CONTRASTIVE DECODING IS CLASSIFIER-FREE GUIDANCE: UNIFYING VISUAL GUIDANCE FOR VISION-LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Classifier-free guidance is standard in image and video generation, yet is not commonly used in vision-language models. We show that the field has effectively used it without recognizing it: existing contrastive decoding methods are classifier-free guidance with different references, distinguished by the bias each reference carries. This unified formulation exposes a common design space for autoregressive generation. Exploring this space, we find that corrupted-image references increasingly align with simply deleting the image, while weaker models and degraded attention provide no better guidance. Guidance operators from diffusion likewise reduce to vanilla guidance at different scales. The key distinction from diffusion lies in the prefix: an autoregressive reference must condition on some answer prefix, but it is unclear which. Existing methods use the model's own prefix, leaking image information into the image-free reference and rapidly collapsing the guidance signal. A prefix-free reference avoids this leakage but becomes stale and overguides long answers. Switching between the two balances these complementary failure modes. Empirically, our approach improves Qwen2.5-VL-7B by 5.7 points and Qwen3-VL-8B by 1.7 points across ten general benchmarks, outperforming four published contrastive decoders at $1.04\times$ the cost of unguided decoding, compared with up to $2\times$ for prior methods.

1 INTRODUCTION

Classifier-free guidance (CFG) (Ho & Salimans, 2022) is a standard mechanism for sampling generative models: at each step, the model is evaluated with and without the conditioning prompt, and the predictions are combined by extrapolating away from the unconditional prediction. Modern text-to-image and video systems widely adopt CFG (Esser et al., 2024; Polyak et al., 2024). Yet unified models reveal a striking asymmetry: they use guidance when generating images, but not when generating text from visual inputs (Chameleon Team, 2024; Chen et al., 2025e).

Contrastive Decoding is classifier-free guidance. Visual contrastive decoding (Leng et al., 2024) subtracts the model's prediction on a noised image, and subsequent methods primarily differ in what the second stream observes. Section 3.1 shows that this is not merely an analogy but an identity: these methods are classifier-free guidance with different reference distributions. The implicit classifier of Ho & Salimans (2022) becomes an implicit *image* classifier, with guidance adjusting each token according to the evidence it carries about the image.

A family of methods becomes a unified design space. Diffusion treats guidance as a broad design space, varying the negative stream, the guidance operator, and its strength and schedule. In contrast, visual contrastive decoding methods typically instantiate individual choices. We place them under a common formulation and ask which principles transfer to autoregressive generation. This view enables techniques from diffusion, such as guidance scheduling and distillation (Meng et al., 2023; Black Forest Labs et al., 2025), to transfer naturally, while allowing existing methods to be compared on common ground. For example, PAI's attention modification (Liu et al., 2024) provides no benefit beyond increasing the guidance scale.

Guidance suffices. We systematically explore the negative reference and guidance operator. Image corruptions increasingly align with simply deleting the image, while weaker models either pro-

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.

duce too little useful contrast or diverge too far from the conditional model. Weakening access to the image through attention improves over unguided generation but does not outperform deletion. Likewise, guidance operators adapted from diffusion (Sadat et al., 2025a; Fan et al., 2025; Kwon et al., 2025; Zhang et al., 2024a) reduce to vanilla guidance at different scales. Together, these results leave a simple construction: delete the image and apply standard guidance.

Autoregressive guidance introduces a new axis: the prefix. Unlike diffusion, where both branches share the same state, an autoregressive image-free reference must condition on some prefix of the generated answer. Existing methods use the guided model's own prefix, passing image-dependent information into the reference intended to exclude the image. After a single generated token, this teacher-forced reference loses roughly two-thirds of its guidance signal, which subsequently falls below its noise floor. A prefix-free reference avoids this leakage and preserves a strong signal, but becomes stale as the answer grows and overguides long responses. The two references therefore succeed in complementary regimes: prefix-free guidance is stronger on short answers, while teacher forcing is better suited to long ones. We combine them by switching references within a single answer.

Our method requires neither training nor benchmark-specific tuning. It improves Qwen2.5-VL-7B by 5.7 points and Qwen3-VL-8B by 1.7 points across ten benchmarks, at only $1.04\times$ the cost of unguided decoding, compared with up to $2\times$ for published contrastive decoders.

2 Related Work

Guidance in generation. Classifier-free guidance (Ho & Salimans, 2022) replaced the separately trained classifier of Dhariwal & Nichol (2021) with one network run with and without its condition, became the default of text-to-image and text-to-video generation (Nichol et al., 2022; Rombach et al., 2022; Ramesh et al., 2022; Saharia et al., 2022; Esser et al., 2024; Polyak et al., 2024), and survived every change of model family, including the token generators that mix logits rather than scores (Chang et al., 2023; Sun et al., 2024a; Tian et al., 2024). The work that followed treats it as a design space rather than a fixed rule: how strong the push should be and for how long (Kynkäänniemi et al., 2024; Wang et al., 2024b; Sadat et al., 2024), how to reshape the update (Sadat et al., 2025a; Fan et al., 2025; Kwon et al., 2025; Zhang et al., 2024a), and what the negative branch should see, which need not be unconditional at all (Hong et al., 2023; Ahn et al., 2024; Hong, 2024; Karras et al., 2024). Visual understanding has no counterpart.

Contrastive decoding. Language modelling reached the same contrast on its own (Li et al., 2023a; Chuang et al., 2024; Shi et al., 2024), and VCD (Leng et al., 2024) brought it to vision-language models by subtracting the prediction on a noised image. Dozens of methods followed, differing mainly in what the contrast stream sees: a distorted instruction (Wang et al., 2024d), no image (Favero et al., 2024; Zhang et al., 2025b), another layer (Tong et al., 2025), or another attention pattern (Liu et al., 2024; Huo et al., 2025; Jo et al., 2026). Each fixes one point in the design space above, evaluates it almost only on hallucination suites (Li et al., 2023b; Rohrbach et al., 2018).

Appendix C gives the full account of both lines.

3 From diffusion guidance to autoregressive decoding

3.1 Guidance as an implicit image classifier

From a classifier to an image. Classifier guidance needs a separately trained classifier for $\nabla \log p(c \mid x)$ (Dhariwal & Nichol, 2021); Ho & Salimans (2022) observe that Bayes' rule already contains one, so a single model run with and without its condition plays both roles. A vision-language model is exactly such a model. With v the image, $h = (q, y_{<t})$ the question and the answer so far, $\ell_c = \log p(\cdot \mid v, h)$, and ℓ_u the same prompt with the image removed, guidance is

$$\tilde{\ell} \;=\; \ell_u + \alpha\,(\ell_c - \ell_u) \;=\; \ell_c + (\alpha - 1)\,(\ell_c - \ell_u), \qquad \tilde{p} = \mathrm{softmax}(\tilde{\ell}), \tag{1}$$

and $\alpha = 1$ is plain decoding. Were the image free pass the true marginal $\bar{p}(y \mid h) = \mathbb{E}_{v \sim p(v \mid h)}\, p(y \mid v, h)$, Bayes' rule would turn the bracket into the pointwise mutual information between the candidate token and the image,

$$\log p(y \mid v, h) - \log \bar{p}(y \mid h) \;=\; \log p(v \mid y, h) - \log p(v \mid h) \;=\; \mathrm{pmi}(y; v \mid h), \tag{2}$$

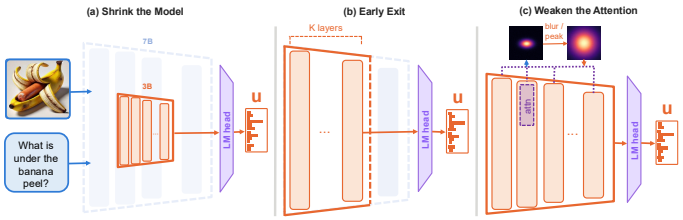

Figure 1: **Three ways to degrade the model instead of the input.** (a) A smaller model, (b) an early exit after L layers, and (c) the same model with its attention to the image edited (Table 1). Each produces the negative stream u.

Table 1: **Weakening the model.** Each edit acts on a text token's scores s over the image tokens, or on its attention p; symbols are defined in Section 3.2.

Variant	Negative stream	Effect
Early exit	Readout after layer L	Fewer layers
Smaller	A 3B model	Less capacity
MUL	$(1+\gamma)\,s$	Flatten
SEG	$G_\sigma * s$	Blur
Over-blur	$p_{\text{SEG}} + a\,(p - p_{\text{SEG}})$	Blur further
Masked	$s = -\infty$	Remove
Soft PAG	$(1-\lambda)\,p + \lambda\,e$	Peak
PAG	e	Peak fully

so the implicit classifier of Ho & Salimans (2022) becomes an implicit *image* classifier. **Classifier-free guidance for a vision-language model pushes each token by how much evidence it carries.**

Every contrastive decoding method is this operator. The operator is the geometric tilt $\tilde{p} \propto p_c^{\alpha}\, p_u^{1-\alpha}$. On two Gaussians with a shared variance, a diffusion reverse step, it returns the update of Ho & Salimans (2022); on two categorical distributions, a vision-language model's next token, it returns Equation 1. The two are not analogous, they are one operator applied to two families (Appendix B.2). No published reference is the marginal $\bar{p}$, and writing $q(y)$ for whatever a method subtracts, the bracket splits exactly as

$$\ell_c(y) - \log q(y) \;=\; \underbrace{\operatorname{pmi}(y; v \mid h)}_{\text{image evidence}} \;+\; \underbrace{\log \bar{p}(y \mid h)/q(y)}_{\text{reference bias } b_q(y)}, \tag{3}$$

so every such method applies the same tilt and differs *only* in the bias b_q, which Appendix B.3 tabulates. Contrastive decoding is not a relative of classifier-free guidance; it *is* classifier-free guidance with a particular reference, and the question worth asking is which reference.

3.2 From corrupted images to autoguidance

Degrading the input. Classifier-free guidance drops the condition (Ho & Salimans, 2022), which for a vision-language model means deleting the image. Contrastive decoding corrupts it instead, as visual contrastive decoding does with noise (Leng et al., 2024). We tested a wide range of corruptions, from black, resized and blurred images to noised, masked, grayscale, patch shuffled and colour inverted ones (Table 14). All of them produce a guidance vector whose alignment with deletion's is far above chance, and that alignment rises with the strength of the corruption: the constructions form a cone that closes onto deletion, with the noised image of visual contrastive decoding as its weakest point. The guidance scale cannot turn one construction into another, because it absorbs the *length* of a construction's vector but never its *angle*: matching deletion's step gives $\alpha^\star - 1 = (\alpha_0 - 1)\cos\theta\,\|\Delta_{\text{drop}}\|/\|\Delta_p\|$, with a residual proportional to $\sin\theta$. Deletion is the only construction with nothing to tune, and the cheapest, since its negative pass carries no image tokens.

Degrading the model. Autoguidance (Karras et al., 2024) proposes a different negative: a weaker copy of the same model, which reproduces the conditional's errors more strongly and so isolates the direction worth following. Taken literally (Figure 1a,b), a weaker copy is either too close to the model to move an argmax (a smaller one) or too far to reflect its errors (an early exit that reads the answer out after layer L). Diffusion can accumulate a small difference over many denoising steps; here, each step is read once. A third variant keeps the premise but weakens the model's *access to the image* rather than its capacity (Figure 1c). In every layer, the scores s that a text token assigns to the image tokens, or its attention $p = \operatorname{softmax}(s)$ over them, are replaced by an edited version, and the negative stream is the same model run with that edit (Table 1). MUL flattens the scores by multiplying them by $1+\gamma$ with $\gamma < 0$; SEG (Hong, 2024) blurs them over the patch grid with a Gaussian kernel G_σ of width σ; and Over-blur extrapolates past the blurred map, away from p, to $p_{\text{SEG}} + a\,(p - p_{\text{SEG}})$ with $a < 0$. All three push the attention towards uniform. Soft PAG instead mixes it, with weight λ, with a one-hot map e on its top patch, and PAG is the case $\lambda = 1$; both are our cross-attention analog of the self-attention perturbation of Ahn et al. (2024), and push the

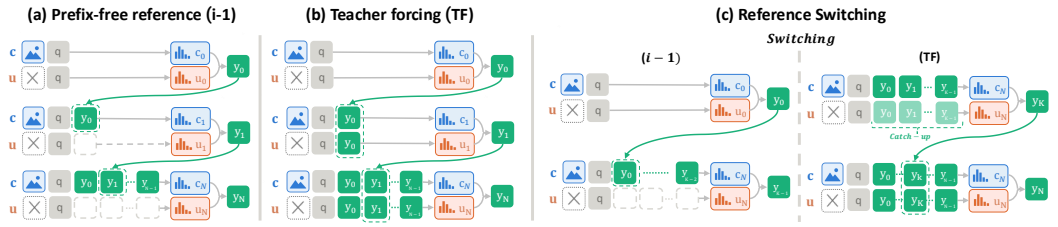

Figure 2: **Three rules for the image free reference.** (a) I-1 reads the question alone and is computed once. (b) TF reruns the reference on the conditional's full prefix at every step. (c) SWITCH runs I-1 for the first K tokens, then catches the reference up and continues with TF.

attention towards a single patch. Finally, masked sets every score to $-\infty$, so no attention reaches the image. Pushed towards uniform, the weakened attention improves on unguided decoding and reaches parity with deletion; pushed towards a single patch, it barely moves the model on most benchmarks. It cannot go further, for a structural reason: even Masked leaves the image's marker tokens, positions, and cache entries, all of which deletion removes.

