha\, \Phi^+(y)$ is subtracted from the logits z_y (Eq. 7), and no parameter is updated.

3 Method

3.1 Preliminaries

An MLLM such as LLaVA (Liu et al., 2023b) encodes the image with a vision encoder, maps the visual features into the embedding space of a language model through a projector, and generates text with a decoder-only transformer of L layers (Touvron et al., 2023). We consider one decoding step at a time. Let $\mathbf{h}_0 \in \mathbb{R}^d$ be the input embedding at the current position and $\mathbf{h}_\ell$ the hidden state after layer ℓ. In the pre-normalization architecture shared by the models we study, layer ℓ applies self-attention and then the FFN, each to a normalized copy of the stream,

$$\mathbf{a}_\ell = \mathrm{Attn}_\ell\big(\mathrm{RMSNorm}(\mathbf{h}_{\ell-1})\big), \qquad \mathbf{f}_\ell = \mathrm{FFN}_\ell\big(\mathrm{RMSNorm}(\mathbf{h}_{\ell-1} + \mathbf{a}_\ell)\big), \tag{1}$$

and adds both outputs to the residual stream,

$$\mathbf{h}_\ell = \mathbf{h}_{\ell-1} + \mathbf{a}_\ell + \mathbf{f}_\ell. \tag{2}$$

We call $\mathbf{a}_\ell$ and $\mathbf{f}_\ell$ the attention write and the FFN write of layer ℓ. The FFN write depends on the attention write through its input, but the two enter the stream as separate terms. After the last layer, the language modeling head produces the logit of each word y in the vocabulary $\mathcal{V}$,

$$z_y = \frac{1}{s}\big[W_U\, \mathrm{diag}(\boldsymbol{\gamma})\, \mathbf{h}_L\big]_y, \qquad s = \|\mathbf{h}_L\|_{\mathrm{RMS}}, \tag{3}$$

where $W_U \in \mathbb{R}^{|\mathcal{V}| \times d}$ is the unembedding matrix, $\boldsymbol{\gamma} \in \mathbb{R}^d$ is the gain of the final RMSNorm (Zhang & Sennrich, 2019), and $\|\mathbf{h}\|_{\mathrm{RMS}}$ is the root mean square of the entries of $\mathbf{h}$. The next-token distribution is $p(y) = \mathrm{softmax}(\mathbf{z})_y$.

3.2 Direct Logit Attribution

Unrolling Eq. 2 expresses the final state as the embedding plus every write,

$$\mathbf{h}_L = \mathbf{h}_0 + \sum_{\ell=1}^{L} \mathbf{a}_\ell + \sum_{\ell=1}^{L} \mathbf{f}_\ell. \tag{4}$$

Since s is a single number at each step, substituting Eq. 4 into Eq. 3 splits every logit exactly into one term per write,

$$z_y = \tilde{\delta}^{\,\mathrm{emb}}(y) + \sum_{\ell=1}^{L} \tilde{\delta}_\ell^{\,\mathrm{attn}}(y) + \sum_{\ell=1}^{L} \tilde{\delta}_\ell^{\,\mathrm{ffn}}(y), \qquad \tilde{\delta}_\ell^{\,\mathrm{ffn}}(y) = \frac{1}{s}\big[W_U \,\mathrm{diag}(\boldsymbol{\gamma})\, \mathbf{f}_\ell\big]_y, \tag{5}$$

with $\tilde{\delta}_\ell^{\,\mathrm{attn}}$ and $\tilde{\delta}^{\,\mathrm{emb}}$ defined in the same way from $\mathbf{a}_\ell$ and $\mathbf{h}_0$. Each term measures how much one write raises or lowers the logit of y directly, which is the direct logit attribution of that write (Elhage et al., 2021; nostalgebraist, 2020; Belrose et al., 2023; Dar et al., 2023). The attribution is to sublayers at the current position, not to input tokens. Mixing across positions happens inside $\mathbf{a}_\ell$, and however much of it occurs, the resulting write is one vector whose contribution is exact. We also use the unnormalized projections

$$\delta_\ell^{\,\mathrm{attn}}(y) = [W_U\, \mathbf{a}_\ell]_y, \qquad \delta_\ell^{\,\mathrm{ffn}}(y) = [W_U\, \mathbf{f}_\ell]_y, \tag{6}$$

which sum to $[W_U \mathbf{h}_L]_y$ and differ from the tilde terms only by the gain $\boldsymbol{\gamma}$ and the scale s shared by all words at a step.

3.3 Component-wise Contrastive Decoding

CCD acts on FFN overcommitment, the pattern in which the late FFN favors hallucinated objects by a larger margin than attention favors faithful ones (Figure 1e,f). It removes the part of the late FFN write that favors a word and keeps everything else. At every decoding step, it corrects the logits as

$$\hat{z}_y = z_y - \Delta z_y, \qquad \Delta z_y = \alpha\, \Phi^+(y), \qquad \Phi^+(y) = \sum_{k=L-K+1}^{L} \big[\delta_k^{\,\mathrm{ffn}}(y)\big]_+, \tag{7}$$

where $[x]_+ = \max(0, x)$, the index k runs over the last K layers, and $\alpha > 0$ sets the strength. We call Δz_y the correction to word y. The model then decodes from $\mathrm{softmax}(\hat{\mathbf{z}})$ with any decoding rule. The model then decodes from $\mathrm{softmax}(\hat{\mathbf{z}})$ with any decoding rule (Figure 2).

Each part of Eq. 7 follows from this pattern. The correction targets the FFN because attention carries the visual evidence for present objects, and its advantage fades once the image is removed (Figure 3a). It is restricted to the last K layers because faithful and hallucinated objects separate only there (Figure 1a,d). It takes the positive part of each FFN write because the negative part lowers the logits of competing candidates, work that the correction should not undo. The correction uses the unnormalized projections of Eq. 6, which require only the unembedding matrix, and the FFN favors hallucinated objects in this reading as well as on the output logit (Figure 1c,f).

Because the correction is additive, its effect on any pair of candidates is exact. For a faithful word y_f and a hallucinated word y_h, the softmax normalizer cancels in their ratio, which gives

$$\log \frac{\hat{p}(y_f)}{\hat{p}(y_h)} = \log \frac{p(y_f)}{p(y_h)} + \alpha\big(\Phi^+(y_h) - \Phi^+(y_f)\big). \tag{8}$$

The correction thus moves probability toward whichever candidate received the smaller positive FFN write, in proportion to the gap, and reverses their ranking once α times the gap exceeds the margin by which y_h leads. FFN overcommitment, $\Phi^+(y_h) > \Phi^+(y_f)$, is exactly the condition under which Eq. 8 favors the faithful word. A single α applied to the whole vocabulary therefore acts mostly on hallucinated objects, and no separate detector is needed (Appendix J). Appendix G characterizes which words lose probability overall and when keeping the negative FFN writes strictly improves the correction.

All the writes that CCD uses are produced by the ordinary forward pass, so it needs neither a second forward pass nor training, and its only extra work is projecting the late FFN writes through the unembedding matrix (Appendix I). Because the correction is applied at every step, including the first, it also acts on single-token answers such as those in POPE.

3.4 Relation to Layer-Level Decoding

Layer-level methods read an intermediate layer through the final RMSNorm and the language modeling head, giving $z_y^{(m)} = \frac{1}{s_m}\big[W_U \,\mathrm{diag}(\boldsymbol{\gamma})\, \mathbf{h}_m\big]_y$ with $s_m = \|\mathbf{h}_m\|_{\mathrm{RMS}}$. Substituting Eq. 4 into

Table 1: Comparison of contrastive decoding methods from the component-wise perspective.

Method	Separates Attn/FFN	Sign-selective	Multi-layer	Training-free	Single pass
DoLa Chuang et al. (2023)	✗	✗	✗	✓	✓
DeCo Wang et al. (2024a)	✗	✗	✗	✓	✓
VCD Leng et al. (2024)	✗	✗	–	✓	✗
CCD (ours)	✓	✓	✓	✓	✓

Eq. 3 gives, with no approximation,

$$z_y = \frac{s_m}{s}\, z_y^{(m)} + \sum_{\ell=m+1}^{L} \Big(\tilde{\delta}_\ell^{\,\mathrm{attn}}(y) + \tilde{\delta}_\ell^{\,\mathrm{ffn}}(y) \Big), \tag{9}$$

The late update $u_y^{(m)} = z_y - z_y^{(m)}$ therefore consists of the attention and FFN writes of every later layer, added with equal weight, minus a rescaled copy $(1 - s_m/s)\, z_y^{(m)}$ of the intermediate prediction.

DoLa (Chuang et al., 2023) selects a premature layer M at each step and, among plausible candidates, scores

$$z_y^{\mathrm{DoLa}} = \log q_L(y) - \log q_M(y) = u_y^{(M)} + c, \tag{10}$$

where $q_m = \mathrm{softmax}(\mathbf{z}^{(m)})$ and c does not depend on y. DoLa thus ranks candidates by the late update alone. DeCo (Wang et al., 2024a) selects an anchor layer m^* and adds its logits to the final ones with weight $\beta > 0$. Up to a positive factor, which leaves the ranking unchanged,

$$z_y + \beta\, z_y^{(m^*)} \;\propto\; z_y^{(m^*)} + \frac{1}{1+\beta}\, u_y^{(m^*)}, \tag{11}$$

so DeCo keeps the earlier prediction and shrinks the late update.

In both methods the late update enters as a whole. DoLa gives it all the weight and DeCo less than its usual share, but neither can change the relative weight of the attention and FFN terms inside it. Since these terms pull faithful and hallucinated candidates in opposite directions, shrinking the update weakens the push toward hallucinated objects only by weakening the support for faithful ones, which costs recall. Whether rescaling helps at all then depends on how the two terms balance over the layers after m, a span chosen anew at each step and a balance that differs across models. CCD instead subtracts only the positive FFN terms of the last K layers, leaving the attention terms, the negative FFN terms and the earlier layers untouched. Appendix G states this comparison in terms of log-odds.

Table 2: **CHAIR evaluation on MSCOCO**. Best in **bold**, second best underlined.

	LLaVA-1.5-7B			InstructBLIP-7B		
Method	$\mathrm{CHAIR}_S \downarrow$	$\mathrm{CHAIR}_I \downarrow$	Recall↑	$\mathrm{CHAIR}_S \downarrow$	$\mathrm{CHAIR}_I \downarrow$	Recall↑
Greedy	54.2	18.5	72.5	54.0	18.1	71.1
+ ICD	56.2	16.3	76.8	56.5	19.5	71.0
+ VCD	51.8	16.2	76.8	57.0	17.0	72.1
+ VAF	52.6	14.2	76.0	54.5	16.5	71.5
+ AGLA	50.0	15.3	76.5	49.0	**12.1**	**72.5**
+ ONLY	49.8	14.3	**78.1**	52.2	15.5	71.0
+ OPERA	48.5	14.9	**78.1**	54.0		
underline12.8	69.8					
+ MEMVR	48.0	14.5	77.1	51.0	15.0	71.5
+ DoLa	48.4	15.9	69.3	60.0	20.1	71.5
+ DeCo	**41.6**	13.4	69.7	44.4	15.2	66.0
+ CCD (ours)	43.8	**13.3**	76.1	**43.6**	12.8	71.7

4 Experiments

4.1 Setup

Models and benchmarks. We evaluate CCD on four models spanning two architectural generations and three backbone families: LLaVA-1.5-7B (Liu et al., 2023b), InstructBLIP-7B (Dai et al., 2023), Qwen3-VL-4B, and Qwen3-VL-8B (Bai et al., 2025). Architectural details are provided in Appendix O.1. We use four benchmarks that evaluate complementary aspects of hallucination. CHAIR (Rohrbach et al., 2018) measures object-level hallucination and recall in free-form captions

Table 3: **CHAIR evaluation on Qwen3-VL**. Best in **bold**, second best underlined.

	Qwen3-VL-4B			Qwen3-VL-8B		
Method	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	Recall↑	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	Recall↑
Greedy	26.0	9.4	49.2	24.4	8.2	47.4
+ VCD	25.8	9.4	49.7	24.0	8.1	47.8
+ VAF	24.8	8.8	49.5	23.4	7.8	47.6
+ DoLa	27.0	9.3	49.8	24.2	8.5	48.0
+ DeCo	27.0	10.0	50.5	23.0	8.0	48.4
+ CCD (ours)	**23.0**	**8.1**	**50.8**	**21.8**	**7.3**	**49.9**

Table 4: **AMBER evaluation** across four MLLM architectures. Best in **bold**, second best underlined.

	Generative				Discriminative				Total	
Method	CHAIR↓	Cover↑	Hal↓	Cog↓	Acc↑	Prec↑	Rec↑	F1↑	**Score**	**Gain**
LLaVA-1.5-7B										
Greedy	11.2	51.6	52.0	7.0	68.7	77.9	71.5	74.6	274.1	**+0.0**
+ VCD	9.8	52.2	47.0	6.8	67.2	76.8	69.9	73.2	275.7	**+1.6**
+ VAF	11.5	47.8	51.0	5.3	66.5	78.5	65.6	71.5	262.1	**-12.0**
+ DoLa	8.1	**53.5**	38.2	5.3	75.7	74.3	**96.9**	84.1	332.9	**+58.8**
+ DeCo	**6.3**	51.2	29.6	4.6	80.0	**93.0**	75.5	83.3	342.5	**+68.4**
+ CCD (ours)	**6.3**	50.0	**28.8**	**4.2**	**81.4**	91.2	79.6	**85.0**	**347.9**	**+73.8**
InstructBLIP-7B										
Greedy	6.4	53.9	31.7	5.0	75.3	92.3	68.5	78.6	325.5	**+0.0**
+ VCD	**5.4**	57.0	30.8	2.8	**77.5**	91.2	72.8	81.0	340.5	**+15.0**
+ VAF	5.6	56.5	31.5	**2.6**	77.2	91.5	72.2	80.7	338.4	**+12.9**
+ DoLa	12.3	**57.3**	56.0	8.3	75.9	74.8	**85.8**	79.9	297.1	**-28.4**
+ DeCo	6.3	49.3	32.0	3.6	67.4	**97.0**	52.4	68.0	292.2	**-33.3**
+ CCD (ours)	5.8	55.1	**27.1**	4.2	74.2	92.6	76.4	**83.7**	**344.9**	**+19.4**
Qwen3-VL-4B										
Greedy	5.9	52.5	31.0	1.9	89.8	91.8	93.0	92.4	380.7	**+0.0**
+ VCD	5.4	51.0	28.5	1.7	90.2	92.5	93.2	92.8	384.1	**+3.4**
+ VAF	5.2	51.2	28.2	**1.6**	90.5	92.8	93.5	93.1	386.1	**+5.4**
+ DoLa	6.5	52.6	34.9	2.3	88.5	88.4	**95.0**	91.6	372.4	**-8.3**
+ DeCo	6.0	54.8	32.0	1.9	90.0	**93.3**	91.4	92.3	381.9	**+1.2**
+ CCD (ours)	**5.1**	**56.7**	**28.0**	2.0	**90.8**	92.9	93.8	**93.3**	**392.4**	**+11.7**
Qwen3-VL-8B										
Greedy	6.0	48.8	30.7	1.8	89.3	91.5	92.4	91.9	375.4	**+0.0**
+ VCD	5.5	48.1	28.2	1.6	90.0	91.2	92.8	92.0	378.8	**+3.4**
+ VAF	5.3	48.3	27.9	**1.4**	90.2	91.5	93.0	92.2	380.6	**+5.2**
+ DoLa	6.7	**49.0**	33.1	2.1	75.9	74.8	**95.8**	84.0	337.6	**-37.8**
+ DeCo	6.7	**49.0**	33.1	2.1	87.7	87.5	95.1	91.1	368.5	**-6.9**
+ CCD (ours)	**4.5**	**49.0**	**26.9**	1.5	**90.6**	**91.9**	93.6	**92.7**	**384.9**	**+9.5**

on 500 MSCOCO images. POPE (Li et al., 2023c) reformulates hallucination as binary object probing under Random, Popular, and Adversarial splits. AMBER (Wang et al., 2023) provides joint generative and discriminative evaluation. MME (Fu et al., 2023) assesses general perception and cognition capabilities. Full metric definitions are given in Appendix T.

Baselines. We compare against nine training-free methods spanning three intervention families: contrastive decoding with ICD (Wang et al., 2024c), VCD (Leng et al., 2024), DoLa (Chuang et al., 2023), and DeCo (Wang et al., 2024a); attention-based intervention with OPERA (Huang et al., 2024), AGLA (An et al., 2025), and ONLY (Wan et al., 2025); and representation correction with MEMVR (Zou et al., 2024) and VAF (Yin et al., 2025). All baselines use released code with reported-best hyperparameters. Implementation details and CCD's model-specific hyperparameters are provided in Appendix O.

4.2 Main Results

Tables 2–5 report the results. On CHAIR, CCD lowers both hallucination metrics on all four models without reducing recall, and the one method that reaches a lower $CHAIR_S$, DeCo on LLaVA-1.5, loses 2.8 recall points. On POPE, CCD obtains the highest F1 on every model and split and moves

Table 5: **POPE evaluation**. Best in **bold**, second best underlined.

	Random				Popular				Adversarial			
Method	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
LLaVA-1.5-7B												
Greedy	83.49	88.84	76.76	82.28	79.98	82.47	76.76	79.34	76.03	76.11	76.80	76.26
+ ICD	84.87	89.96	78.53	83.27	82.93	86.28	78.46	81.45	81.07	**82.64**	78.60	79.96
+ VCD	86.84	87.15	**86.68**	86.83	82.65	87.15	80.59	83.37	77.31	73.43	86.47	79.28
+ OPERA	87.53	**94.52**	79.80	86.45	84.21	**88.00**	79.80	83.50	80.88	82.16	79.76	80.69
+ AGLA	88.54	94.41	82.08	87.71	85.14	87.88	82.08	84.68	81.13	81.20	82.10	81.36
+ DoLa	84.78	87.59	81.27	84.19	79.75	84.11	76.22	80.61	76.32	77.27	75.47	76.16
+ DeCo	85.80	89.20	82.00	85.42	82.40	85.50	79.50	82.38	78.80	79.60	78.80	79.18
+ CCD (ours)	**89.23**	91.47	86.53	**88.93**	**87.17**	87.64	**86.53**	**87.08**	**81.57**	78.64	**86.67**	**82.46**
InstructBLIP-7B												
Greedy	80.42	78.93	83.21	80.94	76.09	73.22	82.94	77.65	72.37	68.78	83.06	75.42
+ ICD	82.50	81.80	84.20	82.80	78.00	75.90	83.60	79.10	73.80	70.20	83.80	76.00
+ VCD	84.11	84.20	85.36	84.13	79.94	77.84	83.33	80.80	76.32	73.24	84.08	78.08
+ OPERA	85.07	88.39	80.73	84.39	78.33	73.85	**87.73**	80.20	75.50	70.49	**87.73**	78.17
+ AGLA	87.30	88.83	85.68	87.07	81.86	80.17	85.68	82.58	77.29	74.09	85.67	79.16
+ DoLa	83.00	83.06	83.13	83.00	78.99	77.12	83.13	79.85	74.67	71.53	83.11	76.68
+ DeCo	84.60	85.40	83.80	84.50	79.90	77.60	84.40	80.80	76.20	72.90	84.30	78.20
+ CCD (ours)	**90.63**	**91.60**	**89.47**	**90.52**	**85.63**	**85.14**	86.33	**85.73**	**82.23**	**79.94**	86.07	**82.89**
Qwen3-VL-4B												
Greedy	90.20	96.80	83.20	89.50	87.30	93.20	81.00	86.60	85.80	90.00	81.00	85.20
+ VCD	90.50	96.50	83.80	89.70	87.60	93.50	81.50	87.10	86.00	90.20	81.50	85.60
+ VAF	91.00	97.00	84.80	90.40	88.20	94.20	82.20	87.80	86.80	91.00	82.50	86.50
+ DoLa	90.40	**98.64**	81.93	89.51	89.03	**95.49**	81.93	88.20	88.30	**93.96**	81.87	87.50
+ DeCo	91.97	98.39	85.33	91.40	89.80	93.70	85.33	89.32	88.33	90.78	85.33	87.97
+ CCD (ours)	**94.00**	94.90	**93.00**	**93.94**	**90.00**	93.42	**86.07**	**89.59**	**88.57**	90.60	**86.07**	**88.27**
Qwen3-VL-8B												
Greedy	88.25	94.82	81.15	87.38	85.85	90.26	81.05	85.41	84.12	86.35	81.20	83.19
+ VCD	89.12	95.36	82.73	88.25	86.63	90.81	82.17	86.28	84.86	87.12	82.34	84.27
+ VAF	89.68	95.74	83.46	88.79	87.19	91.25	82.93	86.83	85.37	87.68	83.16	84.92
+ DoLa	89.80	**98.69**	80.67	88.77	88.50	**95.65**	80.67	87.52	**87.50**	**93.44**	80.67	86.58
+ DeCo	91.83	97.57	85.80	91.31	88.90	91.47	85.80	88.54	86.93	87.89	85.67	86.77
+ CCD (ours)	**93.70**	92.70	**94.87**	**93.77**	**89.33**	89.28	**89.40**	**89.34**	87.40	87.85	**86.80**	**87.32**

the share of yes answers toward 50% in 11 of the 12 settings (Appendix S), so its gain does not come from shifting the answer in a fixed direction. It also has the highest AMBER and MME total scores on every model (MME in Appendix Q). Where the layer-level baselines reduce hallucination most, as DeCo does on LLaVA-1.5 and InstructBLIP, they lose recall, and both lower the AMBER total of InstructBLIP and Qwen3-VL-8B. We further reproduce AD-HH (Yang et al., 2025b) and two methods from 2026 (Jo et al., 2026; Ha et al., 2026), each against greedy decoding in its own run (Appendix P). Every one that reaches a lower CHAIR_S than CCD loses recall, and those whose correction depends on the text generated so far leave almost all POPE answers unchanged.

4.3 Ablation Studies

Table 6 changes one design choice of CCD at a time on LLaVA-1.5 with CHAIR. No variant matches CCD on both CHAIR_S and recall, and each fails in the way the diagnosis predicts. Applying the correction to attention instead of the FFN makes hallucination worse than no intervention and lowers recall, as expected for the sublayer that carries visual evidence. Applying it to both sublayers reduces hallucination less than CCD and lowers recall below that of greedy decoding. Removing the negative FFN writes raises hallucination above that of greedy decoding, and removing the FFN writes of both signs reduces hallucination about as much as CCD but gives up most of its recall gain. Extending the correction to every layer both weakens the reduction and lowers recall, since the two cohorts separate only in the last layers (Figure 1a,d). Every (α, K) in our grid improves on greedy decoding in all three CHAIR metrics (Appendix H), and under beam search and nucleus sampling CCD again raises recall while DoLa and DeCo lower it (Appendix R).

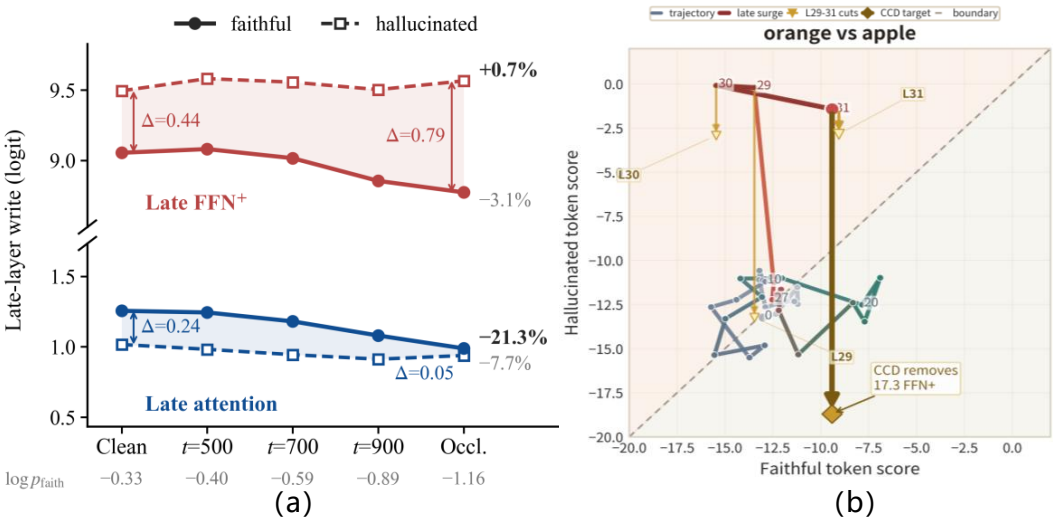


Figure 3: **(a)** Late-layer writes to faithful and hallucinated objects on LLaVA-1.5-7B as the image is degraded with the caption held fixed. The attention advantage of faithful objects nearly vanishes, while the positive FFN write (FFN$^+$) to hallucinated objects stays flat. The bottom row gives the mean log-probability of faithful objects, confirming that visual evidence is removed. **(b)** Layer-by-layer logits of the hallucinated word "orange" against the faithful word "apple". The late FFN writes carry the trajectory across the decision boundary, and the terms that CCD subtracts (gold) bring it back (Appendix N).