3.3 WHERE THE REFERENCE STANDS

Guidance suppresses the language prior, but relative to which prefix? The image-free stream sees the question and the answer so far, but not the image, so it predicts what the model would say without looking: the language prior. This prior has long been known to make VQA models answer from the question rather than the image (Goyal et al., 2017; Agrawal et al., 2018), and to drive hallucination in large vision-language models (Leng et al., 2024; Zhang et al., 2025b). Subtracting it in Equation 1 leaves each token with the evidence it carries about the image (Equation 2). In diffusion both branches see the same noisy sample; in an autoregressive VLM the image-free stream can be conditioned on any prefix of the answer. Figure 2 shows the options: no prefix (a), the conditional's full prefix (b), or a switch from (a) to (b) within one answer (c).

Teacher forcing follows the answer, and leaks it. The intuitive choice keeps the image-free stream on the conditional's full prefix at every step (Figure 2b). This teacher-forced reference (TF) is the rule of visual contrastive decoding (Section 2). It forces that stream to read tokens it did not generate, and those tokens reveal what the image contains: the stream meant to be blind to the image no longer is, and the guidance term, the gap between the two streams, shrinks. Since the first generated token is already a decision about the image, the leak should be immediate; Section 4.1.1 confirms it: guidance collapses after the first token.

The prefix-free reference reads the question once. The alternative runs the image-free stream once, at the first step, and reuses its prediction for the whole answer (I-1, Figure 2a). That is the right place to read the language prior: the model has attended to the entire input and to none of the answer, the state at which decoder-only models are read out as encoders (Jiang et al., 2024; Wang et al., 2024a; Ma et al., 2024). The cached reference is therefore the question's prior, untouched by the answer, and I-1 is cheaper: N tokens take $N + 1$ forward passes, against $2N$ for TF.

Neither rule fits every answer length. At the first step the prefix is empty, so both rules coincide on single-token answers. Beyond it, they fail in opposite regimes. The prefix-free term never fades, so I-1 guides at full strength throughout. On a short answer this helps twice: the signal is strong, and a reference still expecting the answer to begin pushes the model towards stopping. On a long answer it compounds: every guided token enters the cache, so each correction is reapplied at every later step and the answer drifts. TF fails the other way, but its leak is not only a defect. Because the reference has read the prefix, its remaining push measures how much the image still matters at *this* token: once the text has fixed the answer, the reference agrees and guidance vanishes. No fixed schedule supplies this estimate, and none we tried did (Section 4.1.1). We therefore expect I-1 to win on short answers and TF on long ones, unless the answer stays image-determined to its last token; there, the cached push is redundant rather than harmful.

Switching the reference inside one answer. Diffusion meets the same tension along its trajectory and resolves it by scheduling the guidance strength over time (Chang et al., 2023; Kynkäänniemi et al., 2024; Wang et al., 2024b). The reference gives an autoregressive model a schedule of its

own. SWITCH keeps the reference prefix-free for the first K tokens then switches to teacher forcing (Figure 2c): full strength while the answer is being decided, then TF's token-by-token fading once it is underway. At the switch, the reference catches up on the prefix, after which SWITCH is identical to TF (Appendix A.2). The switch point K is the only hyperparameter; Section 4.1.1 describes how we set it. On answers shorter than K the switch never fires, and guidance costs one extra pass.

3.4 WHAT THE GEOMETRY OF THE UPDATE CAN CHANGE

The diffusion literature reshapes the guidance update to repair Equation 1: oversaturated images and samples off the data manifold. Table 2 writes four widely used operators in our notation. APG (Sadat et al., 2025a) rescales the part of Δ parallel to ℓ_c by a factor $\eta < 1$, which we also take above one; CFG-Zero* (Fan et al., 2025) replaces ℓ_u by the multiple $s^\star \ell_u$ that best fits ℓ_c; TCFG (Kwon et al., 2025) keeps only the part of ℓ_u along the leading direction u_1 the two streams share; and EP-CFG (Zhang et al., 2024a) rescales the guided logits back to the norm of ℓ_c.

Every such operator is a guidance scale. Every operator in Table 2 returns a weighted sum of the same two logit vectors, $\tilde{\ell} = a\,\ell_c + b\,\ell_u$, with $a > 0$ and a step dependent b. A greedy decoder takes an argmax, which cannot see a positive rescaling, so dividing by a gives

$$\arg\max_y \tilde{\ell}(y) \;=\; \arg\max_y \big[\ell_c(y) - r\,\ell_u(y)\big], \qquad r = -b/a. \tag{4}$$

The two weights reach the decision only through their ratio r, and for $r < 1$ vanilla guidance produces the same ratio at the scale $\alpha_{\text{eff}} = 1/(1-r)$; every operator step we measured has $a > 0$ and $r < 1$. *An operator that stays in the span of the two streams is therefore a schedule for the guidance scale* (Appendix A.4). Composing operators does not escape this: a composition is again a weighted sum of ℓ_c and ℓ_u; under sampling the factor a survives as an inverse temperature.

What each operator computes. Because α_{eff} depends only on the two streams, we can read it off the model and name what each operator actually does (Appendix A.4). *EP-CFG is the identity*: it scales both weights by one factor, leaves r unchanged, and in exact arithmetic runs at α. *APG is a relabelled constant*: it shifts the scale by about 0.1 and holds it there, up for $\eta < 1$ and down for $\eta > 1$. *TCFG is a strength rule*: with equal norms the direction it keeps is the streams' bisector, giving $\alpha_{\text{eff}} = (\alpha + 1)/2$ whatever the angle between them; unequal norms move it away from that, but only inside $[1, \alpha]$, so TCFG can lower the guidance excess and never raise it. *CFG-Zero* is a duration rule*: on the prefix free reference its fitted $s^\star$ is 0.77 to 0.99 at the first token, and then the stale reference turns orthogonal to the conditional as soon as the answer starts, so $s^\star$ falls to zero and guidance switches off until the reference refreshes. The imported toolkit therefore reduces to the two knobs Section 3.3 derives from the decoding itself: how strong the push is and how long it lasts. An operator can still beat the default scale, but only by landing on a better constant or schedule, so the right test is against the best constant scale, which Section 4.1.3 runs.

Table 2: Geometry operators each written as guided logits $\tilde{\ell} = a\,\ell_c + b\,\ell_u$, with $\Delta = \ell_c - \ell_u$ and $\gamma = \alpha - 1$.

operator	guided logits $\tilde{\ell}$	a	b
vanilla CFG	$\ell_c + \gamma\,\Delta$	$1 + \gamma$	$-\gamma$
APG	$\ell_c + \gamma\,(\eta\,\Delta_\parallel + \Delta_\perp)$	$1 + \gamma\,(1 + (\eta - 1)s)$	$-\gamma$
CFG-Zero*	$\ell_c + \gamma\,(\ell_c - s^\star \ell_u)$	$1 + \gamma$	$-\gamma\, s^\star$
TCFG	$\ell_c + \gamma\,(\ell_c - \hat{\ell}_u)$	$1 + \gamma\,(1 - p)$	$-\gamma\, q$
EP-CFG	$\kappa\,(\ell_c + \gamma\,\Delta)$	$\kappa\,(1 + \gamma)$	$-\kappa\gamma$

$\Delta_\parallel = s\,\ell_c,\ s = \Delta^\top \ell_c / \|\ell_c\|^2,\ \Delta_\perp = \Delta - \Delta_\parallel$
$s^\star = \ell_c^\top \ell_u / \|\ell_u\|^2$, the least squares fit
$\hat{\ell}_u = (\ell_u^\top u_1)\,u_1$
$\quad = p\,\ell_c + q\,\ell_u$, with u_1 the top eigenvector of $\ell_c \ell_c^\top + \ell_u \ell_u^\top$
$\kappa = \|\ell_c\| / \|\ell_c + \gamma\,\Delta\|$

4 EXPERIMENTS

Setup. We run everything in `lmms-eval` (Li et al., 2024) with greedy decoding and each benchmark's default prompt. One configuration is fixed in advance and never tuned per benchmark: the negative stream deletes the image, the blend acts on the last hidden state, which the linear readout makes equivalent to blending the logits (Appendix A.1), the reference switches at $K = 8$, and $\alpha = 1.5$ throughout. Section 4.1 analyses that configuration on Qwen2.5-VL-7B-Instruct (Bai et al., 2025b); Section 4.2 then applies it unchanged to five models from three families at 3B to 8B (Qwen2.5-VL (Bai et al., 2025b), Qwen3-VL (Bai et al., 2025a) and InternVL3.5 (Wang et al., 2025)), against four contrastive decoding baselines at their published settings. Our ten benchmarks

Table 3: **Where the reference stands**, on benchmarks ordered by mean answer length. *+oracle*: the I-1 push rescaled to TF's per token strength (the strength oracle). *+decay*: rescaled by TF's average decay (the decay schedule). *Overall*: mean over the seven benchmarks.

	OK-VQA	VizWiz	O3	PhD-rat.	ReasonSeg	OCRBench	MathVerse	Overall
# tokens	*2.4*	*2.5*	*3.8*	*30.2*	*60.9*	*152.1*	*337.7*	
no guidance	42.60	73.80	57.34	68.73	49.45	53.69	60.28	57.98
TF	52.76	76.01	58.68	71.98	52.01	52.96	62.82	61.03
I-1	54.00	78.06	63.61	69.99	50.18	53.83	56.01	60.81
I-1 +oracle	54.60	78.46	56.91	68.33	51.29	52.41	61.55	60.51
I-1 +decay	54.64	78.49	58.96	69.72	50.91	53.83	58.83	60.77
SWITCH	54.00	78.06	63.61	72.78	51.86	53.77	62.54	62.37

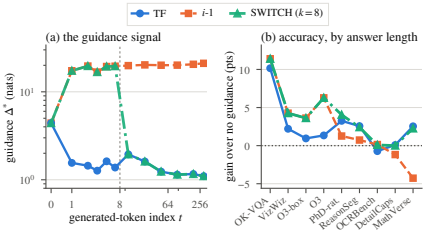

Figure 3: (a) The guidance term $\Delta^\star$ per generated token (OCRBench v2). (b) Gain over unguided decoding, by answer length.

are listed in Appendix B.8; the analysis of Section 4.1 adds longer answers, the axis along which the reference rules separate.

4.1 Analysis

4.1.1 Where the reference stands

Each rule wins in its own regime. Section 3.3 predicts that I-1 wins on short answers and TF on long ones, and Table 3 confirms both, with one instructive exception. On every benchmark whose answer is shorter than five tokens I-1 beats TF, by up to 4.93 points. On long answers the order reverses: on MathVerse I-1 falls 6.81 points below TF and 4.27 below no guidance. Figure 3a shows the reason. After one generated token the teacher forced push on the decision token, $\Delta^\star$, loses about two thirds of its value, while the cached push stays an order of magnitude higher. The drop is the leak described in Section 3.3, not the ease of continuing an answer: with length and step fixed, $\Delta^\star$ falls from 18.35 nats with no prefix to 15.05 with another document's prefix, 7.55 with this document's tokens shuffled, and 1.48 with its real prefix (Table 8). The prefix window closes the loop from the guidance term to accuracy. Letting the reference read only the last w generated tokens interpolates between the two rules, and the whole effect sits in the first token: on VizWiz, where I-1 is the better rule, one token of prefix already costs 2.59 points, and every larger window stays within 0.6 of it; on RefChartQA, where TF is better, one token brings 67% of *its* advantage (Table 9). On short answers the strong cached push, together with the stop pressure it creates, is what helps: on O3, I-1 stops after the first label in 90% of answers, which lifts F1 from 58.68 under TF to 63.61 (Table 3).

The exception is the one the mechanism predicts. OCRBench v2 has long answers, yet TF costs 0.73 points there while I-1 is flat against no guidance ($+0.14$). Reading text off an image keeps the answer dependent on the image at every token, so I-1's undiminished push lands on tokens the conditional already ranks first and changes nothing, while TF's push, 0.45 nats on the leader against a 1.03 nat spread over its rivals, reshuffles near ties. What survives TF's cliff is below its own noise floor: from the first generated token onward, on every benchmark and at every depth, TF pushes the leader less than it displaces its rivals, while the cost of standing on a foreign prefix does not fade with the signal, the reference's excess entropy staying flat at 0.29 to 1.49 nats (Table 8d). Teacher forced guidance does not switch off gracefully: its signal drains while its perturbation remains. On OCRBench's readout subtasks themselves ($n = 692$) both rules gain, by 3.9 to 6.1 points, and only a schedule that removes the late push gains nothing (Appendix A.2).