4.4 Analysis

Which write depends on the image? We teacher-force the original LLaVA-1.5 captions while degrading the image in steps, from diffusion noise as in VCD (Leng et al., 2024) to full occlusion, and recompute the late writes to the same faithful and hallucinated words (Figure 3a). As visual evidence disappears, the attention write to faithful objects falls steadily and its advantage over hallucinated objects nearly vanishes, while the positive FFN write to hallucinated objects stays flat. The push toward hallucinated objects therefore does not come from the image.

Table 6: **Component ablation on CHAIR** with LLaVA-1.5-7B. Each variant modifies one design axis.

Variant	Comp.	Sign	Layers	$CH_S\downarrow$	$CH_I\downarrow$	Rec↑
Greedy	—	—	—	54.2	18.5	72.5
Attn-only	Attn	pos	late	55.8	18.9	68.3
Attn+FFN	Both	pos	late	47.4	15.5	70.2
Neg-only	FFN	neg	late	56.4	19.6	69.8
Both-sign	FFN	all	late	44.6	**13.2**	72.8
All-layers	FFN	pos	all	49.0	15.3	70.6
CCD	**FFN**	**pos**	**late**	**43.8**	13.3	**76.1**

A single case. Figure 3b traces the logits of a hallucinated object and a competing faithful one across layers. The faithful object leads through the middle layers, the late FFN writes carry the hallucinated object past it, and the terms that CCD subtracts restore the faithful object, as in five of the six cases in Appendix N.

5 Conclusion

Object hallucinations in MLLMs are written by late feed-forward layers. In the last few layers of four models, attention favors the objects in the image, while the FFN favors absent objects by a larger margin. On LLaVA-1.5, degrading the image removes the attention advantage but leaves the FFN push intact. Layer-level decoding rescales both writes together, which is why it loses recall and behaves inconsistently across models. CCD removes only the positive late FFN writes and reduces hallucination while preserving recall, with no training and no extra forward pass. More broadly, interventions guided by interpretability can act on the write that carries an error rather than on the layer that contains it, and we expect this granularity to be useful beyond hallucination.

AI USE STATEMENT

Generative AI tools were used solely for language editing and coding assistance. Specifically, they were used to improve the clarity, grammar, and presentation of the manuscript, as well as to assist with code implementation and debugging. They were not used to formulate the research problem, develop the core methodology, conduct experiments, analyze results, or draw scientific conclusions. All AI-assisted text was carefully reviewed and revised by the authors, and all AI-assisted code was inspected, tested, and validated for correctness. The authors take full responsibility for the final content of this work, including all text, claims, results, and code produced with the assistance of generative AI.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

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

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Cong Chen, Mingyu Liu, Chenchen Jing, Yizhou Zhou, Fengyun Rao, Hao Chen, Bo Zhang, and Chunhua Shen. Perturbollava: Reducing multimodal hallucinations with perturbative visual training. *arXiv preprint arXiv:2503.06486*, 2025.

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:2403.00425*, 2024.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*, 2023.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, 2023.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8493–8502, 2022.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. Analyzing transformers in embedding space. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16124–16170, 2023.

Xin Dong, Shichao Dong, Jin Wang, Jing Huang, Li Zhou, Zenghui Sun, Lihua Jing, Jinsong Lan, Xiaoyong Zhu, and Bo Zheng. Inter: Mitigating hallucination in large vision-language models by interaction guidance sampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2534–2544, 2025.

Jinhao Duan, Fei Kong, Hao Cheng, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. Truthprint: Mitigating lvlm object hallucination via latent truthful-guided pre-intervention. *arXiv preprint arXiv:2503.10602*, 2025.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5484–5495, 2021.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pp. 30–45, 2022.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235, 2023.

Sreyan Ghosh, Chandra Kiran Reddy Evuru, Sonal Kumar, Utkarsh Tyagi, Oriol Nieto, Zeyu Jin, and Dinesh Manocha. Visual description grounding reduces hallucinations and boosts reasoning in lvlms. *arXiv preprint arXiv:2405.15683*, 2024.

Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18135–18143, 2024.

Jiwoo Ha, Jongwoo Baek, and Jinhyun So. First logit boosting: Visual grounding method to mitigate object hallucination in large vision-language models. *arXiv preprint arXiv:2604.00455*, 2026.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. Linearity of relation decoding in transformer language models. *arXiv preprint arXiv:2308.09124*, 2023.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427, 2024.

Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-introspective decoding: Alleviating hallucinations for large vision-language models. *arXiv preprint arXiv:2408.02032*, 2024.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27036–27046, 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 Computer Vision and Pattern Recognition Conference*, pp. 25004–25014, 2025.

Yujin Jo, Sangyoon Bae, and Taesup Kim. Attention-space contrastive guidance for efficient hallucination mitigation in lvlms. *arXiv preprint arXiv:2601.13707*, 2026.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024.

Jiaming Li, Jiacheng Zhang, Zequn Jie, Lin Ma, and Guanbin Li. Mitigating hallucination for large vision language model by inter-modality correlation calibration decoding. *arXiv preprint arXiv:2501.01926*, 2025.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023a.

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*, 36:41451–41530, 2023b.

Qiming Li, Zekai Ye, Xiaocheng Feng, Weihong Zhong, Weitao Ma, and Xiachong Feng. Causal tracing of object representations in large vision language models: Mechanistic interpretability and hallucination mitigation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 31645–31653, 2026.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023c.

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

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pp. 366–384. Springer, 2024.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 2023a.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023b.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023c.

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

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. *arXiv preprint arXiv:2407.21771*, 2024b.

Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. Negative object presence evaluation (nope) to measure object hallucination in vision-language models. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pp. 37–58, 2024.

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

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

nostalgebraist. interpreting GPT: the logit lens, 2020. URL https://www.alignmentforum.org/posts/AcKRB8wDpdaN6v6ru.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, 2018.

Pritam Sarkar, Sayna Ebrahimi, Ali Etemad, Ahmad Beirami, Sercan Ö Arik, and Tomas Pfister. Mitigating object hallucination via data augmented contrastive tuning. *CoRR*, 2024.

Yuying Shang, Xinyi Zeng, Yutao Zhu, Xiao Yang, Zhengwei Fang, Jingyuan Zhang, Jiawei Chen, Zinan Liu, and Yu Tian. From pixels to tokens: Revisiting object hallucinations in large vision-language models. *arXiv preprint arXiv:2410.06795*, 2024.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13088–13110, 2024.

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

Feilong Tang, Zile Huang, Chengzhi Liu, Qiang Sun, Harry Yang, and Ser-Nam Lim. Intervening anchor token: Decoding strategy in alleviating hallucinations for mllms. In *The Thirteenth International Conference on Learning Representations*, 2025a.

Feilong Tang, Chengzhi Liu, Zhongxing Xu, Ming Hu, Zile Huang, Haochen Xue, Ziyang Chen, Zelin Peng, Zhiwei Yang, Sijin Zhou, et al. Seeing far and clearly: Mitigating hallucinations in mllms with attention causal decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 26147–26159, 2025b.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9568–9578, 2024.

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

Zifu Wan, Ce Zhang, Silong Yong, Martin Q Ma, Simon Stepputtis, Louis-Philippe Morency, Deva Ramanan, Katia Sycara, and Yaqi Xie. Only: One-layer intervention sufficiently mitigates hallucinations in large vision-language models. *arXiv preprint arXiv:2507.00898*, 2025.

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. Mllm can see? dynamic correction decoding for hallucination mitigation. *arXiv preprint arXiv:2410.11779*, 2024a.

Jiaqi Wang, Yifei Gao, and Jitao Sang. Valid: Mitigating the hallucination of large vision language models by visual layer fusion contrastive decoding. *arXiv preprint arXiv:2411.15839*, 2024b.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, et al. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

Weixing Wang, Zifeng Ding, Jindong Gu, Rui Cao, Christoph Meinel, Gerard de Melo, and Haojin Yang. Image tokens matter: Mitigating hallucination in discrete tokenizer-based large vision-language models via latent editing. *arXiv preprint arXiv:2505.21547*, 2025.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. *arXiv preprint arXiv:2403.18715*, 2024c.

Sangmin Woo, Donguk Kim, Jaehyuk Jang, Yubin Choi, and Changick Kim. Don't miss the forest for the trees: Attentional vision calibration for large vision language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 1927–1951, 2025.

Jinge Wu, Yunsoo Kim, and Honghan Wu. Hallucination benchmark in medical visual question answering. *arXiv preprint arXiv:2401.05827*, 2024.

Le Yang, Ziwei Zheng, Boxu Chen, Zhengyu Zhao, Chenhao Lin, and Chao Shen. Nullu: Mitigating object hallucinations in large vision-language models via halluspace projection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14635–14645, 2025a.

Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. Understanding and mitigating hallucination in large vision-language models via modular attribution and intervention. In *International Conference on Learning Representations*, volume 2025, pp. 51546–51568, 2025b.

Zhihe Yang, Xufang Luo, Dongqi Han, Yunjian Xu, and Dongsheng Li. Mitigating hallucinations in large vision-language models via dpo: On-policy data hold the key. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10610–10620, 2025c.

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

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105, 2024.

Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12944–12953, 2024a.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13807–13816, 2024b.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019.

Kejia Zhang, Keda Tao, Jiasheng Tang, and Huan Wang. Poison as cure: Visual noise for mitigating object hallucinations in lvms. *arXiv preprint arXiv:2501.19164*, 2025.

Xingcheng Zhou, Mingyu Liu, Ekim Yurtsever, Bare Luka Zagar, Walter Zimmer, Hu Cao, and Alois C Knoll. Vision language models in autonomous driving: A survey and outlook. *arXiv preprint arXiv:2310.14414*, 2023a.

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

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

Xianwei Zhuang, Zhihong Zhu, Yuxin Xie, Liming Liang, and Yuexian Zou. Vasparse: Towards efficient visual hallucination mitigation via visual-aware token sparsification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4189–4199, 2025.

Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Kening Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, et al. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. *arXiv preprint arXiv:2410.03577*, 2024.

APPENDIX OVERVIEW

This appendix provides supplementary material organized as follows.

Section	Content
Scope and Impact	
Appendix A	Discusses the broader impact of this work.
Appendix B	Discusses the scope of our analysis and directions for future work.
Diagnostic Analysis	
Appendix C	Describes how the faithful and hallucinated cohorts are built and compared.
Appendix D	Compares the attribution before and after the final normalization.
Appendix E	Reports the diagnosis and its statistics on all four models.
Appendix F	Extends the attribution to single heads and relates it to hallucination heads.
Theory, Design, and Efficiency	
Appendix G	Derives the formal properties of CCD and compares it with layer-level methods.
Appendix H	Reports sensitivity to α and K.
Appendix I	Reports latency and memory.
Appendix J	Shows why a single α acts mostly on hallucinated objects.
Geometric and Token-Level Interpretations	
Appendix K	Places every word in the plane of its attention and FFN contributions.
Appendix L	Tracks the probabilities of hallucinated and faithful words across layers.
Appendix M	Compares the same word when it is faithful and when it is hallucinated.
Appendix N	Follows pairwise logit trajectories and the CCD correction in six cases.
Experimental Protocols and Extended Results	
Appendix O	Lists model architectures, baselines and hyperparameters.
Appendix P	Compares CCD with recent methods under the same protocol.
Appendix Q	Reports general perception and cognition on MME.
Appendix R	Evaluates CCD under beam search and nucleus sampling.
Appendix S	Shows how CCD changes the share of yes answers and of false yes answers on POPE.
Appendix T	Defines the evaluation metrics.
Appendix U	Shows example captions.
Appendix V	Discusses related work on interpreting transformer components.

A BROADER IMPACT

Object hallucination is a practical risk wherever MLLMs describe images to people who cannot check them, as in medical imaging, driving assistance and tools for visually impaired users. CCD reduces such errors at inference time without new data or retraining, so it can be added to a deployed model at little cost, and it leaves the model weights unchanged. It reduces hallucination but does not eliminate it, and outputs used in high-stakes settings still require human verification. The attribution we use is a general way to inspect what each sublayer contributes to a prediction, and it may help audit other failures of multimodal models.

B LIMITATIONS

Our analysis relies on the additive residual stream shared by current MLLMs, and we study four models from 4B to 8B parameters that differ in vision encoder, projector and language backbone. Extending it to larger models and to mixture-of-experts architectures is a natural next step. CCD has two hyperparameters that we set per model. Its results are stable across our grid (Appendix H), and a rule that sets them automatically would make deployment simpler still. Our evaluation centers on object hallucination in English, and other forms of hallucination and other languages are natural extensions.