Strength, not direction, is what goes wrong. If the cached reference fails on long answers because its push stays too strong, and not because it points the wrong way, then shrinking the push alone should repair it. We test this with two variants built on top of I-1 (Table 3). The strength oracle keeps the I-1 direction but gives it the per token strength of TF; it runs the teacher forced stream to do so, so it explains rather than competes. It recovers TF on long answers (MathVerse 61.55 against 62.82, ReasonSeg 51.29 against 52.01, both within noise) and keeps the gain of I-1 on VizWiz. It fails on one axis only: because it carries the I-1 direction it also carries I-1's stopping behaviour, so it loses on the benchmarks whose guided gain is the stop decision itself, O3 and PhD with a rationale. Strength alone reproduces what TF does to the content of an answer, not what it does to its length. The decay schedule, a fixed curve that needs no second stream, falls 3.99 points short on MathVerse, and the other single stream schedules fall short by comparable margins (Appendix A.2). The right strength after the first tokens has to be read token by token from the teacher forced stream.

Switching takes the better of both rules. Together, the two results call for one rule: keep the cached reference while it is still close to exact, and read the teacher forced stream once it goes stale. SWITCH (Section 3.3) does exactly this, with $K = 8$ set from answer length statistics before reading any score (any K from 4 to 32 ties TF on MathVerse, Table 11). It follows the better rule at every answer length (Figure 3b) and beats both on average, 62.37 against 61.03 for TF and 60.81 for I-1 (Table 3). It is also cheap where it can be: on short answers the switch never fires, so it costs what I-1 costs, 1.19 to 1.31× unguided decoding against 1.74× for TF, and never more (Table 12).

Finding 1. The teacher forced reference leaks the answer and collapses after one token; the cached one never fades and overguides long answers. Switching between them takes the better of both, and single token benchmarks cannot see the difference.

4.1.2 FROM CORRUPTED IMAGES TO AUTOGUIDANCE

Weakening the model's capacity misses on both sides. Every corruption of the image we tested points towards simply deleting it (Appendix A.3), so we turn to autoguidance, which weakens the model instead (Section 3.2). Its two literal forms, an early exit after layer 15 and the smaller 3B model of the same family, both fail at $\alpha = 1.5$, and on opposite sides (Table 5). The 3B negative is too close to the model to move an argmax, so subtracting it removes almost nothing and VizWiz falls to 69.79. The early exit is too far: its top token never matches the conditional's, on any of the 1,727 documents we checked, so it is a different model rather than a weaker one, and it scores 73.57 against 73.80 unguided (Table 15).

Weakening attention reaches parity, not more. The third variant weakens the model's access to the image instead of its capacity, through the six edits of Table 1 (Section 3.2). On VizWiz and BLINK every edit improves on unguided decoding and the two sides score alike, between 75.22 and 77.76 and between 26.67 and 28.67, all below deletion. MME-RealWorld separates them along the side the edit pushes towards, as Section 3.2 anticipates: the edits towards uniform gain 4 to 6 points, while the two PAG variants stay at or below the unguided level, because a map already peaked barely changes when pushed further (its negative agrees with the conditional on 74.6% of documents). Against deletion no edit does better than parity, and only tuning both sides makes that visible: an apparent 1.62 point win for the blurred negative at a shared α disappears once deletion is tuned too, leaving three ties and one significant loss across four benchmarks (Table 16). The edit that comes closest also changes from benchmark to benchmark (Masked on VizWiz, PAG on BLINK, SEG on MME-RealWorld), so weakening attention forces a per-benchmark choice where deletion forces none. And deletion is the limit these edits approach rather than one they pass: the residue they leave behind grows with image size, to 2.34 points on MME-RealWorld (Appendix A.3).

The mirage of composition. Weakened attention and deletion degrade different things, so it is tempting to use both. The headroom looks real: choosing the better of the two per document with hindsight gains 1.86 points, positive on all eight benchmarks. No combination captures it. The best recovers 0.15 points, and of 37 benchmark and combiner pairs 3 are significant wins and 7 significant losses, because the blurred attention's guidance vector lies inside the cone of the pixel corruptions, so every combination lands on the scale curve of deletion alone, matched in score by a single α to within 0.3 points (Appendix A.3). A second negative is a larger guidance scale in disguise.

Finding 2. Every negative stream we built, from the pixels, from a weaker model or from degraded attention, either points towards image deletion or fails. A weaker model is too close to the conditional or too far from it; weakened attention reaches parity with deletion, and combining the two only rescales the guidance.

4.1.3 THE GEOMETRY OF THE UPDATE

Every operator lands on the scale ladder. Section 3.4 shows that every operator of Table 2 is vanilla guidance at another scale. Table 4 runs them on top of SWITCH. Averaged over five benchmarks, the default scale, TCFG, EP-CFG and APG with $\eta = 2$ score 64.02, 64.02, 64.00 and 64.18: three published operators and the default, one number. Single cells do move, and no operator is significantly better than the best constant α on any benchmark (Table 18). The nearest miss shows why the comparison must be against the tuned scale: on O3, TCFG beats the default scale by 2 points, but the best rung, $\alpha = 1.25$, by only 0.34, which is not significant.

Table 4: **Guidance operators against the scale.** Five benchmarks ordered by answer length, all on SWITCH ($K = 8$) at $\alpha = 1.5$; *Overall* is the unweighted mean. The full-scale ladder is in Table 18.

	VizWiz	O3	PhD-rat.	ReasonSeg	OCRBench	Overall
unguided, α=1	73.80	57.34	68.73	49.45	53.69	60.60
VLM-CFG	78.06	63.61	72.78	51.86	53.77	64.02
APG, $\eta = 0.5$	77.59	63.15	72.18	51.65	54.11	63.74
APG, $\eta = 2$	78.43	64.13	72.84	51.67	53.82	64.18
CFG-Zero*	78.70	56.15	71.05	51.95	54.01	62.37
TCFG	78.11	65.61	71.18	51.35	53.87	64.02
EP-CFG	78.01	63.68	72.18	51.84	54.30	64.00

Table 5: **Autoguidance:** early exit, smaller model, attention perturbation.

Negative stream	VizWiz	BLINK IQ	MME-RW
Unguided	73.80	24.67	41.38
VLM-CFG	78.06	28.67	45.86
Early exit, layer 15	73.57	24.67	39.81
Smaller, 3B	69.79	24.00	41.01
MUL, $\gamma = -0.5$	75.54	27.33	45.81
SEG	75.22	27.33	47.47
Soft PAG, $\lambda = 0.25$	76.83	26.67	41.12
PAG, $\lambda = 1.0$	77.16	28.67	40.33
Over-blur, $a = -0.5$	77.16	26.67	46.64
Masked ($-\infty$)	77.76	26.67	43.51

Table 6: Comparison with existing contrastive decoding methods on ten VQA benchmarks.

	OK-VQA	VizWiz	O3	O3-box	PhD	MMStar	TreeBench	BLINK IQ	V*	MME-RW	Overall	Cost
Qwen2.5-7B	42.36	73.49	57.39	60.24	67.50	**62.43**	40.74	25.33	70.16	39.24	53.89	1.00×
VCD (Leng et al., 2024)	47.24	74.37	55.52	59.80	70.90	61.73	41.98	26.67	70.16	42.42	55.08	1.96×
M3ID (Favero et al., 2024)	42.16	75.06	55.61	57.73	67.51	58.01	37.04	25.33	70.68	46.17	53.53	1.08×
PAI (Liu et al., 2024)	50.68	75.81	**65.28**	**67.29**	70.49	62.26	41.48	28.00	70.16	41.84	57.33	2.05×
SID (Huo et al., 2025)	48.24	73.49	56.74	61.74	73.04	61.19	41.73	26.67	72.25	45.49	56.06	2.00×
VLM-CFG (ours)	**54.00**	**78.01**	63.94	65.86	**79.15**	62.17	**42.96**	**29.33**	**73.30**	**47.11**	**59.58**	1.04×
Qwen3-8B	**52.96**	81.06	59.62	70.61	74.52	60.16	44.69	28.00	79.06	46.12	59.68	1.00×
VCD (Leng et al., 2024)	51.76	79.28	60.28	70.76	74.46	59.50	44.69	28.67	**81.15**	48.51	59.91	1.96×
M3ID (Favero et al., 2024)	52.68	81.09	59.63	70.49	74.65	60.20	45.68	26.67	78.53	46.22	59.58	1.08×
PAI (Liu et al., 2024)	51.72	79.71	62.09	71.08	76.22	**61.97**	44.44	30.00	79.58	47.37	60.42	2.05×
SID (Huo et al., 2025)	51.60	80.95	62.30	**72.64**	77.55	60.77	**46.42**	28.00	76.96	47.99	60.52	2.00×
VLM-CFG (ours)	52.36	**81.13**	**63.80**	72.43	**79.01**	61.16	45.19	**30.67**	79.58	**48.62**	**61.40**	1.04×

Predictions read off the states, and the two that were wrong. Because α_{eff} is measured on stored trajectories, every cell was predicted before its score existed, and four of the six predictions held: the code path with the operator switched off reproduced SWITCH on 1,131 of 1,131 answers, and TCFG and both APG rungs scored at the ladder rung their scale named, to within the noise floor. Two failed, and both corrections are in the table. EP-CFG was predicted to emit identical answers; in bf16 it changes 0.8% of VizWiz answers and 24% of PhD rationales, which is what gives the table its noise floor. CFG-Zero* was predicted to lose on VizWiz, from a median α_{eff} of 1.03, and it gained 0.64 instead: the median had averaged one guided step together with many unguided ones and described no step at all. Resolving it by step index gives the duration rule of Section 3.4 and a sharper prediction, which the next paragraph tests.

The two operators that move the scale are known rules in disguise. Section 3.4 names what each operator computes, and the measurements confirm it (Appendix A.4). When the two streams have equal norms, TCFG runs at the predicted $\alpha_{\text{eff}} = (\alpha+1)/2$: we measure 1.251 against 1.25, and on VizWiz TCFG's answers match those of $\alpha = 1.25$ on 96.6% of documents. CFG-Zero* acts on duration instead: its fitted $s^\star$ runs from 0.77 to 0.99 at the first token and falls to zero afterwards, so the operator guides the first token and then nothing until SWITCH refreshes the reference (Table 17). It fails exactly where that rule should. It holds on VizWiz and ReasonSeg, where the answer is decided early; it loses 1.73 points on the 30 token rationales of PhD, and on O3, whose gain lies in the stop decision, it falls 7.46 points, below unguided decoding. The imported toolkit thus adds no direction; it rediscovers the two knobs of Section 3.3: how strong the push is and how long it lasts.

Finding 3. Every guidance operator imported from diffusion lands on the scale ladder, and none beats the best constant scale on any benchmark. The two that move the scale appreciably are a halving of it (TCFG) and guidance on the first token until the reference refreshes (CFG-Zero*).

4.2 MAIN RESULTS

The contrastive decoding family on a general suite. Each contrastive decoding method is one point in the design space above (Section 3.1), and the line is evaluated almost entirely on hallucination suites and on backbones two generations old, with reproduction studies questioning even those gains (Yin et al., 2025b; Bendre et al., 2026). We re-run four members at their published settings on ten general benchmarks (Table 6; every model in Table 19), and the suite splits them in two. The methods that act on the conditional stream gain on every model: PAI (Liu et al., 2024) by 0.1 to 3.4

points and SID (Huo et al., 2025) by 0.6 to 2.2. PAI is the mirage of composition of Section 4.1.2, an attention edit on top of image deletion guidance that a tuned scale matches (Appendix A.6). The two that build a fixed negative do not gain. VCD (Leng et al., 2024) loses on two models of five, as Section 3.2 predicts of a noised image, the weakest point of the corruption cone. M3ID (Favero et al., 2024) is below the unguided model on all five.

Guidance chosen by the analysis. VLM-CFG is the design Section 4.1 arrives at: image deletion, the last hidden state, the switching reference at $K = 8$, and $\alpha = 1.5$ everywhere. It is the strongest arm on both blocks of Table 6, gaining 5.69 points over the unguided Qwen2.5-7B against 3.44 for the best published method, and a further 1.72 on Qwen3-8B, whose unguided score of 59.68 already exceeds what any of the four baselines reaches on Qwen2.5-7B. Qwen2.5-3B behaves like the 7B, at 3.42 points against 1.91. On the hallucination benchmark PhD-CCS, prompting the same models to reason before answering is worth at most 0.72 points, while guidance adds 2.65 to 11.65 (Table 20). The same fixed configuration transfers across two model families and four sizes. One model does not follow: on InternVL3.5-8B guidance is worth 0.07 points and both PAI and SID edge past it.

The cheapest guided decoder. A guided decoder pays for its negative branch, and the four do not pay alike. VCD and SID re-encode the image and carry every vision token in theirs, so each of their forwards costs a full conditional pass. M3ID drops the image instead, as we do, so it costs what a teacher forced second stream costs. PAI's negative is image free too, but its attention edit forces an eager kernel on the conditional stream, where the cost is. On the short answers of this suite, prefill included, SWITCH costs $1.04\times$ unguided decoding, against $1.96\times$ for VCD, $2.00\times$ for SID, $2.05\times$ for PAI and $1.08\times$ for M3ID (Table 12). Against M3ID the margin is four points of overhead, so we do not rest that comparison on cost: M3ID picks the right negative branch but leaves the reference rule unstated, pairing it with a weight that grows without bound in the token index, where our scale ladder finds no schedule beating the best constant.

4.3 LIMITATIONS

Adaptive guidance strength. We use a single guidance scale everywhere, which keeps the method a one line change to the decoding loop but leaves gains on the table: the best scale differs from benchmark to benchmark, and picking it per benchmark beats our fixed choice (Table 18). Picking it per sample should do better still, but we cannot yet: our best adaptive rule needs an oracle, and no feature available at inference time reproduces it (Appendix A.2). Diffusion faced the same question and answered it by scheduling the scale along the trajectory (Chang et al., 2023; Kynkäänniemi et al., 2024; Wang et al., 2024b). Our reference rule is already a schedule of that kind, over the answer rather than over noise levels; a schedule over the scale itself is the obvious next one.

Localized-CFG. Our negative deletes the whole image, the limit every corruption approaches and the only construction with nothing to tune (Section 3.2). If we knew which region carried the answer, we could delete it alone. Appendix A.7 measures what this *Localized-CFG* could be worth: blacking out the annotated region beats deleting the whole image on two of the three benchmarks we tried, and matches it on the third. That is an upper bound, not a method. Every variant there reads the dataset's own annotation, and when the model predicts the region itself the gain almost disappears. Whether a predicted region can localize the negative stream, rather than the attention, is open.

5 CONCLUSION

Contrastive decoding for vision-language models is classifier-free guidance with a particular reference, and reading it as one equation turns a list of methods into a design space. Most of what diffusion built on that equation does not survive autoregressive decoding: every negative stream we tried points towards deleting the image or fails, and every imported operator is the guidance scale in disguise. The one axis with no diffusion counterpart is where the reference stands, and the field's silent answer leaks the image into the stream meant to be blind to it, collapsing guidance after a single token. Switching the reference inside one answer takes the better of both rules, adding 5.7 points to Qwen2.5-VL-7B and 1.7 to Qwen3-VL-8B at $1.04\times$ the cost of unguided decoding, with no training and one scale everywhere.

AI USE STATEMENT

We used generative AI tools as writing assistants, under author direction and on material whose substance the authors had already fixed. Their use was confined to language and internal consistency: polishing prose, correcting grammar and typography, condensing sections to the page limit, drafting and redrafting passages from author-supplied content and findings, and auditing the manuscript for contradictions, duplicated claims, and numbers that disagreed between the text, the tables and the appendix. Every issue raised this way was checked by the authors against the original experiment logs before any edit was made, and several were rejected as incorrect.

We did not use generative AI to generate research ideas, to design experiments, to write or debug the experimental code, to run or analyse experiments, or to produce any number reported in this paper. All results come from the authors' own runs. No AI system is an author. We have reviewed all AI-assisted text and take full responsibility for the final content of this work, including every claim, number and artifact.

ETHICS STATEMENT

This work studies a decoding rule for existing, publicly released vision-language models. It involves no human subjects and no new data collection: every benchmark we use is public and is used as released, under its own licence and evaluation protocol. We release no model weights and no dataset, and the method requires no training.

Two risks are worth stating plainly. First, guidance amplifies whatever the two streams disagree on. We use it to strengthen the evidence an image carries, and it raises accuracy on the benchmarks we report, but it verifies nothing: it can sharpen a confident error as readily as a correct answer, and it inherits whatever biases the underlying model already carries. A gain on a hallucination or VQA benchmark is not a guarantee of factuality, and we would not deploy this in a safety-critical setting without task-specific validation. Second, an inference-time method that improves accuracy at near-zero extra cost lowers the barrier to deploying vision-language models at scale, and carries the dual-use considerations that come with them.

We have tried to report negative outcomes as carefully as positive ones. The paper names the model on which our method does not gain, reports the constructions we tried that failed, and discloses two defects in our own parameter sweeps (Appendix A).

REPRODUCIBILITY STATEMENT

Section 4 states the complete configuration, fixed in advance and identical on every benchmark and every model: image deletion as the negative stream, the blend on the last hidden state, the switching reference at $K = 8$, and $\alpha = 1.5$ throughout. Nothing is tuned per benchmark, so there are no per-dataset hyperparameters to recover. All five models and all ten benchmarks are public, and every run uses `lmms-eval` (Li et al., 2024) with greedy decoding and each benchmark's default prompt.

Appendix B.8 lists every benchmark with its split and size. Appendix B.7 gives the evaluation protocol and the paired significance tests. Appendix A.4 states and proves the span lemma in full, with the assumptions it needs. Appendix A reports the measurements behind each claim of Section 4.1, together with the arms we tried that did not work and the defects we found in our own sweeps.

Two numerical caveats matter for exact replication. Qwen2.5-VL runs in bf16 and Qwen3-VL in fp16, a choice forced by hardware, and we verified that the two precisions give matching scores. Separately, bf16 rounding can flip a token whose top two logits sit within one unit in the last place, which on continuation tasks diverts the rest of the answer; Appendix A.1 quantifies how often this occurs, and every comparison we draw is paired so that it is not confounded by it.

Our code, including the decoding implementations of the four baselines exactly as we ran them, will be made publicly available upon acceptance.

REFERENCES

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Donghoon Ahn, Hyoungwon Cho, Jaewon Min, Wooseok Jang, Jungwoo Kim, SeonHwa Kim, Hyun Hee Park, Kyong Hwan Jin, and Seungryong Kim. Self-rectifying diffusion sampling with perturbed-attention guidance. In *European Conference on Computer Vision (ECCV)*, 2024.

Donghoon Ahn, Jiwon Kang, Sanghyun Lee, Minjae Kim, Jaewon Min, Wooseok Jang, Sangwu Lee, Sayak Paul, and Seungryong Kim. Where and how to perturb: On the design of perturbation guidance in diffusion and flow models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29915–29926, 2025.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

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

Yuanhao Ban, Ruochen Wang, Tianyi Zhou, Minhao Cheng, Boqing Gong, and Cho-Jui Hsieh. Understanding the impact of negative prompts: When and how do they take effect? In *European Conference on Computer Vision (ECCV)*, 2024.

Arnav Bendre, Guneesh Gupta, Kavish Grover, Chayan Aggarwal, and Shreyansh Modi. Rethinking CD: A reproducibility study and extension on the ineffectiveness of contrastive decoding at mitigating object hallucinations in MLLMs. *arXiv preprint arXiv:2607.25196*, 2026.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. FLUX.1 Kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. SEGA: Instructing text-to-image models using semantic guidance. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Arwen Bradley and Preetum Nakkiran. Classifier-free guidance is a predictor-corrector. *arXiv preprint arXiv:2408.09000*, 2024.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. InstructPix2Pix: Learning to follow image editing instructions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Jinjin Cao, Zhiyang Chen, Zijun Wang, Liyuan Ma, Weijian Luo, and Guojun Qi. When images speak louder: Mitigating language bias-induced hallucinations in VLMs through cross-modal guidance. *arXiv preprint arXiv:2510.10466*, 2025.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T. Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan. Muse: Text-to-image generation via masked generative transformers. In *International Conference on Machine Learning (ICML)*, 2023.

Chongyan Chen, Samreen Anjum, and Danna Gurari. Grounding answers for visual questions asked by visually impaired people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19098–19107, 2022.

Dar-Yen Chen, Hmrishav Bandyopadhyay, Kai Zou, and Yi-Zhe Song. Normalized attention guidance: Universal negative guidance for diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025a. arXiv:2505.21179.

Huayu Chen, Kai Jiang, Kaiwen Zheng, Jianfei Chen, Hang Su, and Jun Zhu. Visual generation without guidance. In *International Conference on Machine Learning (ICML)*, 2025b.

Huayu Chen, Hang Su, Peize Sun, and Jun Zhu. Toward guidance-free AR visual generation via condition contrastive alignment. In *International Conference on Learning Representations (ICLR)*, 2025c.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a.

Wei Chen, Xin Yan, Bin Wen, Fan Yang, Tingting Gao, Di Zhang, and Long Chen. Decoupling contrastive decoding: Robust hallucination mitigation in multimodal large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025d.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-Pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025e.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. HALC: Object hallucination reduction via adaptive focal-contrast decoding. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 7824–7846, 2024b.

Muthu Chidambaram, Khashayar Gatmiry, Sitan Chen, Holden Lee, and Jianfeng Lu. What does guidance do? A fine-grained analysis in a simple setting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 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. In *International Conference on Learning Representations (ICLR)*, 2024.

Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. CFG++: Manifold-constrained classifier free guidance for diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yasser Dahou, Ngoc Dung Huynh, Phuc H. Le-Khac, Wamiq Reyaz Para, Ankit Singh, and Sanath Narayan. Vision-language models can't see the obvious. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24159–24169, 2025.

Prafulla Dhariwal and Alex Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Hongyuan Dong, Jiawen Li, Bohong Wu, Jiacong Wang, Yuan Zhang, and Haoyuan Guo. Benchmarking and improving detail image caption. *arXiv preprint arXiv:2405.19092*, 2024.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning (ICML)*, 2024.

Weichen Fan, Amber Yijia Zheng, Raymond A. Yeh, and Ziwei Liu. CFG-Zero*: Improved classifier-free guidance for flow matching models. *arXiv preprint arXiv:2503.18886*, 2025.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14303–14312, 2024.

Ling Fu, Zhebin Kuang, Jiajun Song, Mingxin Huang, Biao Yang, Yuzhe Li, Linghao Zhu, Qidi Luo, Xinyu Wang, Hao Lu, Zhang Li, Guozhi Tang, Bin Shan, Chunhui Lin, Qi Liu, Binghong Wu, Hao Feng, Hao Liu, Can Huang, Jingqun Tang, Wei Chen, Lianwen Jin, Yuliang Liu, and Xiang Bai. OCRBench v2: An improved benchmark for evaluating large multimodal models on visual text localization and reasoning. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. BLINK: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision (ECCV)*, 2024.

Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-A-Scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision (ECCV)*, 2022.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Xuan Gong, Tianshi Ming, Xinpeng Wang, and Zhihua Wei. DAMRO: Dive into the attention mechanism of LVLM to reduce object hallucination. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7696–7712, 2024.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. HallusionBench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385, 2024.

Wenqi Guo and Shan Du. VSF: Simple, efficient, and effective negative guidance in few-step image generation models by value sign flip. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2508.10931.

Danna Gurari, Qing Li, Abigale J. Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. VizWiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. A short version appeared in the NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications.

Susung Hong. Smoothed energy guidance: Guiding diffusion models with reduced energy curvature of attention. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Susung Hong, Gyuseong Lee, Wooseok Jang, and Seungryong Kim. Improving sample quality of diffusion models using self-attention guidance. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 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 (CVPR)*, 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. In *International Conference on Learning Representations (ICLR)*, 2025.

Junha Hyung, Kinam Kim, Susung Hong, Min-Jung Kim, and Jaegul Choo. Spatiotemporal skip guidance for enhanced video diffusion sampling. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Ting Jiang, Shaohan Huang, Zhongzhi Luan, Deqing Wang, and Fuzhen Zhuang. Scaling sentence embeddings with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 3182–3196, 2024.

Yujin Jo, Sangyoon Bae, and Taesup Kim. Attention-space contrastive guidance for efficient hallucination mitigation in LVLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 9706–9715, 2026.

Zhehan Kan, Ce Zhang, Zihan Liao, Yapeng Tian, Wenming Yang, Junyuan Xiao, Xu Li, Dongmei Jiang, Yaowei Wang, and Qingmin Liao. CATCH: Complementary adaptive token-level contrastive decoding to mitigate hallucinations in LVLMs. *arXiv preprint arXiv:2411.12713*, 2024.

Tero Karras, Miika Aittala, Tuomas Kynkäänniemi, Jaakko Lehtinen, Timo Aila, and Samuli Laine. Guiding a diffusion model with a bad version of itself. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision*, 128(7):1956–1981, 2020.

Mingi Kwon, Shin seong Kim, Jaeseok Jeong, Yi Ting Hsiao, and Youngjung Uh. TCFG: Tangential damping classifier-free guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2620–2629, 2025.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. LISA: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9579–9589, 2024.

Krunoslav Lehman Pavasovic, Jakob Verbeek, Giulio Biroli, and Marc Mezard. Classifier-free guidance: From high-dimensional analysis to generalized guidance forms. *arXiv preprint arXiv:2502.07849*, 2025.

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

Bo Li, Peiyuan Zhang, Kaichen Zhang, Fanyi Pu, Xinrun Du, Yuhao Dong, Haotian Liu, Yuanhan Zhang, Ge Zhang, Chunyuan Li, and Ziwei Liu. LMMs-Eval: Accelerating the development of large multimodal models, March 2024. URL https://github.com/EvolvingLMMs-Lab/lmms-eval.

Pengxiang Li, Shilin Yan, Joey Tsai, Renrui Zhang, Ruichuan An, Ziyu Guo, and Xiaowei Gao. Adaptive classifier-free guidance via dynamic low-confidence masking. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12286–12312, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 292–305, 2023b.

Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D. Plumbley. AudioLDM: Text-to-audio generation with latent diffusion models. In *International Conference on Machine Learning (ICML)*, 2023.