C Cohort Construction and Statistical Tests

For each model, we caption 500 MSCOCO validation images (Lin et al., 2014) with greedy decoding and record the attention and FFN writes of every layer at every step. Each generated object word is labeled as faithful or hallucinated with the CHAIR pipeline (Rohrbach et al., 2018). The faithful and hallucinated cohorts are drawn from the same captions, cover a shared set of COCO categories and are equal in size, so that differences between them do not come from sample size. Contributions are summed over the last W layers, with $W = 5$ for the 32-layer models and $W = 3$ for the 36-layer Qwen3-VL models. This diagnostic window is separate from the hyperparameter K of CCD.

The two cohorts are independent by construction. We compare them with a two-sided Welch t-test, use the Mann-Whitney U test as a distribution-free check, and apply Holm correction across the four models within each sublayer. Because one caption can contain several object words, we also resample whole images rather than individual words, and every effect keeps its sign and significance under this image-level bootstrap.

D Before and After the Final Normalization

Direct logit attribution can be read before the final RMSNorm, through the projections δ of Eq. 6, or on the output logit, through the contributions $\tilde{\delta}$ of Section 3.2. Table 7 compares the two readings on LLaVA-1.5-7B.

Table 7: Mean contributions summed over the last five layers on LLaVA-1.5-7B. $\Delta = \mathrm{H} - \mathrm{F}$, so positive values favor hallucinated objects.

	Before normalization			**Output logit**		
	Faithful	Halluc.	Δ	Faithful	Halluc.	Δ
Attention	-0.73	-0.59	$+0.14$	$+19.62$	$+10.38$	-9.24
FFN	$+2.32$	$+6.51$	$+4.19$	-18.02	-4.45	$+13.57$

Before normalization, attention contributes almost equally to the two cohorts, while the FFN already gives hallucinated objects a larger contribution. On the output logit, which determines the prediction, attention favors faithful objects and the FFN favors hallucinated objects by a larger margin. The two readings differ because the gain γ weights hidden dimensions unequally. The FFN separates the cohorts in both readings, which is why CCD can operate on the projections δ, while the diagnosis is based on the output logit.

E Diagnosis Across Models

Table 8 extends the diagnosis to InstructBLIP-7B and the two Qwen3-VL models, which differ from LLaVA-1.5-7B in vision encoder, projector and language backbone (Appendix O.1).The pattern is the same in all four. Over the window, faithful tokens lose between 0.11 and 0.21 in probability on average and hallucinated tokens gain about as much, and on the output logit attention favors faithful objects while the FFN favors hallucinated objects by a larger margin. All eight differentials remain significant after Holm correction. In both Qwen3-VL models, the FFN contribution even changes sign between the cohorts, lowering faithful objects and raising hallucinated ones. Figures 4 and 5 show the full attribution for InstructBLIP-7B and Qwen3-VL-4B in the layout of Figure 1.

F Head-Level Attribution

The attention write of a layer is the sum of the outputs of its heads, so direct logit attribution applies to each head exactly. On LLaVA-1.5-7B, we rank all heads by how much more they contribute to hallucinated than to faithful words, and compare the ranking with the hallucination heads and non-hallucination heads released by Yang et al. (2025b), which were identified by attribution search. Table 9 shows that our ranking recovers both sets without any ablation or search.

Table 8: Diagnosis on four models over the last W layers, with $W = 5$ for the 32-layer models and $W = 3$ for the 36-layer Qwen3-VL models. n is the size of each balanced cohort. Drift is the mean change in the probability of the generated token from the first to the last of these layers, as shown in the insets of Figures 1, 4 and 5. The attention and FFN columns give mean contributions to the output logit for faithful (F) and hallucinated (H) tokens, and $\Delta = \mathrm{H} - \mathrm{F}$ with its 95% confidence interval.

			Drift	Attention		FFN	
Model	n	W	F / H	F / H	Δ [95% CI]	F / H	Δ [95% CI]
LLaVA-1.5-7B	908	5	-0.112 / $+0.107$	19.62 / 10.38	-9.24 [-10.29, -8.17]	-18.02 / -4.45	$+13.57$ [12.21, 14.94]
InstructBLIP-7B	529	5	-0.200 / $+0.207$	21.23 / 10.23	-11.00 [-11.88, -10.14]	-18.67 / -2.13	$+16.54$ [15.51, 17.62]
Qwen3-VL-4B	81	3	-0.207 / $+0.214$	15.12 / 11.50	-3.62 [-6.31, -0.93]	-5.10 / $+1.21$	$+6.31$ [3.92, 8.75]
Qwen3-VL-8B	54	3	-0.184 / $+0.195$	17.34 / 10.11	-7.23 [-13.00, -1.46]	-6.18 / $+3.18$	$+9.20$ [3.39, 15.01]

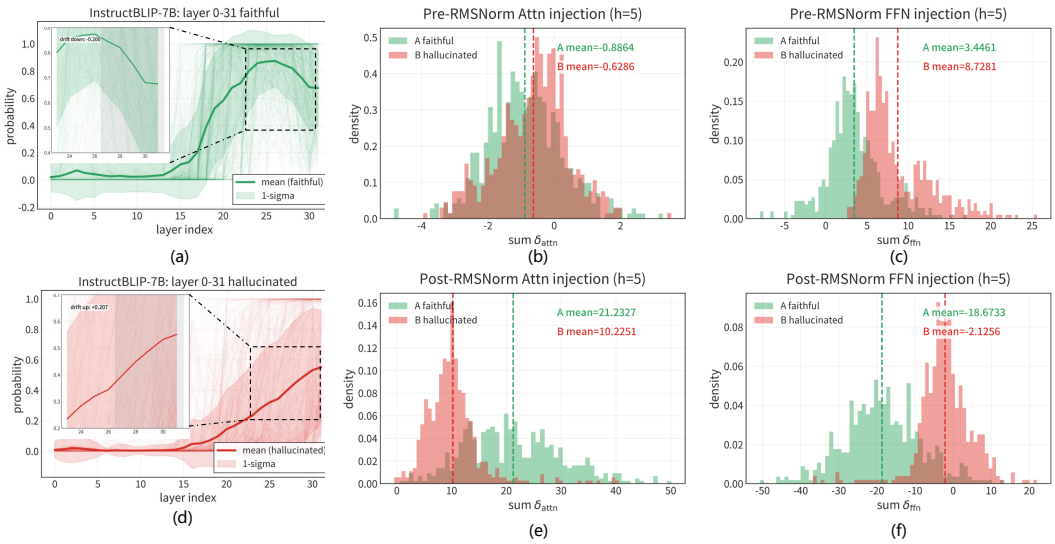


Figure 4: Direct logit attribution on InstructBLIP-7B, in the layout of Figure 1.

Table 9: Agreement between our head ranking on LLaVA-1.5-7B and the head sets of Yang et al. (2025b). Percentiles run from the most hallucination-favoring head (0%) to the most faithful-favoring head (100%).

Test	**Ours**	**Chance**
Hallucination heads in our most hallucination-favoring 5%	12 / 20	1 / 20
Non-hallucination heads in our most faithful-favoring 10%	12 / 20	2 / 20
Median percentile of the two sets	1.7% / 95.7%	50%

The two findings are compatible. Nearly half of all heads lean toward hallucinated objects, but together they carry much less weight than the heads leaning toward faithful objects. The layer sum therefore favors faithful objects, while the strongest heads of the first group are those that Yang et al. (2025b) select.

G Theoretical Analysis of CCD

Eq. 8 gives the effect of CCD on the log-odds between two candidates at the same step. This appendix derives three further consequences and restates the comparison with layer-level methods of Section 3.4 in the same terms. All results follow from the additivity of the correction and the softmax. The only empirical premise is the pairwise condition of Definition 1, the form of FFN overcommitment that CCD acts on. Appendix J shows that it holds on average on LLaVA-1.5-7B, where Φ^+ is 38% larger for hallucinated tokens.

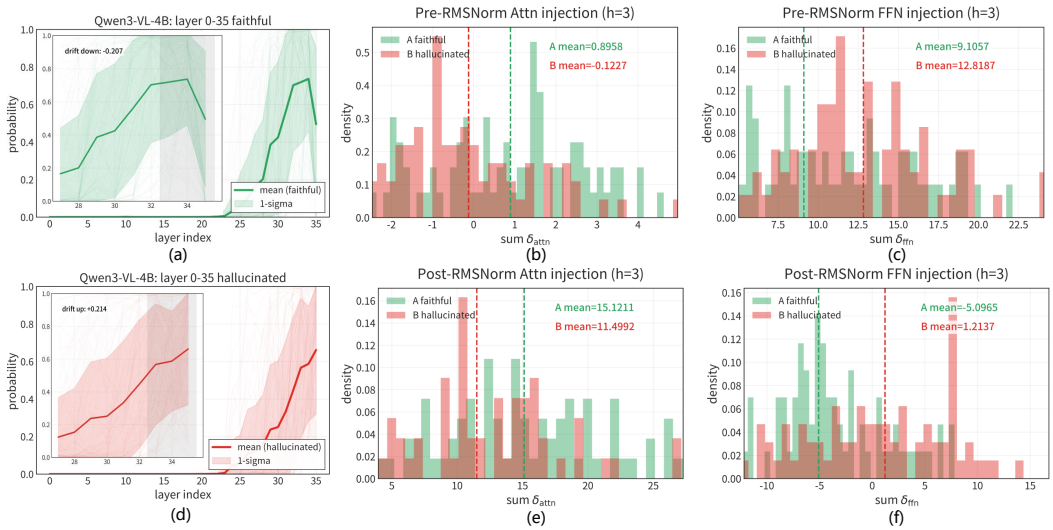


Figure 5: Direct logit attribution on Qwen3-VL-4B, in the layout of Figure 1. On the output logit, the FFN lowers faithful objects and raises hallucinated ones.

G.1 Setup

Consider one decoding step with logits z_y, distribution $p(y) = \mathrm{softmax}(\mathbf{z})_y$, and CCD-corrected logits $\hat{z}_y = z_y - \alpha\,\Phi^+(y)$ with distribution $\hat{p}$ (Eq. 7). Alongside Φ^+, we define the negative part of the late FFN writes and their total,

$$\Phi^-(y) = \sum_{k=L-K+1}^{L} \min\big(0,\, \delta_k^{\mathrm{ffn}}(y)\big), \qquad \Phi(y) = \Phi^+(y) + \Phi^-(y). \tag{12}$$

Definition 1 (FFN overcommitment)**.** *A word y_h exhibits FFN overcommitment relative to a word y_f if $\Phi^+(y_h) > \Phi^+(y_f)$. The overcommitment gap is $\Delta\Phi = \Phi^+(y_h) - \Phi^+(y_f)$.*

Under this condition, Eq. 8 raises the log-odds of y_f over y_h by exactly $\alpha\,\Delta\Phi$.

G.2 Rank reversal

Corollary 1 (Rank reversal)**.** *Let $m = z_{y_h} - z_{y_f}$ be the logit margin by which y_h leads before correction. If $\alpha\,\Delta\Phi > m$, then $\hat{z}_{y_f} > \hat{z}_{y_h}$.*

Proof. The corrected difference is $\hat{z}_{y_f} - \hat{z}_{y_h} = -m + \alpha\,\Delta\Phi$, which is positive whenever $\alpha\,\Delta\Phi > m$.

A hallucinated word is therefore easiest to overturn when it combines a large overcommitment gap with a small margin, which is where hallucinated words concentrate (Appendix J).

G.3 Which words lose probability

Proposition 1. *Let $\hat{Z}_\alpha = \sum_{y'} p(y') \exp\big(-\alpha\,\Phi^+(y')\big)$. For any word y,*

$$\hat{p}(y) < p(y) \quad \Longleftrightarrow \quad \Phi^+(y) > -\tfrac{1}{\alpha}\log \hat{Z}_\alpha, \tag{13}$$

and $-\frac{1}{\alpha}\log \hat{Z}_\alpha \le \mathbb{E}_p[\Phi^+]$. Every word whose positive late FFN write exceeds its expectation under p therefore loses probability.

Proof. Writing the corrected distribution through the original one gives $\hat{p}(y) = p(y)\exp\big(-\alpha\,\Phi^+(y)\big)/\hat{Z}_\alpha$, so $\hat{p}(y) < p(y)$ exactly when $\exp\big(-\alpha\,\Phi^+(y)\big) < \hat{Z}_\alpha$. Taking logarithms gives Eq. 13. The bound follows from Jensen's inequality, $\hat{Z}_\alpha = \mathbb{E}_p\big[\exp(-\alpha\,\Phi^+)\big] \ge \exp\big(-\alpha\,\mathbb{E}_p[\Phi^+]\big)$.

G.4 WHY ONLY THE POSITIVE PART

An alternative correction subtracts the whole late FFN write, $\hat{z}_y^{\text{full}} = z_y - \alpha\,\Phi(y)$. Since $\Phi^-(y) \le 0$, this adds $-\alpha\,\Phi^-(y) \ge 0$ back to each logit and cancels the suppressive part of the FFN write.