Jiazhen Liu, Yuhan Fu, Ruobing Xie, Runquan Xie, Xingwu Sun, Fengzong Lian, Zhanhui Kang, and Xirong Li. PhD: A ChatGPT-prompted visual hallucination evaluation dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19857–19866, 2025.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B. Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision (ECCV)*, 2022.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in LVLMs. In *European Conference on Computer Vision (ECCV)*, 2024.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. Fine-tuning LLaMA for multi-stage text retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik P. Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Viet Nguyen, Anh Nguyen, Trung Dao, Khoi Nguyen, Cuong Pham, Toan Tran, and Anh Tran. Supercharged one-step text-to-image diffusion models with negative prompts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. arXiv:2412.02687 (first titled "SNOOPI: Supercharged One-step Diffusion Distillation with Proper Guidance").

Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning (ICML)*, 2022.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, David Yan, Dhruv Choudhary, Dingkang Wang, Geet Sethi, Guan Pang, Haoyu Ma, Ishan Misra, Ji Hou, Jialiang Wang, Kiran Jagadeesh, Kunpeng Li, Luxin Zhang, Mannat Singh, Mary Williamson, Matt Le, Matthew Yu, Mitesh Kumar Singh, Peizhao Zhang, Peter Vajda, Quentin Duval, Rohit Girdhar, Roshan Sumbaly, Sai Saketh Rambhatla, Sam Tsai, Samaneh Azadi, Samyak Datta, Sanyuan Chen, Sean Bell, Sharadh Ramaswamy, Shelly Sheynin, Siddharth Bhattacharya, Simran Motwani, Tao Xu, Tianhe Li, Tingbo Hou, Wei-Ning Hsu, Xi Yin, Xiaoliang Dai, Yaniv Taigman, Yaqiao Luo, Yen-Cheng Liu, Yi-Chiao Wu, Yue Zhao, Yuval Kirstain, Zecheng He, Zijian He, Albert Pumarola, Ali Thabet, Artsiom Sanakoyeu, Arun Mallya, Baishan Guo, Boris Araya, Breena Kerr, Carleigh Wood, Ce Liu, Cen Peng, Dimitry Vengertsev, Edgar Schonfeld, Elliot Blanchard, Felix Juefei-Xu, Fraylie Nord, Jeff Liang, John Hoffman, Jonas Kohler, Kaolin Fire, Karthik Sivakumar, Lawrence Chen, Licheng Yu, Luya Gao, Markos Georgopoulos, Rashel Moritz, Sara K. Sampson, Shikai Li, Simone Parmeggiani, Steve Fine, Tara Fowler, Vladan Petrovic, and Yuming Du. Movie Gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. DreamFusion: Text-to-3D using 2D diffusion. In *International Conference on Learning Representations (ICLR)*, 2023.

Qwen Team. Qwen3.5 model card. `https://huggingface.co/Qwen/Qwen3.5-397B-A17B`, 2026.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 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 (EMNLP)*, pp. 4035–4045, 2018.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Seyedmorteza Sadat, Jakob Buhmann, Derek Bradley, Otmar Hilliges, and Romann M. Weber. CADS: Unleashing the diversity of diffusion models through condition-annealed sampling. In *International Conference on Learning Representations (ICLR)*, 2024.

Seyedmorteza Sadat, Otmar Hilliges, and Romann M. Weber. Eliminating oversaturation and artifacts of high guidance scales in diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025a.

Seyedmorteza Sadat, Manuel Kansy, Otmar Hilliges, and Romann M. Weber. No training, no problem: Rethinking classifier-free guidance for diffusion models. In *International Conference on Learning Representations (ICLR)*, 2025b.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Guillaume Sanchez, Alexander Spangher, Honglu Fan, Elad Levi, and Stella Biderman. Stay on topic with classifier-free guidance. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 43197–43234, 2024.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 783–791, 2024.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024a.

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

Lexiang Tang, Weihao Gao, Bingchen Zhao, Lu Ma, Qiao Jin, Bang Yang, and Yuexian Zou. Thinking by subtraction: Confidence-driven contrastive decoding for LLM reasoning. *arXiv preprint arXiv:2602.18232*, 2026.

Zhicong Tang, Jianmin Bao, Dong Chen, and Baining Guo. Diffusion models without classifier-free guidance. *arXiv preprint arXiv:2502.12154*, 2025.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Bingkui Tong, Jiaer Xia, and Kaiyang Zhou. Mitigating hallucination in multimodal LLMs with layer contrastive decoding. *arXiv preprint arXiv:2509.25177*, 2025.

Alexander Vogel, Omar Moured, Yufan Chen, Jiaming Zhang, and Rainer Stiefelhagen. RefChartQA: Grounding visual answer on chart images through instruction tuning. In *Document Analysis and Recognition – ICDAR 2025*, volume 16026 of *Lecture Notes in Computer Science*, pp. 523–537. Springer, 2025.

Haochen Wang, Xiangtai Li, Zilong Huang, Anran Wang, Jiacong Wang, Tao Zhang, Jiani Zheng, Sule Bai, Zijian Kang, Jiashi Feng, Zhuochen Wang, and Zhaoxiang Zhang. Traceable evidence enhanced visual grounded reasoning: Evaluation and method. In *International Conference on Learning Representations (ICLR)*, 2026.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. AMBER: An LLM-free multi-dimensional benchmark for MLLMs hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11897–11916, 2024a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Xi Wang, Nicolas Dufour, Nefeli Andreou, Marie-Paule Cani, Victoria Fernandez Abrevaya, David Picard, and Vicky Kalogeiton. Analysis of classifier-free guidance weight schedulers. *Transactions on Machine Learning Research (TMLR)*, 2024b.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024c.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15840–15853, 2024d.

Sangmin Woo, Jaehyuk Jang, Donguk Kim, Yubin Choi, and Changick Kim. RITUAL: Random image transformations as a universal anti-hallucination lever in large vision language models. *arXiv preprint arXiv:2405.17821*, 2024.

Penghao Wu and Saining Xie. V^*: Guided visual search as a core mechanism in multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13084–13094, 2024.

Yuchen Wu, Minshuo Chen, Zihao Li, Mengdi Wang, and Yuting Wei. Theoretical insights for diffusion guidance: A case study for Gaussian mixture models. In *International Conference on Machine Learning (ICML)*, 2024.

Yuxuan Xia, Siheng Wang, and Peng Li. SDCD: Structure-disrupted contrastive decoding for mitigating hallucinations in large vision-language models. *arXiv preprint arXiv:2601.03500*, 2026.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One single transformer to unify multimodal understanding and generation. In *International Conference on Learning Representations (ICLR)*, 2025.

Dongli Xu, Aleksei Tiulpin, and Matthew B. Blaschko. SoftCFG: Uncertainty-guided stable guidance for visual autoregressive model. In *International Conference on Learning Representations (ICLR)*, 2026.

Hao Yin, Guangzong Si, and Zilei Wang. ClearSight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14625–14634, 2025a.

Hao Yin, Guangzong Si, and Zilei Wang. The mirage of performance gains: Why contrastive decoding fails to mitigate object hallucinations in MLLMs? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025b.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *Transactions on Machine Learning Research (TMLR)*, 2022.

Xin Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Song-Hai Zhang, and Xiaojuan Qi. Text-to-3D with classifier score distillation. In *International Conference on Learning Representations (ICLR)*, 2024.

Kai Zhang, Fujun Luan, Sai Bi, and Jianming Zhang. EP-CFG: Energy-preserving classifier-free guidance. *arXiv preprint arXiv:2412.09966*, 2024a.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, Peng Gao, and Hongsheng Li. MathVerse: Does your multi-modal LLM truly see the diagrams in visual math problems? In *European Conference on Computer Vision (ECCV)*, Lecture Notes in Computer Science, pp. 169–186. Springer, 2024b.

Yi-Fan Zhang, Huanyu Zhang, Haochen Tian, Chaoyou Fu, Shuangqing Zhang, Junfei Wu, Feng Li, Kun Wang, Qingsong Wen, Zhang Zhang, Liang Wang, and Rong Jin. MME-RealWorld: Could your multimodal LLM challenge high-resolution real-world scenarios that are difficult for humans? In *International Conference on Learning Representations (ICLR)*, 2025a.

YiFan Zhang, Yang Shi, Weichen Yu, Qingsong Wen, Xue Wang, Wenjing Yang, Zhang Zhang, Liang Wang, and Rong Jin. Debiasing multimodal large language models via penalization of language priors. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM)*, 2025b. doi: 10.1145/3746027.3755364.

Yuechen Zhang, Shengju Qian, Bohao Peng, Shu Liu, and Jiaya Jia. Prompt highlighter: Interactive control for multi-modal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13215–13224, 2024c.

Candi Zheng and Yuan Lan. Characteristic guidance: Non-linear correction for diffusion model at large guidance scale. In *International Conference on Machine Learning (ICML)*, 2024.

Lanyun Zhu, Deyi Ji, Tianrun Chen, Peng Xu, Jieping Ye, and Jun Liu. IBD: Alleviating hallucinations in large vision-language models via image-biased decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 1624–1633, 2025.

A Details of the analysis

This appendix supplements Section 4.1. Appendix A.1 settles the one axis the main text states in a line rather than a section, where the guidance equation acts; the rest follows the order of Section 4.1: where the reference stands (Section 4.1.1), the negative stream (Section 4.1.2), and the geometry of the update (Section 4.1.3).

A.1 Where the equation acts

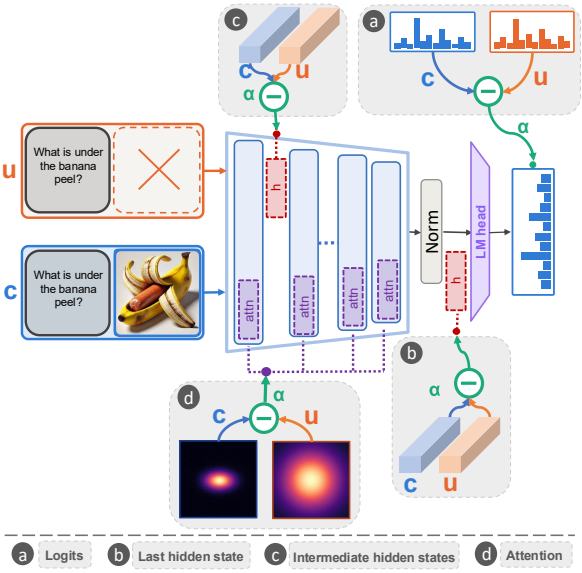

Figure 4: **Four feature spaces for guidance.** The conditional stream c sees the image and the negative stream u does not. Their blend can act on (a) the logits, (b) the last hidden state, (c) an intermediate hidden state, or (d) the attention to the image.

The four spaces, and why two of them are ruled out. Classifier-free guidance was defined on the output of the denoising network (Ho & Salimans, 2022), and recent diffusion work also extrapolates inside attention (Chen et al., 2025a; Nguyen et al., 2025; Guo & Du, 2026), as does one work for vision-language models (Jo et al., 2026). Of the four spaces of Figure 4, the logits and the last hidden state are one: the language model head is a bias free linear map W applied after the final normalisation, and a linear map commutes with an affine combination, $W\big(h_u + \alpha(h_c - h_u)\big) = \ell_u + \alpha(\ell_c - \ell_u)$, with h_c, h_u the two streams' normalised states. A bias in the head passes through unchanged, because the weights α and $1-\alpha$ sum to one; it would not for the operators of Section 3.4, whose weights do not (Appendix A.4). An intermediate hidden state is not guidance at all, since below the readout the network recomputes its answer from whatever is written there, and the axis therefore resolves against the leverage: by every measure we have, the readout is the weakest place in the stack to intervene, because the two streams have nearly converged by the output, and it is also the only place where the operator is defined. A well posed small lever beats an ill posed large one. In attention space a second restriction follows from the simplex: a blended attention row has to be renormalised, and the natural choice restores the image block's own total, which makes the edit a pure redistribution within the image, able to change where the model looks but never how much. That is the choice the published implementations make, and it leaves guidance nearly inert; releasing the image mass does let it act, but only on the one benchmark whose answer is a continuation. The negative attention can build natively, the conditional's own blurred map, carries no signal either, and the best variant falls short of the last hidden state at twice the cost (Table 7).

Why the two equivalent sites differ at all. That identity is exact in real arithmetic, and the two sites agree to within 0.25 points on every benchmark of Table 7. The two implementations differ only in the order of a bf16 rounding: one blends 3584 coordinates and then multiplies by the $152\text{k} \times 3584$ head, the other multiplies twice and then blends 152k coordinates. Greedy decoding then discretises

the rounding: whenever the top two logits sit within it, the token flips and the continuation can diverge. A duplicate run gives 0 flips on V* (191) and on TreeBench (405), so the flips are not run to run jitter. Changing the site changes the answer on 2 of 191 and 10 of 405 documents, and its correctness on 0 and 5 (three one way, two the other, $p = 1.0$). The flips sit exactly where bf16 runs out of resolution: on TreeBench every flipped document has a margin of at most 0.188 between its top two guided logits (median 0.094), against a median of 1.250 on the documents that do not flip, and these margins come in steps of 0.0625, one bf16 unit in the last place at this logit scale. The flipped documents are 0 to 3 such units from a tie, and only 13.6% of documents are close enough to be at risk at all. On VizWiz, where answers are continuations, the same rounding can also fire the end of sequence token early and diverge the rest of the answer (*milk thistle dietary supplement* becomes *milk thistle*); the single token suites have no continuation to diverge, which is why they are stable run to run.

Why we guide the last hidden state. Since the two sites make the same decisions, the choice rests on what each space allows. First, only the hidden state gives the guidance vector a geometry. Softmax ignores a constant added to every logit, so a guidance vector in logit space carries a component that no decision can see, 60% to 97% of its length depending on the benchmark, and its 152k coordinates carry no natural metric. The operators of Section 3.4, which rescale or project Δ by its length and angle, are well defined only in the hidden space. Second, the hidden state is 42 times smaller to store, 3584 coordinates against 152k. Third, because the readout is linear, one stored run replays the whole ladder of guidance scales offline on single token benchmarks, exactly in arithmetic and to within one document of the live run, which lets us treat the scale as a free axis.

What the four sites score. One block below the readout, at layer 27, guidance collapses accuracy by 51 points on V* and 27 on TreeBench, measured against guidance at the readout (Table 7). The collapse is not the decision: guiding only the first generated token restores accuracy to within noise of the readout, while the collapsed variants emit 10 to 29 words on tasks whose answers are one or two. Rebuilding the guidance vector from the two streams' own layer 27 states makes the collapse deeper on all three benchmarks, and a ladder of magnitudes expressed as a fraction of the layer's norm collapses at every point. In attention space, holding the image block's total attention fixed leaves guidance worth $+0.2$ points on VizWiz; releasing it raises the gain to $+2.1$ at $\alpha = 3$, still 2.2 points below the last hidden state ($p = 0.004$) and at 2.05 times the time of unguided decoding against 1.09 for a hidden state blend. Releasing the mass is a VizWiz effect: on V* and TreeBench the free mass variants stay within half a point of unguided decoding at every scale, with no trend in α. The variants in Table 7 come from a campaign whose own $\alpha = 1$ run scores 73.39 on VizWiz rather than the 73.80 the table reports; against that baseline the two attention gains read $+0.6$ and $+2.5$.

Table 7: **Where the guidance equation acts.** Accuracy (%), image deletion, $\alpha = 1.5$ except the attention row ($\alpha = 3$, with the image block's total mass left free); layer 27 is one block below the readout.

site the blend acts in	V*	TreeBench	VizWiz
no guidance	*70.16*	*40.99*	*73.80*
logits	70.68	42.72	77.83
last hidden state	70.68	42.96	78.06
layer 27	19.37	15.80	37.04
attention probabilities	70.16	41.48	75.85

Three variants that must not be confused. An *early exit* builds the negative stream from an intermediate layer and blends at the final hidden state; it varies the negative stream and belongs to Section 3.2. A *layer injection* adds the final layer guidance vector at layer L. A *local injection* builds the vector from the two streams' own layer L states and adds it there. The last two also differ in their coefficient: the default path injects $(\alpha - 1)\Delta$ and the local path $\alpha\Delta_L$, a factor of three apart at $\alpha = 1.5$. Every number in Table 7 uses the matched coefficient.

The vector is in the wrong basis, and fixing that does not help. A guidance vector computed at the readout is expressed in a basis the residual stream does not share: $\cos(h_L, h_{\text{final}}) < 0.07$ for every $L < 22$, and at $L = 1$ the vector is 7.4 times the norm of the whole residual state and orthogonal to it ($\cos = 0.005$). Rebuilding the vector in the target layer, or sweeping its magnitude as a fraction τ of that layer's norm, still collapses: at $\tau = 0.3$ the three benchmarks score 6.81, 14.07 and 17.52, and at $\tau = 1.2$ they score 0.00, 7.41 and 2.49.

A retraction. We first read the layer 27 collapse as depth washing out the correction. Classifying the outputs showed otherwise: the first token matches the readout variant on 18 of 20 documents, 11 of 50 outputs are truncated prefixes, 21 of 50 share a first character and then diverge (*laptop* becomes *lado*), and mean length falls from 1.70 to 1.06 words. The end of sequence token fires too early. A second retraction concerns attention space: an earlier implementation blended postsoftmax probabilities into presoftmax scores, which reduces to a temperature change rather than guidance. Under the corrected blend the generation length stays flat at 1.53 words up to $\alpha = 3$. We report the corrected behaviour only.

A.2 Where the reference stands

The probe. The probe decodes the conditional stream without guidance and, at every step, records what each reference would push. With $w^\star$ the unguided argmax, $\Delta^\star_{\text{TF}}(t) = \log p_c(w^\star \mid y_{<t}) - \log p_u(w^\star \mid y_{<t})$ and $\Delta^\star_{\text{I-1}}(t) = \log p_c(w^\star \mid y_{<t}) - \log p_u(w^\star \mid \emptyset)$. Both references are measured on the same trajectory, which makes the comparison controlled. We measure the push on the decision token rather than the norm of the whole guidance vector: on OCRBench v2 that norm runs from 1548 to 2286, flat or rising along the answer, because it is dominated by about 152k tail coordinates that no decision depends on. Table 8 gives the cliff, its depth, and the prefix swap control of Section 4.1.1. Figure 3a uses a second probe run on 210 OCRBench v2 documents, where the drop at the first token is 65% against 63% here.

Table 8: **Guidance at the decision token** ($\Delta^\star$, nats). **(a)** The cliff between step 0 and step 1. **(b)** Depth on OCRBench v2. **(c)** The prefix swap control: mean over steps after the first, holding length and step fixed and changing only what the reference reads. **(d)** The push on the decision token against the spread of pushes over the conditional's top 50 rivals (SNR), and the reference's excess entropy over a free running image free stream, in nats; each range runs from the first generated token to the deepest step bucket.

(a) the cliff

	n	TF @0	TF @1	drop	I-1 @1
OCRBench v2	100	4.676	1.720	−63%	15.793
VizWiz	200	11.646	2.506	−78%	19.603
RefChartQA	200	5.149	2.724	−47%	11.905

(b) depth, OCRBench v2

step	TF	I-1	ratio
0	4.676	4.676	1.0
1	1.720	15.793	9.2
8 to 15	1.742	18.776	10.8
64 to 127	1.012	19.757	19.5
128 to 255	0.893	20.476	22.9

(c) what the reference reads

	nothing	another document	this document, shuffled	this document	generic share	leak share
OCRBench v2	18.354	15.053	7.553	1.475	20%	80%
VizWiz	19.704	10.747	4.693	1.967	51%	49%
RefChartQA	12.581	7.844	3.564	1.781	44%	56%

(d) signal to noise, and the cost of standing on a foreign prefix

	SNR, TF @0	SNR, TF 1 to deepest	SNR, I-1 1 to deepest	excess H, TF	excess H, I-1
OCRBench v2	1.19	0.58 to 0.29	4.52 to 5.96	0.43 to 0.82	0.64 to 1.08
VizWiz	2.90	0.77 to 0.33	5.67 to 6.62	0.29 to 1.49	−0.27 to 0.28
RefChartQA	1.37	0.78 to 0.41	3.03 to 3.77	0.35 to 0.73	2.07 to 2.34

The prefix window. Section 4.1.1 reports the sweep that lets the reference read only the last w generated tokens; Table 9 gives it in full, with $w = 0$ the I-1 rule and $w = \infty$ the TF one. Beyond the first token the curve is flat: on RefChartQA four tokens bring 97% of TF's advantage. On OCRBench v2 the sweep is complete and uninformative, its total range of 0.72 points falling below its own plateau scatter of 0.40 points, which is what the mechanism predicts for a benchmark where I-1 is inert and TF costs less than a point.

What the second stream supplies. The strength oracle of Table 3 shows that on long answers the I-1 direction is right and only its strength is wrong. Table 10 asks whether that strength can be set without the teacher forced stream. It cannot. The oracle, which reads the strength from that stream token by token, comes within 1.3 points of TF on MathVerse, a difference that is not significant; it also ties TF on ReasonSeg and OCRBench v2 and beats it on VizWiz. It loses only where the

Table 9: **The prefix window.** Score when the reference reads only the last w generated tokens; $w = 0$ is I-1 and $w = \infty$ is TF. The end points on VizWiz are the runs of Table 3; the other columns come from an earlier run of the same rule.

w	0	1	2	4	8	16	∞
VizWiz	78.06	75.47	75.69	76.09	75.98	76.07	76.01
RefChartQA	82.44	83.63	83.69	84.16	84.19	84.19	84.22

Table 10: **What the second stream supplies** (MathVerse). Every row except TF keeps the I-1 direction and changes only its strength.

	second stream	score
no guidance	–	60.28
TF	yes	62.82
I-1	no	56.01
I-1, strength oracle	yes	61.55
I-1, first token only	no	59.49
I-1, capped at step 0	no	59.39
I-1, decay schedule	no	58.83

Table 11: **The switch point.** SWITCH at switch point K against the two rules it switches between. $K = 16$ and 32 were not run on the short answer benchmarks, where the switch fires on at most 1.5% of answers at $K = 8$. At $K = 32$ only 53.5% of PhD rationales reach the switch.

	MathVerse	PhD-rat.	VizWiz	OK-VQA	O3
TF	62.82	71.98	76.01	52.76	58.68
I-1	56.01	69.99	78.06	54.00	63.61
SWITCH, $K = 4$	61.93	71.98	77.51	54.00	63.85
SWITCH, $K = 8$	62.54	72.78	78.06	54.00	63.61
SWITCH, $K = 16$	61.55	71.91	–	–	–
SWITCH, $K = 32$	63.47	69.79	–	–	–

guided gain is the stop decision itself: because it keeps the I-1 direction it keeps I-1's stopping behaviour, and it falls 1.77 points below TF on O3 and 3.65 on PhD with a rationale. Strength reproduces what TF does to the content of an answer, not what it does to its length. Nor is the late push inert everywhere: on OCRBench v2's pre-registered readout subtasks ($n = 692$) every rule that keeps guiding late gains 3.9 to 6.1 points, while the oracle, which removes it, gains 0.23 and is not significant.

Why no single stream schedule reaches it. The three schedules that use one stream all fall 3.3 to 4 points short of TF: guiding the first token only, capping the push at its step 0 size, and the decay schedule, which rescales the push by the average teacher forced decay. The cap shows why: it binds on 99.7% of steps and cuts the push to a fifth, so it removes the long answer harm by removing the guidance. The right strength changes from token to token, and only the second stream measures it. SWITCH uses that stream directly, once the cached reference has gone stale.

The switch. At the switch the reference is advanced from token 0 to token K, after which SWITCH is identical to TF; $K = 1$ reproduces TF and $K = 10000$ reproduces I-1 on 30 of 30 answers each. Table 11 sweeps the switch point, and Table 12 gives the cost of each rule. Advancing the reference one token at a time and advancing it over the K queued tokens in a single batched pass are the same computation in exact arithmetic; they differ only in the order in which the kernel reduces, and that drift changed answers at tokens 67 to 295 on 6 of 15 MathVerse documents. The accuracy tables use the sequential build, which is the one that is bitwise identical to TF after the switch and so passes the invariant above. The costs in Table 12 are the batched ones, computed from exact forward counts rather than wall clocked: the reference stream issues T passes under TF, one under I-1, and $T-K+1$ under SWITCH once the answer reaches the switch, at a measured price of 0.74 conditional forwards per reference forward, the reference pass carrying no image tokens and a much shorter cache.

Handing a decision back to the conditional. Two variants of guided decoding let the conditional stream alone make part of the decision (Table 13). The *EOS guard* takes the decision to stop from the conditional and lets guidance choose only among the other tokens. The *adaptive plausibility constraint* (APC) of contrastive decoding (Li et al., 2023a; Leng et al., 2024) restricts the guided choice to tokens the conditional already finds plausible, $p_c(y) \geq 0.1 \max p_c$. The EOS guard hands back one decision, the stop; APC hands back every decision where guidance would pick an implausible token. The EOS guard confirms that the stop decision is where the reference rules differ. It lifts TF on VizWiz from 76.01 to 78.21, the level of I-1, because the whole deficit of TF there is answers that run on (*ketchup* becomes *ketchup packet*). On lists it does the opposite: stopping early is the gain on O3, and the guard costs I-1 9.5 points there. APC acts as a veto. It changes the token on only 0.04% to 7% of steps, but a changed token rewrites the rest of the answer, and among changed answers it restores the unguided one on 74% of VizWiz, 78% of OK-VQA and all PhD answers.

Table 12: **End-to-end decoding cost**, relative to unguided decoding, on benchmarks ordered by mean answer length. Prefill is included: a request costs one conditional prefill plus T decode steps, and each method is charged for whatever second stream it runs. A conditional prefill is priced at 22 decode steps (vision tower plus 2k prompt tokens at the 1.6 Mpx cap, estimated for one A100), and a negative forward that carries no vision tokens at 0.74 of a conditional one. TF runs that second forward at every step, I-1 reuses one cached pass, and SWITCH pays I-1's cost until the switch fires. VCD (Leng et al., 2024) and SID (Huo et al., 2025) instead re-encode the image and carry every vision token in their negative stream, so each of their forwards — prefill included — costs a full conditional pass; their cost is $2 - 1/(P + T)$ for *any* prefill price P, whereas ours falls towards 1 as P grows. The rows below price forward passes only. M3ID's negative is image deletion, exactly as ours is, so its forward count is TF's. †PAI's negative is also image free, and its attention edit is switched off while that pass runs, so its forward count too is TF's; but the edit forces an eager kernel on the conditional stream, where the cost is. Priced at the 1.88× we measure for our own attention space route (2.05× against 1.09× for a fused deletion route) that puts PAI near 2.05× on a short answer, which is what Table 6 reports. SID's negative cannot use a fused kernel either, but that penalty lands on prefill and memory rather than per-token decode and is not priced here.