Proposition 2. *If a hallucinated word receives stronger suppressive FFN writes than a faithful word,*

$$\Phi^-(y_h) < \Phi^-(y_f) \le 0, \tag{14}$$

then CCD yields a strictly larger log-odds of y_f over y_h than subtracting the whole FFN write.

Proof. As in Eq. 8, the normalizer cancels in the ratio, and the two corrections differ only by the negative part, giving

$$\log \frac{\hat{p}^{\,\text{full}}(y_f)}{\hat{p}^{\,\text{full}}(y_h)} - \log \frac{\hat{p}(y_f)}{\hat{p}(y_h)} = \alpha\big(\Phi^-(y_h) - \Phi^-(y_f)\big), \tag{15}$$

which is negative under Eq. 14.

Within one FFN window, some layers can promote a word while others push it down (Geva et al., 2021; 2023; Dai et al., 2022). Subtracting the whole write removes these internal vetoes along with the promotion, while CCD keeps them. Section 4.3 compares the two rules empirically.

G.5 LAYER-LEVEL METHODS

By Eq. 9, the late update splits exactly into an attention part $A_m(y) = \sum_{\ell>m} \tilde{\delta}_\ell^{\text{attn}}(y)$, an FFN part $F_m(y) = \sum_{\ell>m} \tilde{\delta}_\ell^{\text{ffn}}(y)$ and a rescaled copy of the intermediate logits, $u_y^{(m)} = A_m(y) + F_m(y) - (1 - s_m/s)\, z_y^{(m)}$. Since the normalizers cancel in each ratio, Eqs. 10 and 11 give

$$\log \frac{p^{\text{DoLa}}(y_f)}{p^{\text{DoLa}}(y_h)} = \big[A_M(y_f) - A_M(y_h)\big] + \big[F_M(y_f) - F_M(y_h)\big] - \Big(1 - \frac{s_M}{s}\Big)\big[z_{y_f}^{(M)} - z_{y_h}^{(M)}\big], \tag{16}$$

$$\log \frac{p^{\text{DeCo}}(y_f)}{p^{\text{DeCo}}(y_h)} = \Big(\beta + \frac{s_{m^*}}{s}\Big)\big[z_{y_f}^{(m^*)} - z_{y_h}^{(m^*)}\big] + \big[A_{m^*}(y_f) - A_{m^*}(y_h)\big] + \big[F_{m^*}(y_f) - F_{m^*}(y_h)\big], \tag{17}$$

where DoLa is restricted to its plausible candidates. In both, the intermediate prediction is reweighted as a whole. On average over the cohorts, the attention bracket is positive and the FFN bracket negative in the late layers. Both methods weight the two brackets equally, so neither can raise the log-odds of the faithful word through one bracket without the other. CCD changes the log-odds only by $\alpha\,\Delta\Phi$, which comes from the positive FFN writes of the last K layers and leaves the attention terms and the negative FFN terms in place.

H HYPERPARAMETER SENSITIVITY

Table 10 varies α with K fixed and K with α fixed on LLaVA-1.5-7B. Every setting in the grid improves on greedy decoding in all three CHAIR metrics, and the results change little around the chosen values. Larger α gradually costs recall, and larger K extends the correction into layers where faithful and hallucinated objects receive similar writes. The values used for each model are listed in Appendix O.2.

I COMPUTATIONAL OVERHEAD

CCD reuses the writes computed by the forward pass, so its only extra work at each step is projecting the late FFN writes through the unembedding matrix, clipping them and subtracting a weighted sum. DoLa instead computes a divergence against several candidate layers, and DeCo projects an additional intermediate hidden state. Table 11 reports the resulting cost.

Table 10: Hyperparameter sensitivity on CHAIR with LLaVA-1.5-7B. The setting used in the main experiments is highlighted.

α	K	$\text{CHAIR}_S \downarrow$	$\text{CHAIR}_I \downarrow$	Recall↑
Fixed $K = 3$, varying α				
0.05	3	44.40	13.36	75.36
0.10	**3**	**43.80**	**13.37**	**76.06**
0.15	3	44.20	13.59	75.43
0.20	3	45.20	13.99	73.72
0.25	3	45.80	14.35	73.22
Fixed $\alpha = 0.10$, varying K				
0.10	2	44.40	14.00	75.36
0.10	**3**	**43.80**	13.37	**76.06**
0.10	4	44.60	14.03	75.11
0.10	5	45.20	**13.00**	75.05
0.10	6	45.60	13.16	74.17

Table 11: Per-token latency and peak GPU memory on LLaVA-1.5-7B, measured on a single NVIDIA RTX 3090 and averaged over the 500 CHAIR captions.

Method	**ms / token**	**Δ (ms)**	**Δ (%)**	**Peak GPU (MB)**
Greedy	52.6	—	—	14 716
DoLa	63.5	+10.9	+20.7%	14 704
DeCo	56.4	+3.8	+7.2%	14 744
CCD (ours)	53.2	+0.6	+1.1%	14 716

J Implicit Selectivity

CCD applies the same α to every word in the vocabulary, yet it acts mostly on hallucinated objects. Figure 6 follows this from the writes to the decisions on LLaVA-1.5-7B. Hallucinated words receive larger positive late FFN writes than faithful words, so the same correction lowers them more, and CCD changes the prediction about six times as often when the original choice is hallucinated as when it is faithful. Figure 7 shows why the difference in decisions exceeds the difference in writes. Faithful words usually lead by wide margins and absorb the correction, while hallucinated words combine large writes with narrow margins, the region where Corollary 1 predicts a reversal.

K Attention and FFN Contributions per Word

Figure 8 places every word of the two LLaVA-1.5-7B cohorts in a plane whose axes are its attention and FFN contributions over the last three layers, read before the final normalization as in CCD (Eq. 6). The two cohorts overlap along the attention axis in this reading, while hallucinated words sit higher along the FFN axis, and many fall outside the region occupied by faithful words. CCD moves each word straight down, by an amount that grows with its positive FFN write, so the words displaced furthest are corrected most. A layer-level method instead rescales the late update, which moves each word along the line through the origin and changes its attention coordinate together with its FFN coordinate.

L Layer-Wise Trajectories of Individual Words

Figures 9 and 10 track, layer by layer, the probability of a hallucinated word and of a faithful competitor at the same position. Both stay near zero through the early and middle layers, and the hallucinated word gains its probability only in the last few layers. In the cases of Figure 10, the faithful word is ahead for several layers before this late rise. For example, “banana” leads through the middle layers before “apple” overtakes it, and “broccoli” holds most of the probability until “carrot”

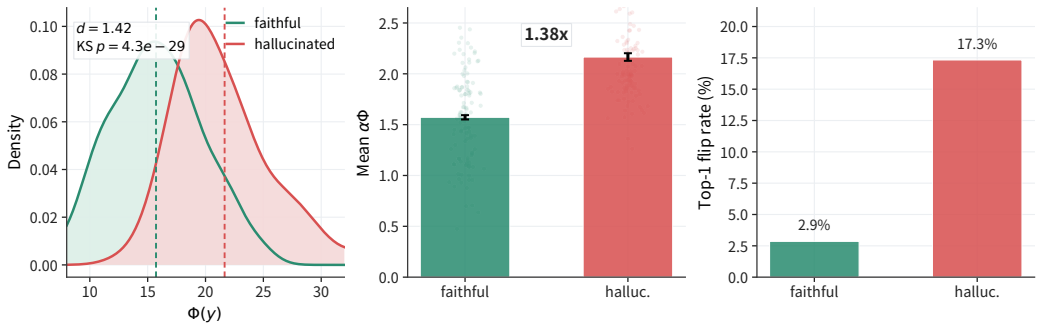


Figure 6: Implicit selectivity on LLaVA-1.5-7B. Left: distribution of the positive late FFN write Φ^+. Center: mean suppression $\alpha\,\Phi^+$. Right: rate at which CCD changes the top-1 prediction.

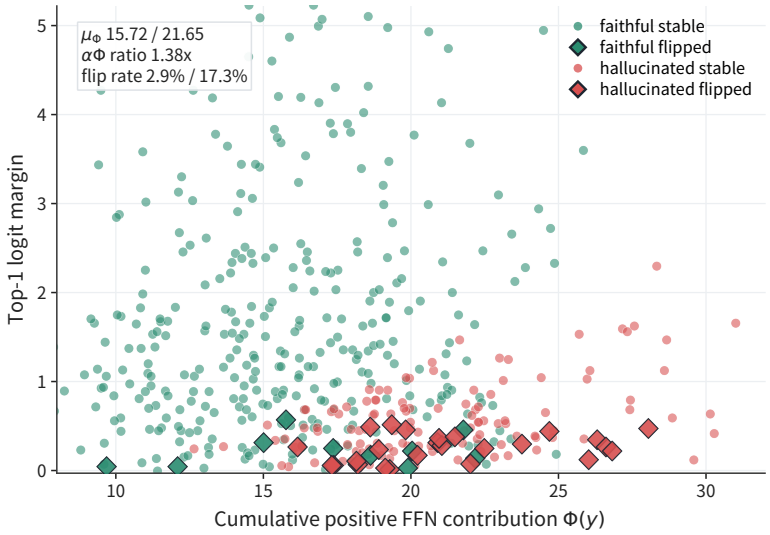


Figure 7: Positive late FFN write Φ^+ against the top-1 logit margin on LLaVA-1.5-7B. Hallucinated words cluster at large Φ^+ and small margins, where CCD changes the prediction.

replaces it in the final layers. The model has already ranked the correct object first by the time the late layers change its decision.

M The Same Word as Faithful and Hallucinated

A late FFN surge could in principle belong to certain words rather than to the situations in which they are hallucinated. To separate the two, we take eight object categories and, for each, one caption in which the model names the object correctly and one in which it names the object although the image does not contain it. Figures 11 and 12 show the per-layer attention and FFN contributions to the logit of the same word in both cases. When the object is present, the FFN contributions grow through the middle layers and level off, with no late surge. When the object is absent, the FFN adds a sharp late surge while attention flattens or pulls the logit down. The same pattern holds in all eight categories, so the surge belongs to the context in which a word is hallucinated rather than to the word itself.

N Pairwise Logit Trajectories and the CCD Correction

Figure 13 follows six hallucination cases on LLaVA-1.5-7B, the first of which appears in Figure 3b. Each panel plots the logit of the hallucinated word against that of the faithful word at every layer,

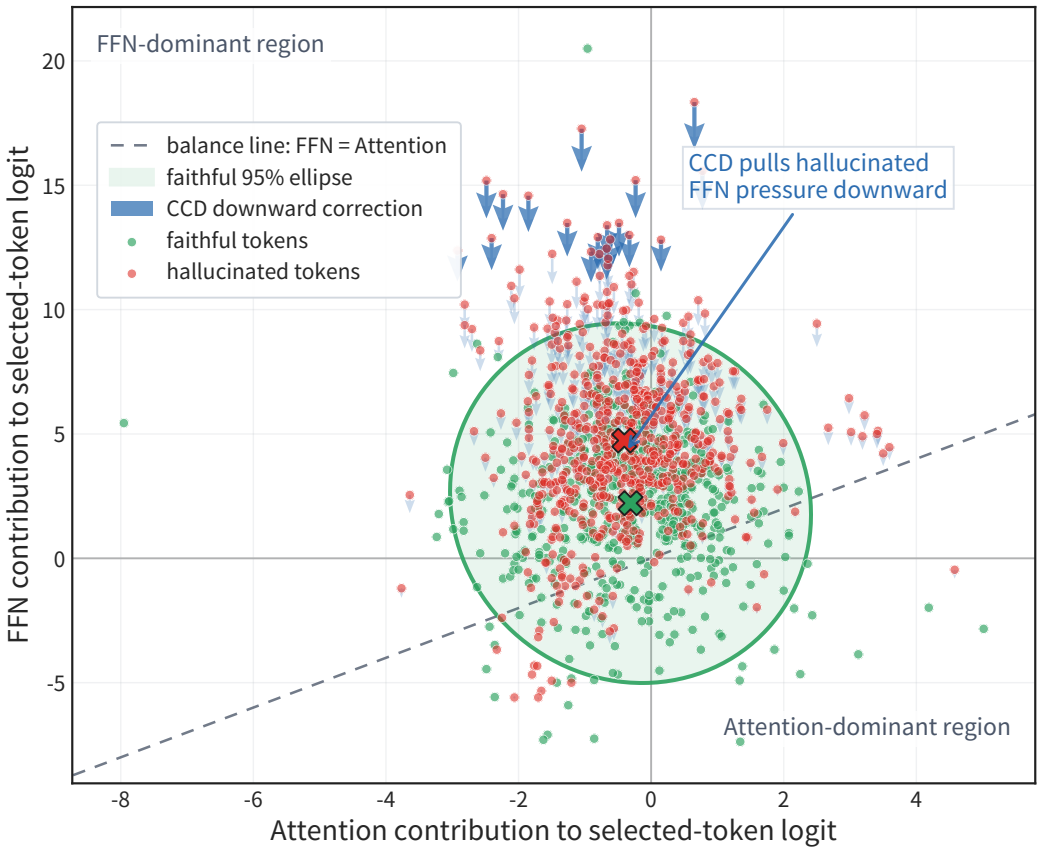


Figure 8: Attention and FFN contributions over the last three layers, before the final normalization, for 908 faithful and 908 hallucinated words on LLaVA-1.5-7B. The ellipse encloses 95% of faithful words, and the arrows show the CCD correction.

so that the diagonal is the decision boundary. The trajectories stay in or near the region where the faithful word wins through the middle layers and cross the boundary in the last few layers, moving mostly upward, as the hallucinated word gains while the faithful word changes little. The gold arrows show the positive FFN terms that CCD subtracts at each of the last three layers. Together they move the trajectory back below the boundary in five of the six cases, restoring the faithful word as the prediction.

O Experimental Details

O.1 Models

Table 12 summarizes the four models. LLaVA-1.5-7B and InstructBLIP-7B pair a frozen CLIP-family vision encoder with a Vicuna decoder and differ in the projector, a two-layer MLP for LLaVA and a Q-Former (Li et al., 2023a) for InstructBLIP. The Qwen3-VL models use a vision encoder trained jointly with the language model and 36-layer Qwen3 backbones. All four decoders are pre-normalization transformers with RMSNorm.

O.2 Baselines and implementation

We compare against nine training-free methods in three families: contrastive decoding with ICD (Wang et al., 2024c), VCD (Leng et al., 2024), DoLa (Chuang et al., 2023) and DeCo (Wang et al., 2024a); attention intervention with OPERA (Huang et al., 2024), AGLA (An et al., 2025) and ONLY (Wan et al., 2025); and representation correction with MEMVR (Zou et al., 2024) and

Table 12: Architectures of the evaluated MLLMs.

Model	Vision Encoder	Projection	LLM Backbone	Layers	Hidden Dim	Params
LLaVA-1.5-7B	CLIP ViT-L/14@336	2-layer MLP	Vicuna-7B v1.5	32	4096	7B
InstructBLIP-7B	EVA-CLIP ViT-G/14	Q-Former	Vicuna-7B v1.1	32	4096	7B
Qwen3-VL-4B	ViT-600M (native)	MLP	Qwen3-4B	36	2560	4B
Qwen3-VL-8B	ViT-600M (native)	MLP	Qwen3-8B	36	4096	8B

VAF (Yin et al., 2025). Each baseline uses its officially released code and reported best hyperparameters. Unless a method prescribes otherwise, decoding is greedy with up to 512 new tokens for captioning. CCD uses $(\alpha, K) = (0.10, 3)$ for LLaVA-1.5-7B, $(0.01, 5)$ for InstructBLIP-7B and $(0.05, 3)$ for both Qwen3-VL models.

P Comparison with Recent Methods

We reproduce AD-HH (Yang et al., 2025b), ACG (Jo et al., 2026) in both of its modes, and FLB (Ha et al., 2026) on LLaVA-1.5-7B, each at its published settings. ACG and FLB share the checkpoint, CHAIR images, POPE splits, greedy decoding and evaluator of our main experiments. AD-HH is run in the protocol of its authors with their released head list, which closely reproduces their published results. Table 13 reports the change of each method relative to greedy decoding in the same run.

Table 13: Change relative to greedy decoding on CHAIR with LLaVA-1.5-7B.

Method	$\Delta\text{CHAIR}_S \downarrow$	$\Delta\text{Recall} \uparrow$
AD-HH (Yang et al., 2025b)	-17.80	-5.04
ACG-Fast (Jo et al., 2026)	-11.60	-5.29
ACG-Full (Jo et al., 2026)	-23.20	-8.32
FLB (Ha et al., 2026)	-16.00	-5.11
CCD (ours)	-10.40	$+3.60$

Every method that lowers CHAIR_S more than CCD also lowers recall, while CCD raises it. On POPE, AD-HH and FLB leave almost every answer unchanged, and neither mode of ACG moves the average F1 by more than one point, whereas CCD improves F1 on every split (Table 5). AD-HH deactivates a head only when it attends strongly to the text generated so far, which rarely happens before a one-word answer, and the boost of FLB is zero at the first generated token, where a POPE answer is decided. CCD corrects the logits at every step, including the first.

Q General Capability on MME

Table 14 reports MME, which tests perception and cognition beyond hallucination. CCD obtains the highest total score on every model, so removing the positive late FFN contributions does not reduce general capability. On LLaVA-1.5-7B its cognition score falls slightly while its perception score rises.

R Decoding Strategy

CCD modifies the logits before a token is selected, so it applies to any decoding rule. Table 15 repeats the CHAIR comparison under beam search and nucleus sampling. Under both, CCD again reduces hallucination while raising recall, whereas DoLa and DeCo lower recall.

S Answer Balance on POPE

A higher POPE F1 can come from a shift in how often the model answers yes rather than from better judgments about individual objects. Each POPE split contains 1,500 questions about objects that are

Table 14: **MME benchmark** results. Best in **bold**, second best underlined.

	LLaVA-1.5-7B			InstructBLIP-7B			Qwen3-VL-4B			Qwen3-VL-8B		
Method	Per	Cog	Total	Per	Cog	Total	Per	Cog	Total	Per	Cog	Total
Greedy	1380.5	298.2	1678.7	947.1	213.9	1161.0	1688.3	583.4	2271.7	1717.1	639.3	2356.4
+ VCD	1418.3	**302.4**	1720.7	1048.7	248.6	1297.3	1689.5	586.5	2276.0	1734.2	648.2	2382.4
+ VAF	1418.9	299.8	1718.7	1065.4	253.1	1318.5	1690.2	590.2	2280.4	1737.5	650.1	2387.6
+ DoLa	1390.2	295.8	1686.0	985.6	226.8	1212.4	1643.7	**620.7**	2264.4	1679.6	652.5	2332.1
+ DeCo	1415.5	296.0	1711.5	1012.3	235.2	1247.5	1686.0	604.3	2290.3	1731.9	650.0	2381.9
+ CCD (ours)	**1459.2**	293.6	**1752.7**	**1116.4**	**270.4**	**1386.8**	**1707.6**	617.5	**2325.1**	**1749.8**	**655.0**	**2404.8**

Table 15: CHAIR on LLaVA-1.5-7B under different decoding rules. Best in **bold**, second best underlined within each group.

Decoding	**Method**	$\text{CHAIR}_S \downarrow$	$\text{CHAIR}_I \downarrow$	Recall↑	ΔCH_S
Greedy decoding					
Greedy	Vanilla	54.2	18.5	72.5	—
Greedy	+ DoLa	48.4	15.9	69.3	−5.8
Greedy	+ DeCo	**41.6**	13.4	69.7	−12.6
Greedy	**+ CCD (ours)**	43.8	**13.3**	**76.1**	−10.4
Beam search (b=3)					
Beam	Vanilla	52.9	17.8	70.9	—
Beam	+ DoLa	47.1	15.2	67.4	−5.8
Beam	+ DeCo	**39.7**	**12.6**	67.9	−13.2
Beam	**+ CCD (ours)**	42.2	12.8	**74.3**	−10.7
Nucleus sampling (p=0.9)					
Nucleus	Vanilla	56.7	19.8	74.1	—
Nucleus	+ DoLa	51.0	17.2	70.5	−5.7
Nucleus	+ DeCo	46.0	14.9	71.5	−10.7
Nucleus	**+ CCD (ours)**	**45.2**	**14.1**	**76.7**	−11.5

present and 1,500 about objects that are absent, so the share of yes answers and the rate at which absent objects are affirmed both follow from the precision P and recall R in Table 5,

$$\text{Yes} = \frac{R}{2P}, \qquad \text{False yes} = R\Big(\frac{1}{P} - 1\Big), \tag{18}$$

where False yes is the fraction of absent objects answered with yes. Table 16 reports both for greedy decoding and CCD.

Greedy decoding is biased in opposite directions on different models. InstructBLIP-7B answers yes too often, up to 60.4% of the time on the Adversarial split, while the two Qwen3-VL models answer yes to only 42.8% to 47.0% of the questions. CCD moves the share of yes answers toward 50% in 11 of the 12 settings and lowers its mean distance from 50% from 5.4 to 2.0 points, so its gain on POPE does not come from answering either yes or no more often. The exception is LLaVA-1.5-7B on the Adversarial split, where greedy decoding is already balanced and CCD raises the share of yes answers to 55.1%. The false yes rate still falls there, from 24.1% to 23.5%, so the additional yes answers go to objects that are present.

The false yes rate separates two sources of gain. On LLaVA-1.5-7B and InstructBLIP-7B, CCD lowers it on every split, and on InstructBLIP-7B by 14 to 16 points. The two Qwen3-VL models rarely affirm absent objects under greedy decoding, and most of their errors are present objects answered with no. CCD recovers many of these and raises recall by up to 13.7 points, while the false yes rate rises slightly on the Random and Popular splits and falls on the Adversarial split. On these models, the POPE gain comes mainly from present objects that CCD no longer denies.

Table 16: Share of questions answered yes (Yes) and fraction of absent objects answered yes (False yes) on POPE, in percent, computed from Table 5 with Eq. 18. A balanced model answers yes to 50% of the questions.

	Random		Popular		Adversarial	
Method	Yes	False yes	Yes	False yes	Yes	False yes
LLaVA-1.5-7B						
Greedy	43.2	9.6	46.5	16.3	50.5	24.1
+ CCD (ours)	47.3	8.1	49.4	12.2	55.1	23.5
InstructBLIP-7B						
Greedy	52.7	22.2	56.6	30.3	60.4	37.7
+ CCD (ours)	48.8	8.2	50.7	15.1	53.8	21.6
Qwen3-VL-4B						
Greedy	43.0	2.8	43.5	5.9	45.0	9.0
+ CCD (ours)	49.0	5.0	46.1	6.1	47.5	8.9
Qwen3-VL-8B						
Greedy	42.8	4.4	44.9	8.7	47.0	12.8
+ CCD (ours)	51.2	7.5	50.1	10.7	49.4	12.0

Because CCD lowers whichever answer token received the larger positive late FFN contribution (Eq. 8), the direction of the shift is not fixed in advance. That it points toward balance both for a model biased toward yes and for models biased toward no suggests that, in the yes-or-no setting as well, the late FFN over-promotes the answer the model would give by default.

T Evaluation Metrics

CHAIR. CHAIR (Rohrbach et al., 2018) compares the objects mentioned in a caption with the MSCOCO annotations. For image i, let $\mathcal{O}_i$ be the set of objects extracted from its caption and mapped to MSCOCO categories, and $\mathcal{G}_i$ its annotated objects. Over M captions,

$$\mathrm{CHAIR}_S = \frac{\left|\{i : \mathcal{O}_i \not\subseteq \mathcal{G}_i\}\right|}{M}, \qquad \mathrm{CHAIR}_I = \frac{\sum_i |\mathcal{O}_i \setminus \mathcal{G}_i|}{\sum_i |\mathcal{O}_i|}, \qquad \mathrm{Recall} = \frac{\sum_i |\mathcal{O}_i \cap \mathcal{G}_i|}{\sum_i |\mathcal{G}_i|}. \tag{19}$$

CHAIR_S is the fraction of captions with at least one hallucinated object, CHAIR_I the fraction of mentioned objects that are hallucinated, and Recall the fraction of annotated objects that are mentioned.

POPE. POPE (Li et al., 2023c) asks "Is there a {object} in the image?" and scores yes-or-no answers. Absent objects are drawn uniformly in the Random split, from the most frequent MSCOCO categories in the Popular split, and from the categories that most often co-occur with the present objects in the Adversarial split. Each split contains 3,000 balanced questions, and we report accuracy, precision, recall and F1.

AMBER. AMBER (Wang et al., 2023) evaluates 1,004 images with a generative task, scored by CHAIR, Cover, Hal and Cog, and a discriminative task, scored by accuracy, precision, recall and F1. We summarize both with the signed total

$$\text{Total Score} = \text{Cover} + \text{Acc} + \text{Prec} + \text{Rec} + \text{F1} - \text{CHAIR} - \text{Hal} - \text{Cog}. \tag{20}$$

MME. MME (Fu et al., 2023) measures general perception and cognition over 14 subtasks. Each image comes with two questions whose correct answers are opposite, and each subtask is scored out of 200. We report the Perception, Cognition and total scores.