	OK-VQA	VizWiz	O3	PhD-rat.	ReasonSeg	OCRBench	MathVerse
# tokens	*2.4*	*2.5*	*3.8*	*30.2*	*60.9*	*152.1*	*337.7*
TF	1.08×	1.09×	1.12×	1.43×	1.55×	1.65×	1.70×
I-1	1.04×	1.04×	1.04×	1.02×	1.01×	1.01×	1.00×
SWITCH	1.04×	1.04×	1.04×	1.35×	1.49×	1.62×	1.68×
M3ID (Favero et al., 2024)	1.08×	1.09×	1.12×	1.43×	1.55×	1.65×	1.70×
PAI†	1.08×	1.09×	1.12×	1.43×	1.55×	1.65×	1.70×
VCD	1.96×	1.96×	1.96×	1.98×	1.99×	1.99×	2.00×
SID†	2.00×	2.00×	2.00×	2.00×	2.00×	2.00×	2.00×

That helps where guidance's largest corrections are errors (TF on VizWiz, OCRBench v2) and hurts where they are the point (OK-VQA, where I-1 loses 2.76 points). Neither variant helps everywhere, so the main results use neither.

Table 13: **Every reference rule with the two decision token variants.** The rules of Table 3, each with an EOS guard (+EOS: the stop decision is taken from the conditional alone and guidance chooses among the other tokens) and with the adaptive plausibility constraint (+APC). Qwen2.5-VL-7B-Instruct, greedy, $\alpha = 1.5$, image deletion; SWITCH uses $K = 8$; benchmarks ordered by answer length, MathVerse pooled over its five versions.

	OK-VQA	VizWiz	O3	PhD-rat.	ReasonSeg	OCRBench	MathVerse
no guidance	42.60	73.80	57.34	68.73	49.45	53.69	60.28
TF	52.76	76.01	58.68	71.98	52.01	52.96	62.82
TF+EOS	53.96	78.21	55.71	71.98	52.01	52.84	62.82
TF+APC	50.60	77.46	58.91	72.51	51.87	53.92	62.58
I-1	54.00	78.06	63.61	69.99	50.18	53.83	56.01
I-1+EOS	54.08	78.05	54.13	69.79	49.93	53.81	55.77
I-1+APC	51.24	78.01	63.92	69.19	49.87	54.23	57.23
SWITCH	54.00	78.06	63.61	72.78	51.86	53.77	62.54
SWITCH+APC	54.00	78.01	63.94	72.71	51.72	55.07	61.83

What we falsified. Four repairs we expected to work do not, and each rules out a class of fix. The norm of the guidance vector is not a measure of guidance, being flat or rising along the answer, which is why the cap, whose signal is that norm, removes the harm by removing the guidance. Sharpening the reference does not help either: guidance subtracts $(\alpha - 1)\ell_u$, so a sharper reference re-ranks more, not less. Gating guidance on disagreement with the image free stream is significantly worse than both rules on ReasonSeg, DetailCaps and AMBER. And routing per document between the two rules has a real ceiling, an oracle gaining 1.87 and 1.37 points against a 0.26 point noise floor, but no inference time feature we tried survives correction for multiple comparisons; gating on the conditional's confidence fails outright, because I-1's long answer harm is spread uniformly across confidence levels and tokens the conditional already gives at least 0.99 probability are never flipped. A fixed evidence budget, which would have predicted the smooth decay we first looked for, is refuted in Appendix B.3.

A.3 THE NEGATIVE STREAM

Every corruption of the image points towards its deletion. Section 3.2 argues that input corruptions form a cone around deletion. Measured on V*, TreeBench and VizWiz, the alignment of a corruption's guidance vector with deletion's rises with the corruption's strength (Pearson $+0.74$, $+0.81$ and $+0.97$) while its spread over documents narrows (Figure 5). The noised image of visual contrastive decoding has the shortest and least aligned vector ($\cos = 0.18$ on V*; Table 14). Tuning the scale per corruption closes the gap to deletion by 2.2 to 3.7 times, but the gap that remains follows the sine of the angle (Figure 6).

Table 14: **Nine constructions of the negative stream on V***, ordered by alignment with deletion. $r = \|\Delta\|/\|h_c\|$; $\alpha^\star$ is the scale whose guided distribution is closest to deletion's at $\alpha_0 = 1.5$; KL is measured at $\alpha^\star$ and at 1.5. The noised image of visual contrastive decoding is the shortest and least aligned vector.

negative	r	$\cos\theta$	$\sin\theta$	$\alpha^\star$	KL at $\alpha^\star$	KL at 1.5
noised (VCD)	0.114	0.175	0.950	1.35	0.0054	0.0143
grayscale	0.180	0.275	0.937	1.25	0.0036	0.0134
resized	0.217	0.389	0.916	1.35	0.0033	0.0092
masked	0.208	0.478	0.872	1.40	0.0021	0.0078
inverted	0.203	0.534	0.839	1.35	0.0017	0.0064
black	0.238	0.586	0.810	1.35	0.0014	0.0061
blurred	0.507	0.634	0.773	1.40	0.0012	0.0052
shuffled	0.358	0.804	0.595	1.45	0.0011	0.0039
deleted	–	1	0	1.5	0	0

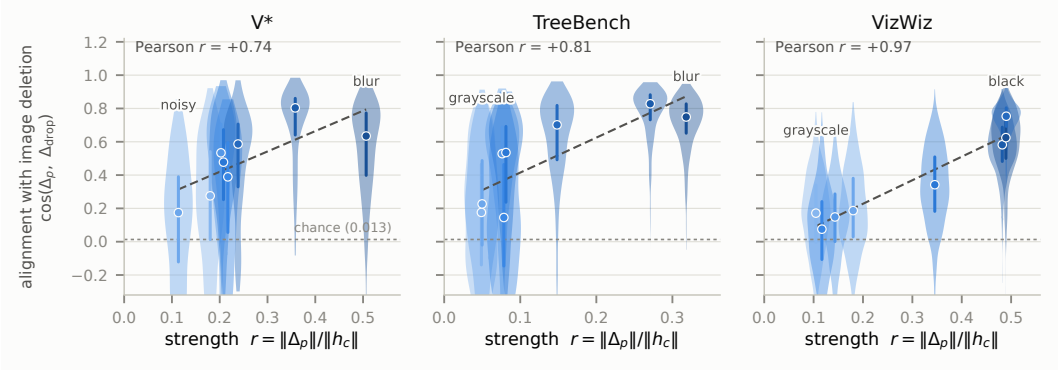

Figure 5: **Every corruption of the image points towards its deletion.** For eight corruptions of the negative image, the alignment of each document's guidance vector with the deletion direction, plotted against the corruption's strength $r = \|\Delta\|/\|h_c\|$. Violins show the distribution over documents; the dashed line fits the medians; chance is 0.013. Alignment rises with strength and the spread narrows: the family is a cone that contracts onto deletion.

Why length and angle separate. Matching a construction's guidance step to deletion's, $\min_\alpha \|(\alpha-1)\Delta_p - (\alpha_0-1)\Delta_{\text{drop}}\|$, gives $\alpha^\star - 1 = (\alpha_0-1)\cos\theta\,\|\Delta_{\text{drop}}\|/\|\Delta_p\|$ with a residual proportional to $\sin\theta$. The scale can remove the length mismatch and nothing else. Under a leave one construction out fit, accuracy as a function of the guidance step $(\alpha-1)\,r$ predicts no better than the global mean (held out error 1.52, 1.23 and 1.53 points against 1.46, 1.19 and 1.82), so the construction axis is not a disguised scale.

Why a weaker model fails. Table 15 compares each literal weak model with the conditional at the first generated token. The smaller model is too close: on VizWiz its guidance vector is 18 times shorter than deletion's and it agrees with the conditional on 62% of documents, against 1%. The early exit is too far: at either exit layer it agrees on none of 1,727 documents. The comparison is decisive on VizWiz. On the multiple choice suites deletion itself agrees with the conditional on about half the documents, because the language prior often picks the same option letter, so only the ordering of the rows carries over, not their values. The 3B model is also separately trained, with 36

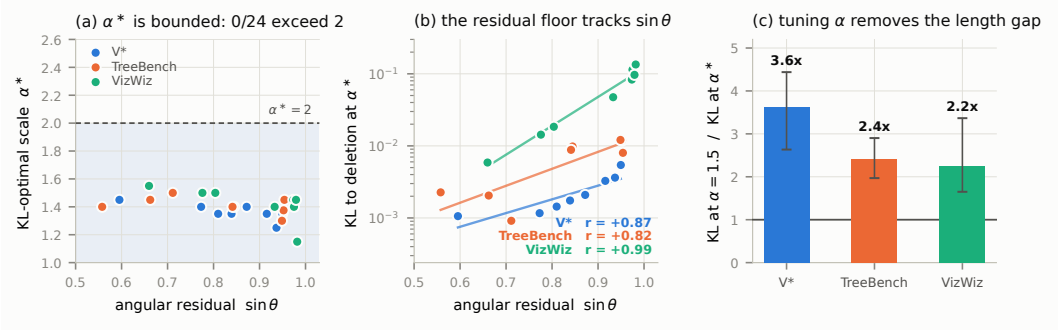

Figure 6: **The guidance scale absorbs the length of the negative direction, never its angle.** Each point is one construction on one benchmark, against its angle to deletion. (a) The scale that best matches deletion's guided distribution stays below 2 on all 24 pairs. (b) The divergence left at that scale rises with $\sin\theta$. (c) Tuning the scale reduces the divergence by 2.2 to 3.7 times.

layers of width 2048 against 28 of width 3584, so its difference from the 7B model mixes "weaker" with "different".

Table 15: **How far each negative stream is from the conditional**, at the first generated token, on full splits. **KL**: median KL divergence between the two next token distributions. **Agr.**: fraction of documents (%) on which their top tokens agree.

	V*		TreeBench		VizWiz	
Negative stream	**KL**	**Agr.**	**KL**	**Agr.**	**KL**	**Agr.**
Image deletion	0.59	48.2	0.29	50.6	10.38	1.1
Early exit, layer 10	8.71	0.0	8.77	0.0	7.58	0.0
Early exit, layer 15	6.79	0.0	6.37	0.0	6.79	0.0
Smaller, 3B	0.21	75.4	0.22	50.6	0.59	61.9

The usable band. The two failures are one rule. A negative is usable when its guidance vector is large enough to move a decision and aligned enough to move it the right way, which on MME-RealWorld means a median step 0 KL of 1 to 2 at 34% to 42% argmax agreement with the conditional. Image deletion sits at 1.50 and 34.2%, and every attention edit that works brackets it. Both failure modes sit outside the band: too small is inert, which is the 3B model and also the peaked attention edits, and too large is destructive, which is the elementwise map at $\gamma = -0.8$, at KL 18.75 and 0.1% agreement and 27.6 points below the baseline on VizWiz, because a useful negative disagrees selectively rather than everywhere. The band is a screen on perturbation design at a fixed scale, not a law, and it has two limits. It is blind to α by construction, since $\Delta = \ell_c - \ell_u$ does not depend on the scale: six image deletion arms at $\alpha \in [1.5, 8]$ share identical geometry while their accuracy spans 45.86 to 50.08. And its numeric range is benchmark specific, since on VizWiz the best arm sits at KL 12.4 and 1.1% agreement, which the rule as calibrated on MME-RealWorld would reject. Its one real test was out of sample: the band was fixed on MME-RealWorld before the early exit and 3B runs of Table 15 were made, and it predicted both failures.

Attention weakened negatives. For a text query row, a layer in $[1, 28)$ and a head, a map Φ_θ acts on the presoftmax scores over the image columns. Blur (Hong, 2024) smooths them over the patch grid and drives the row towards uniform while keeping the block mean; the elementwise map $(1+\gamma)s$ is a temperature that moves mean and variance together; the peaked map, our cross-attention analogue of the perturbation Ahn et al. (2024) apply to self-attention, moves mass onto each row's argmax patch. The peaked end fails by barely perturbing: its negative agrees with the conditional on 74.6% of documents, with median KL 0.14, and on V* its guidance vector is orthogonal to deletion's ($\cos = -0.018$). Blur is robust where the elementwise map is brittle: across a sixfold range of σ the V* predictions are identical on 191 of 191 documents, while moving γ from -0.35 to -0.8 costs 13.1 points on V* and 25.3 on VizWiz. Splitting the blur vector into its part along deletion and the rest, and guiding with each alone, the aligned part matches the full vector in all 32 cells we tested, and the rest is never better than deletion.