U Qualitative Examples

Figures 14 and 15 compare captions from greedy decoding and CCD on LLaVA-1.5-7B. In each example, CCD removes the hallucinated objects, such as the skis in the snow scene and the chair and

couch in the dog scene, and keeps the objects that are present. Its captions are about as long as the originals, and some add details that greedy decoding missed, such as the dog's collar.

V More Related Work

Mechanistic interpretability of transformer components. Our analysis builds on work revealing the functional specialization of transformer sublayers. FFN layers act as key-value memories that store factual associations (Geva et al., 2021) and promote concepts in the vocabulary (Geva et al., 2022), localizable to specific layers (Meng et al., 2022) and individual neurons (Dai et al., 2022), with recall mechanisms that can be dissected step by step (Geva et al., 2023; Hernandez et al., 2023). The residual stream formalization (Elhage et al., 2021) treats the output of each sublayer as an additive contribution to a shared channel, and the logit lens and its variants (nostalgebraist, 2020; Belrose et al., 2023) project intermediate hidden states through the unembedding matrix to inspect evolving predictions. Circuit-level analyses (Olsson et al., 2022; Conmy et al., 2023; Dar et al., 2023) identify specific computational motifs behind downstream behaviors. Projecting the write of each sublayer, rather than a whole hidden state, gives the direct logit attribution we use. What we add is a contrast between faithful and hallucinated objects, read separately by sublayer, sign and depth, which reveals the opposing roles of the two sublayers and prescribes the design of CCD.

Component-level analysis of hallucination in MLLMs. Causal tracing, which measures a different quantity, likewise implicates deep-layer FFNs at the final token in representing objects (Li et al., 2026). Modular attribution identifies hallucination heads, attention heads that favor hallucinated objects (Yang et al., 2025b), which may seem at odds with our finding that attention as a whole favors faithful objects. Extended to single heads, our attribution recovers both of their head sets, and the two findings describe the same distribution at different resolutions (Appendix F).

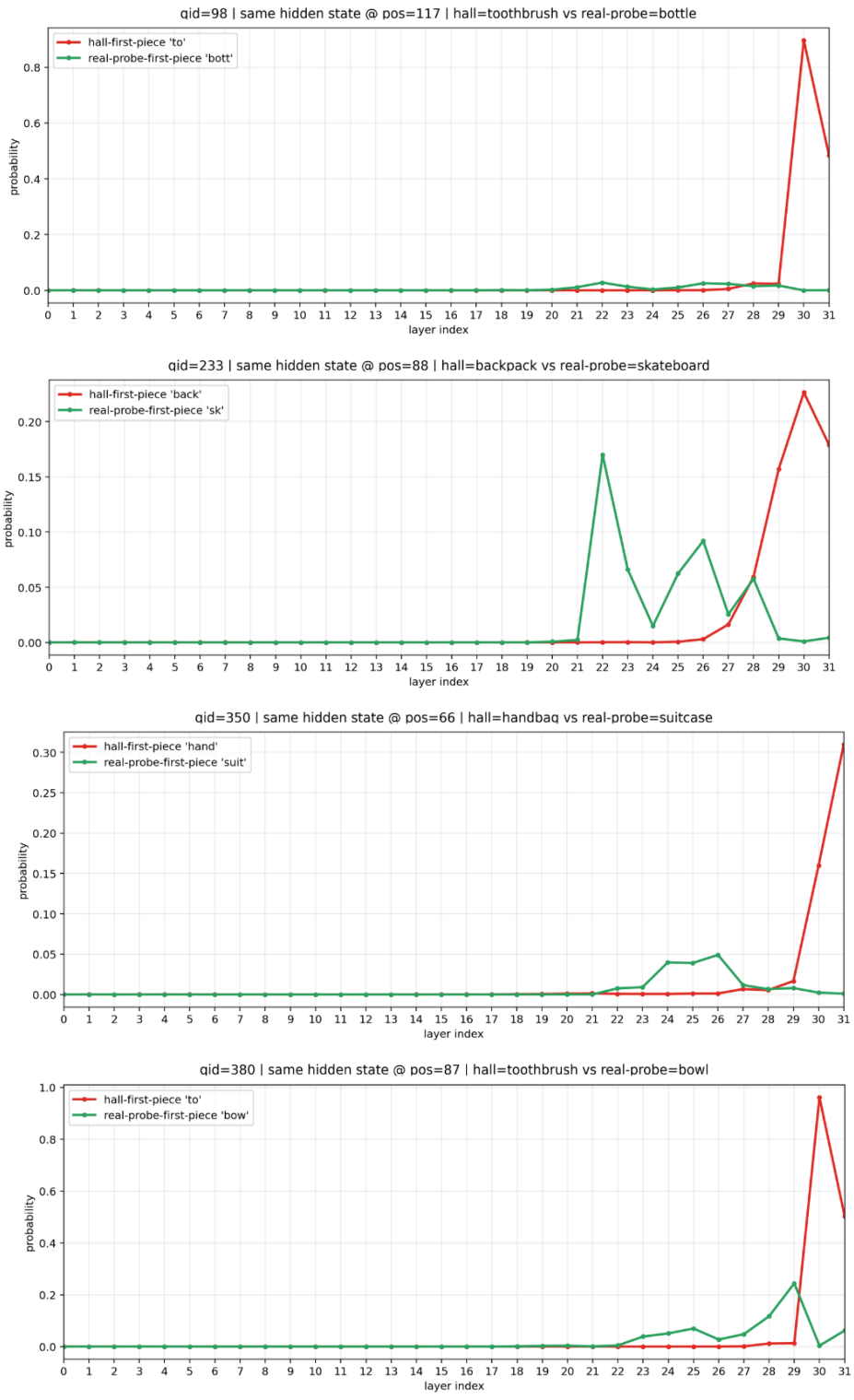


Figure 9: Layer-wise probabilities of a hallucinated word and a faithful competitor at the same position, for four cases on LLaVA-1.5-7B. The hallucinated word gains its probability in the last layers.

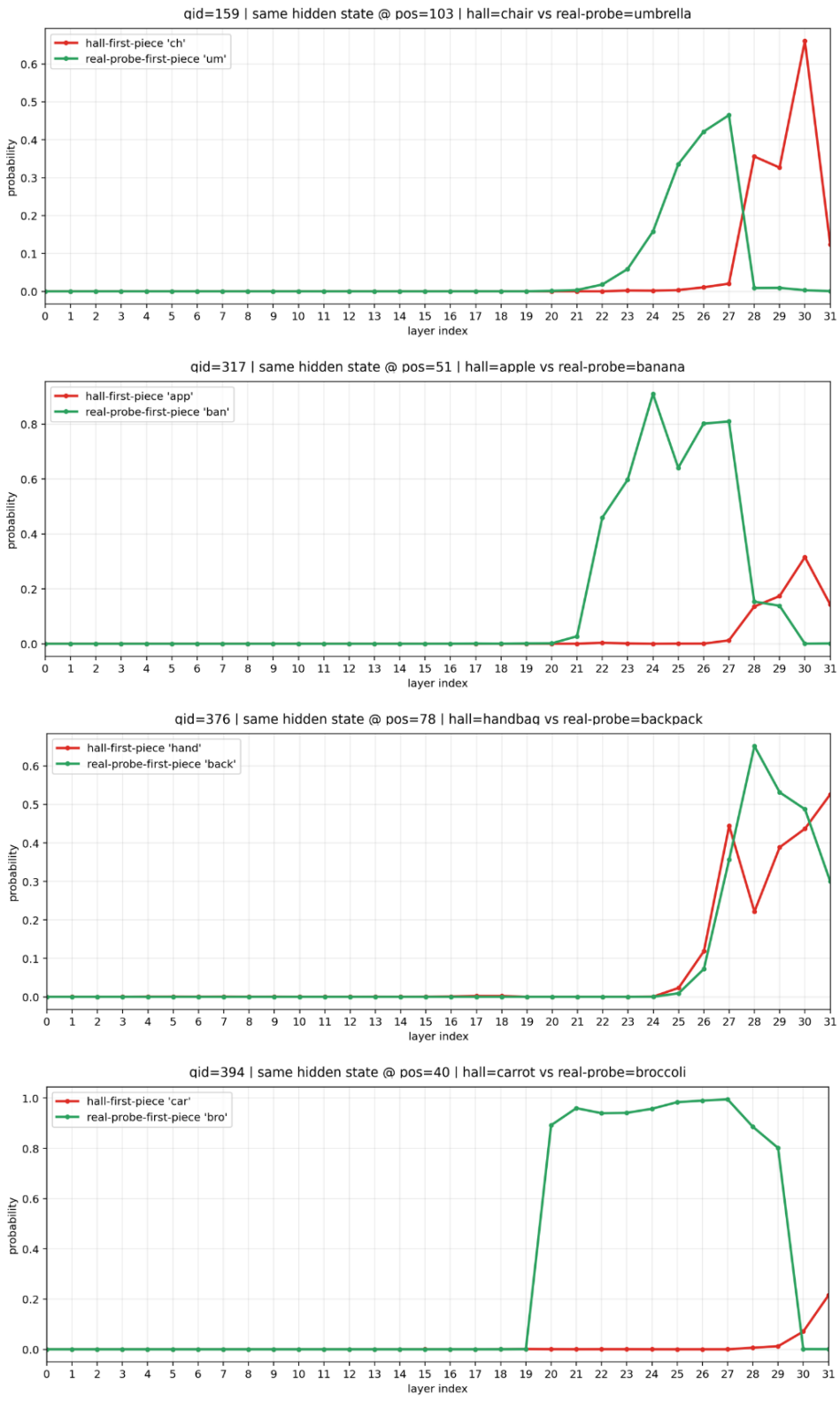


Figure 10: Cases in which the faithful word leads through the middle layers and is overtaken in the last layers.

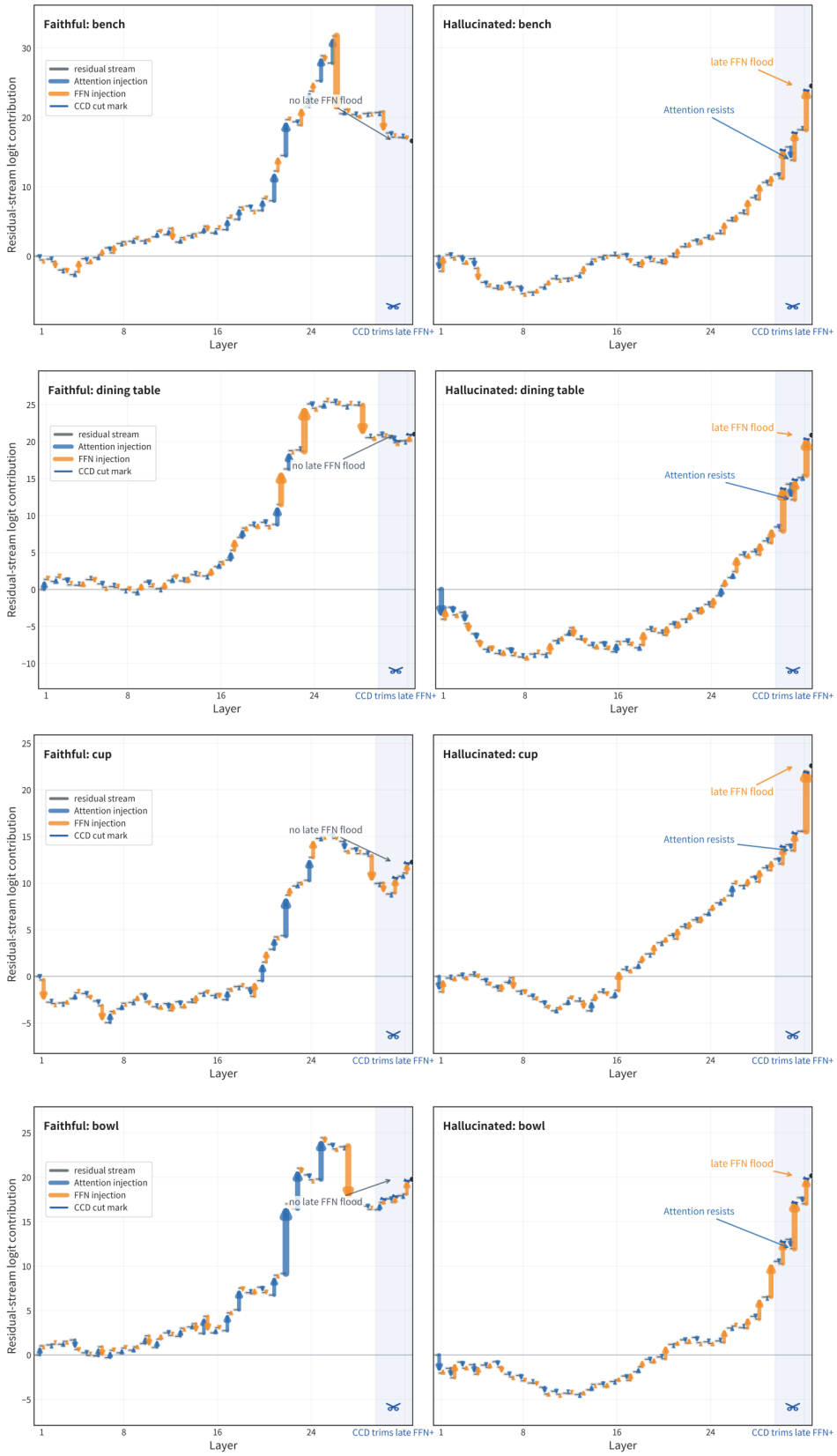


Figure 11: Per-layer attention and FFN contributions to the logit of the same word when it is faithful (left) and hallucinated (right), for four categories on LLaVA-1.5-7B. The late FFN surge appears only in the hallucinated instances.

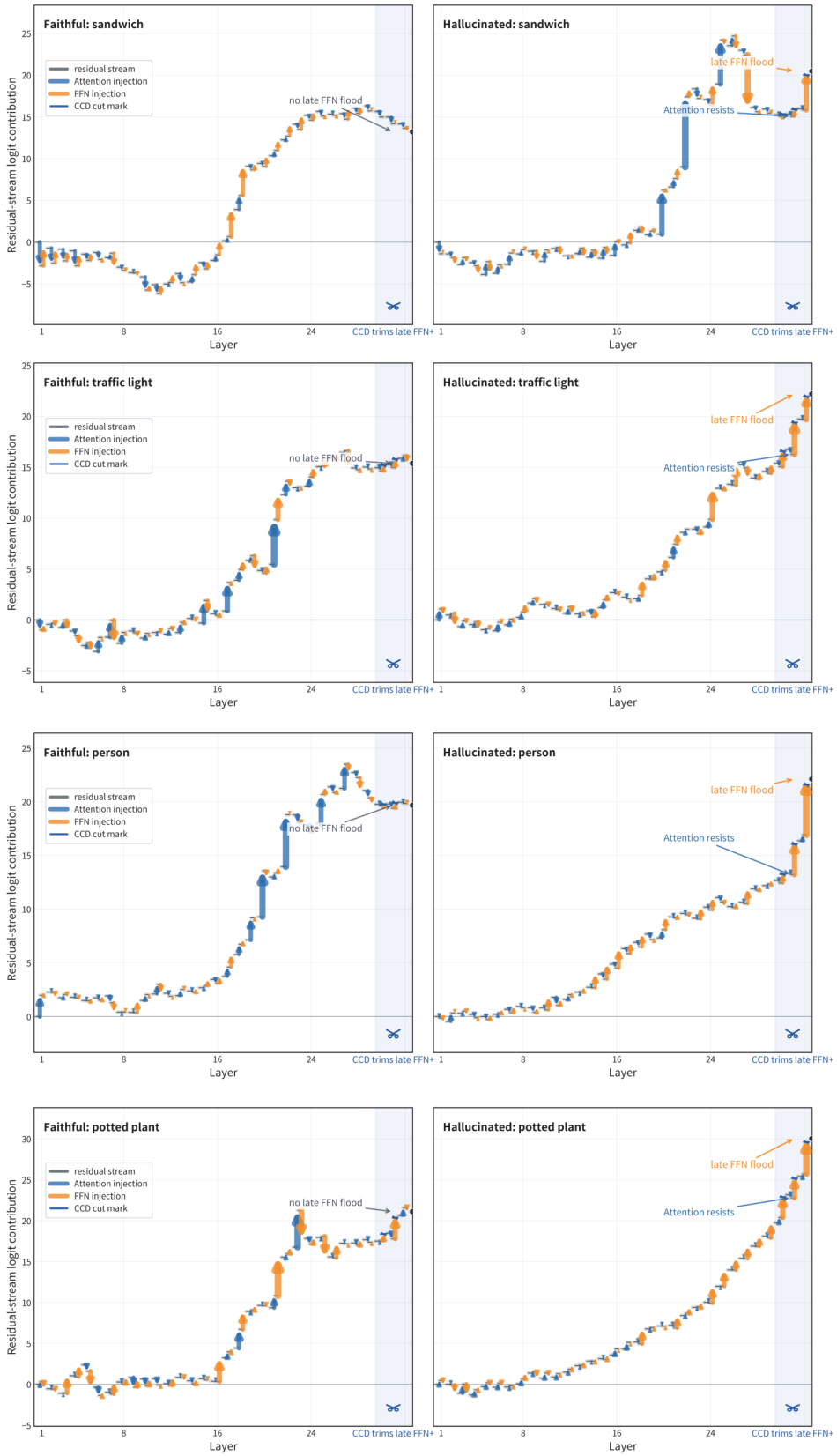


Figure 12: The same comparison for four further categories.

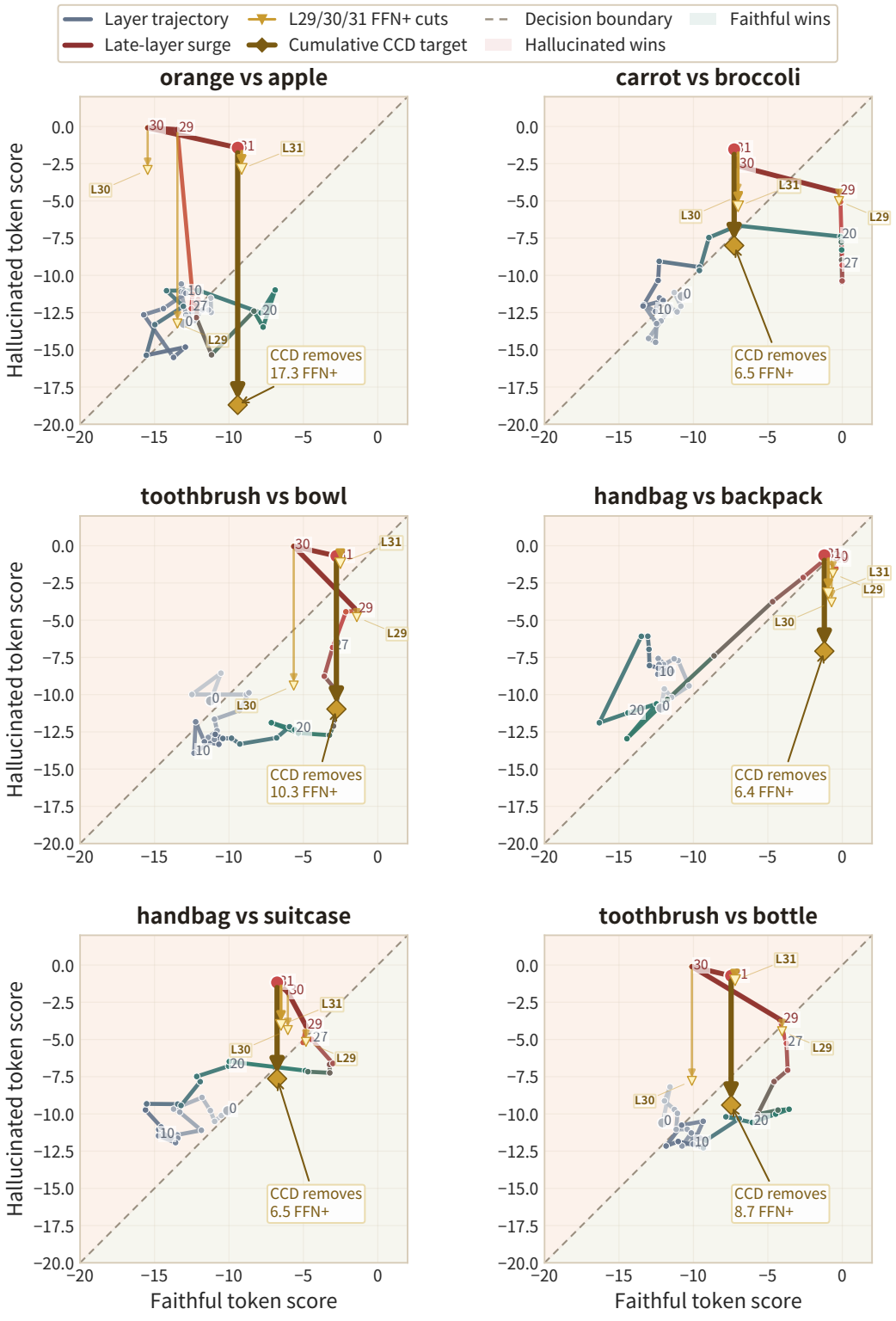


Figure 13: Pairwise logit trajectories for six hallucination cases on LLaVA-1.5-7B. Red segments mark the last layers, gold arrows show the FFN terms subtracted by CCD at each of the last three layers, and gold diamonds mark the corrected position.



Prompt: Describe this image.

Baseline Answer: The image features a woman lying on her back in the snow, with her arms and legs spread out. She appears to be enjoying the snowy weather and having fun. The woman is wearing a red jacket, which stands out against the white snow. There are several snow-covered objects in the scene, including a pair of skis located near the woman's legs and a snowboard placed further away. The woman's position and the presence of these winter sports equipment suggest that she might be taking a break from skiing or snowboarding.

CCD Answer: The image features a woman wearing a red jacket, lying on her back in the snow. She is smiling and appears to be enjoying the experience. The woman is positioned in the center of the scene, with her arms and legs spread out, creating a sense of relaxation and playfulness. The snow around her is quite deep, with a significant amount of it covering the ground. The woman's body is partially covered by the snow, adding to the overall wintery atmosphere of the scene.



Prompt: Describe this image.

Baseline Answer: The image features a white bicycle parked on a sidewalk next to a brick wall. The bicycle is positioned in front of a bench, which is located on the left side of the scene. The bench appears to be empty, and the bicycle is the main focus of the image. The scene is set in a public area, possibly a park or a street corner, where people can enjoy outdoor activities.

CCD Answer: The image features a white bicycle parked on a sidewalk next to a brick wall. The bicycle is positioned in front of a bench, which is located on the left side of the scene. The bench appears to be empty, and the bicycle is the main focus of the image. The scene is set outdoors, with the bicycle and bench being the main objects in the picture.

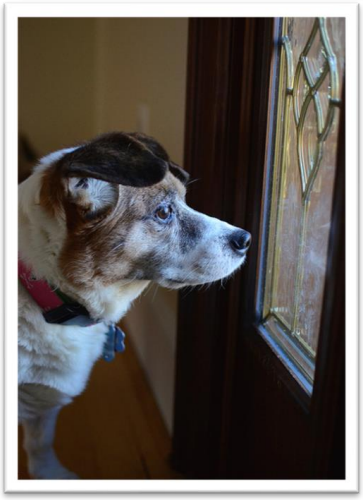
Figure 14: Captions from greedy decoding and CCD on LLaVA-1.5-7B. Green marks grounded objects and pink marks hallucinated ones.



Prompt: Describe this photo in detail.

Baseline Answer: The image depicts a kitchen scene with a black blender or food processor placed on a counter. This appliance appears to be churning out a shake, and there are additional ingredients placed nearby, including milk and bananas that have been sliced. In the room, there are a couple of more bananas that aren't shown in the image. Several bottles can be seen, with one placed on a surface near the left edge of the image and the others scattered throughout. On the ground, a bowl is located, possibly ready for adding ingredients to the blended drink. A clock is also visible in the background, providing a sense of time during the process.

CCD Answer: The image features a blender sitting on a kitchen counter. The blender is filled with a brown liquid, possibly a smoothie or a milkshake. The blender is placed on the counter, ready to be used. In addition to the blender, there are several bottles scattered around the kitchen. Some of these bottles are placed on the counter, while others are located in different areas of the kitchen. The presence of these bottles suggests that the kitchen is well-equipped for various cooking and food preparation tasks.



Prompt: Describe this image.

Baseline Answer: The image features a brown and white dog standing in a room, looking out of a window with a longing expression. The dog appears to be waiting for someone or something, possibly a family member or a friend. The dog is positioned near the center of the scene, with the window in the background. In the room, there is a chair located on the left side, and a couch can be seen in the background on the right side. The scene gives the impression of a cozy and comfortable living space where the dog enjoys spending time.

CCD Answer: The image features a brown and white dog standing in a room, looking out of a window with a longing expression. The dog appears to be waiting for someone or something, possibly a family member or a friend. The dog is wearing a collar, which adds to its adorable appearance. The room has a doorway visible in the background, and the dog is positioned near the door, looking out with anticipation. The scene captures the dog's curiosity and desire to be outside or to see what's happening beyond the window.

Figure 15: Further examples in the same format.