Scale. Every negative stream in Table 5 is compared at $\alpha = 1.5$ on every benchmark, which under-tunes image deletion on MME-RealWorld, whose ladder keeps rising to $\alpha \approx 4$ (45.86, 47.73, 49.71, 50.08, 50.08 and 50.03 at $\alpha = 1.5$ to 8). Table 16 therefore repeats the comparison with both arms at their own optimum; each scale is chosen on the test set, so both entries are upper bounds. The attention route then ties image deletion on V*, TreeBench and MME-RealWorld and loses on VizWiz. The 1.62 point advantage the blur arm shows at a shared $\alpha = 1.5$ on MME-RealWorld ($p = 0.001$) falls to 0.99 points ($p = 0.20$) once deletion is tuned; we extended deletion's ladder to $\alpha = 8$ in advance in order to catch exactly this.

Table 16: **Image deletion against attention weakening, each at its own guidance scale.** Accuracy (%), with the scale in parentheses; Δ is the attention route minus deletion, paired ($^{**}p < 0.01$). Parity on three benchmarks, a significant win for image deletion on the fourth. At a shared $\alpha = 1.5$ the MME-RealWorld comparison reverses, because deletion's optimum there lies at $\alpha \approx 4$ while the blur arm's lies at $\alpha = 3$. The rest of the paper uses $\alpha = 1.5$ on every benchmark.

Benchmark	Image deletion	Attention blur	Δ	p
V*	71.20 ($\alpha = 1.5$)	71.73 ($\alpha = 1.5$)	$+0.52$	n.s.
TreeBench	42.47 ($\alpha = 1.25$)	42.96 ($\alpha = 2.0$)	$+0.49$	n.s.
MME-RealWorld lite	50.08 ($\alpha = 4.0$)	51.07 ($\alpha = 3.0$)	$+0.99$	0.20
VizWiz	78.06 ($\alpha = 1.5$)	76.15 ($\alpha = 1.5$)	-1.91	**

Composition. The combiner guides with both negatives, $\tilde{\ell} = \ell_c + \alpha_0(\ell_c - \ell_{\text{drop}}) + \alpha_1(\ell_c - \ell_{\text{attn}})$; at $\alpha_1 = 0$ it reproduces deletion on 191 of 191 V* answers, which checks the wiring. Choosing per document between deletion and blur with hindsight would gain 1.86 points on average, positive on 8 of 8 benchmarks, and the gain is exact because greedy decoding is bitwise reproducible on these runs. No combiner we built recovers more than 0.15; exact McNemar tests over 37 benchmark and combiner pairs give 3 significant wins, 7 significant losses and 27 ties. The reason is redundancy: the blurred attention's guidance vector lies inside the cone of the pixel corruptions ($\cos = 0.64$ with deletion, against 0.63 for a pixel blur), and the part of it that deletion lacks never beats deletion on its own. Every combination is reproduced in score by a plain scale to within 0.3 points and agrees with one rung of the α ladder on 78% to 96% of predictions: a second negative adds a reparametrisation of α, not a new mechanism. We measure redundancy as the fraction of documents on which the two negatives' top tokens agree. Across the eight benchmarks it predicts the benefit of the second term (Spearman $+0.762$): the term pays where the negatives are interchangeable (MME-RealWorld 93%, BLINK 97%) and costs 5.14 points on VizWiz (4.4%) and 2.62 on RefChartQA (25%), where they disagree. The obvious alternative, averaging the two negatives' logits, is ordinary guidance against their midpoint, which cancels exactly the disagreement that composition is meant to exploit.

How close attention can get to deletion. Setting every image score to $-\infty$ closes the gap exactly on V* (0.00), TreeBench (-0.25) and VizWiz ($+0.07$), and leaves -2.34 points ($p < 0.001$) on MME-RealWorld, whose images carry 16,224 tokens. The residual is sequence structure that no attention edit can remove: the vision marker tokens stay in the prompt, the question sits about a thousand rotary positions later, and the key cache is longer.

Cost and portability. Attention weakened guidance takes 2.05 times the time of unguided decoding against 1.09 times for deletion, whose negative is cached after the first step. Editing the scores of every text row requires materialising the full attention matrix, which forces the eager kernel; eager against fused attention alone moves logits by up to 1.22 in this model. Editing only the last query row keeps the fused kernel but is inert: removing every image column from it changes 0 of 50 answers on V*, because the visual evidence was absorbed into the text tokens during prefill. The route also depends on the architecture exposing softmax scores at all. Hybrid models such as Qwen3.5 (Qwen Team, 2026) replace three of every four attention layers of the language model with linear attention, which forms no score over image tokens, so an attention space negative can reach only a quarter of the layers. A hidden state blend needs one hook on the final hidden state and transfers unchanged.

Two defects in our own sweeps. We parameterise σ in patch units rather than as a fraction of the image, so the same $\sigma = 16$ covers 15% of MME-RealWorld's short side and 46% of VizWiz's. The blur sweeps are therefore not comparable across benchmarks as parameterised, and VizWiz's optimum was never bracketed, its curve still rising at the largest σ we ran; every claim we make about blur is within a benchmark. Separately, the elementwise map corrupts the output format at

large α, emitting " C" or "(D)" where the task wants "B", so part of its measured collapse is a formatting failure rather than a wrong choice among the options.

Considered and not included. HeadHunter (Ahn et al., 2025) searches for the layer and head pairs to weaken. We ran it on MME-RealWorld, where it reaches parity (46.22 against 45.86), but it is a search over where to apply a perturbation rather than a new negative, and a head set chosen on the evaluation benchmark invites an overfitting objection. SAG (Hong et al., 2023) degrades input pixels and uses the model's attention only as a mask for where to degrade, so it belongs to the input corruption axis rather than to this one.

A.4 THE GEOMETRY OF THE UPDATE

The operators. Write $\Delta = \ell_c - \ell_u$ and $\gamma = \alpha - 1$. APG splits Δ into $\Delta_\parallel = s\,\ell_c$, with $s = \Delta^\top \ell_c / \|\ell_c\|^2$, and $\Delta_\perp = \Delta - \Delta_\parallel$, and rescales the first by η. CFG-Zero* replaces ℓ_u by $s^\star \ell_u$ with $s^\star = \ell_c^\top \ell_u / \|\ell_u\|^2$, the least squares fit of ℓ_c. TCFG takes the top eigenvector u_1 of $\ell_c \ell_c^\top + \ell_u \ell_u^\top$, which lies in the plane of the two streams, and projects ℓ_u onto it, $\hat{\ell}_u = (\ell_u^\top u_1)\, u_1 = p\,\ell_c + q\,\ell_u$. EP-CFG multiplies the guided logits by $\kappa = \|\ell_c\| / \|\ell_c + \gamma\Delta\|$. Collecting the coefficients of ℓ_c and ℓ_u gives the columns a and b of Table 2.

The span lemma. *Let an operator produce guided logits $\tilde{\ell} = a\,\ell_c + b\,\ell_u$ with scalars $a > 0$ and b, which may depend on the step. Under greedy decoding the emitted token depends on the operator only through $r = -b/a$, and vanilla guidance emits the same token at $\alpha_{\text{eff}} = 1/(1-r)$ for every $r < 1$.* Proof: since $a > 0$, $\arg\max_y \tilde{\ell}(y) = \arg\max_y \big[\ell_c(y) - r\,\ell_u(y)\big]$. Vanilla guidance has $a = \alpha$ and $b = -(\alpha - 1)$, so $r = (\alpha - 1)/\alpha$, which is increasing in $\alpha > 0$ and inverts to $\alpha = 1/(1-r)$ on $r < 1$. Every operator step we measured has $a > 0$ and $r < 1$. Under sampling the factor a does not cancel and acts as an inverse temperature. We apply the operators to the last hidden states; with a bias free head, $W(a\,h_c + b\,h_u) = a\,\ell_c + b\,\ell_u$, so the coefficients are the same in both spaces. A head with a bias β would add $(1 - a - b)\,\beta$, a third direction for every operator whose coefficients do not sum to one; vanilla guidance is unaffected.

TCFG halves the guidance excess. Let $\rho = \|\ell_u\| / \|\ell_c\|$. If $\rho = 1$, then $\ell_c \ell_c^\top + \ell_u \ell_u^\top$ is proportional to $\hat{c}\hat{c}^\top + \hat{u}\hat{u}^\top$, whose top eigenvector is the bisector of the two unit vectors. Projecting ℓ_u onto it gives $\hat{\ell}_u = \frac{1}{2}\ell_c + \frac{1}{2}\ell_u$, so $p = q = \frac{1}{2}$, $r = (\alpha - 1)/(\alpha + 1)$ and $\alpha_{\text{eff}} = (\alpha + 1)/2$. The angle between the streams cancels. In the limits, $\rho \to 0$ rotates u_1 onto ℓ_c and gives $\alpha_{\text{eff}} \to 1$, and $\rho \to \infty$ rotates it onto ℓ_u and gives $\alpha_{\text{eff}} \to \alpha$, so TCFG can only weaken guidance. On real states with $0.98 \le \rho \le 1.02$ we measure $p = 0.5000$, $q = 0.5018$ and $\alpha_{\text{eff}} = 1.2510$ against the predicted 1.25, with 96.1% of those steps within 0.02 of it (Table 17c). Real states sit mostly at $\rho \approx 1.05$ to 1.25, which is why TCFG runs at 1.25 to 1.46. On VizWiz its answers match those of $\alpha = 1.25$ on 96.6% of documents and those of $\alpha = 1.5$ on 93.6%.

CFG-Zero* guides the first token, then waits for the refresh. On the cached reference, $s^\star$ is 0.77 to 0.99 at step 0 and about zero from step 1 on (Table 17b): once the answer starts, the stale reference becomes orthogonal to the conditional state (on VizWiz exactly), so $s^\star \to 0$ and $\alpha_{\text{eff}} \to 1$. That orthogonality holds between the two streams, not between the guidance vector and the state it is added to, so it does not bear on where Δ points (Section 4.1.1). Guidance returns only when SWITCH refreshes the reference, at $t = K$. CFG-Zero* is therefore a duration rule, guiding the first token and then nothing until that refresh, and it inherits that rule's failure wherever the gain comes after the first token: on PhD with a rationale it loses 1.73 points, and on O3, whose gain lives in the stop decision, it scores 56.15 against 63.61 for SWITCH and 57.34 unguided. On a teacher forced reference $s^\star$ instead stays between 0.76 and 0.99 throughout, so the duration rule is a property of the prefix free reference and not of the operator alone.

The scale ladder and the tuned comparison. Table 18 gives the full ladder. No operator is significantly better than the best constant α on any benchmark. The nearest misses are TCFG on O3 ($+0.34$ over $\alpha = 1.25$, $p = 0.27$) and CFG-Zero* on VizWiz ($+0.48$ over $\alpha = 1.25$, $p = 0.23$). Where an operator is significant against the best rung, it is worse: TCFG loses 2.39 to $\alpha = 2$ on PhD with a rationale, and CFG-Zero* loses 9.12 to $\alpha = 1.25$ on O3. The best constant scale itself differs by benchmark (1.25 on O3, 1.5 on ReasonSeg, 2 on PhD with a rationale), so neither the operators nor any single scale wins everywhere.

Table 17: **The scale each operator runs at**, $\alpha_{\text{eff}} = 1/(1-r)$ at a nominal $\alpha = 1.5$, read from the stored decoding states of SWITCH ($K = 8$). **(a)** Medians for the operators whose scale barely moves; *early* is the first eight steps (cached reference), *late* the steps after the switch. **(b)** CFG-Zero* by step: full guidance at step 0, none after, until the reference is refreshed. **(c)** TCFG against the norm ratio $\rho = \|\ell_u\|/\|\ell_c\|$: at $\rho \approx 1$ it sits at the predicted $(\alpha+1)/2 = 1.25$.

(a) median α_{eff}, early / late

	VizWiz	O3	ReasonSeg	PhD-rat.
APG, $\eta = 0.5$	1.63 / 1.52	1.62	1.65 / 1.51	1.65 / 1.50
APG, $\eta = 2$	1.36 / 1.47	1.36	1.34 / 1.49	1.34 / 1.49
TCFG	1.44 / 1.28	1.46	1.51 / 1.28	1.39 / 1.25
EP-CFG	1.50 / 1.50	1.50	1.50 / 1.50	1.50 / 1.50

(b) CFG-Zero*: $s^\star$ and α_{eff} by step

step	VizWiz	O3	PhD-rat.
0	0.77 / 1.35	0.94 / 1.46	0.99 / 1.49
1	0.02 / 1.01	0.06 / 1.02	0.07 / 1.03
2 to 5	≈ 0 / 1.00	≈ 0 / 1.00	≈ 0 / 1.00
≥ 6, after the switch	–	–	0.95 / 1.46

(c) TCFG against the norm ratio (9,888 steps, four benchmarks)

ρ	0.90–0.98	**0.98–1.02**	1.02–1.10	1.10–1.25	1.25–1.60	> 1.60
steps	1,077	2,030	2,775	1,891	1,092	715
p / q	0.50 / 0.48	0.50 / 0.50	0.50 / 0.53	0.49 / 0.61	0.32 / 0.89	0.35 / 0.86
α_{eff}	1.235	**1.251**	1.269	1.319	1.461	1.479

Table 18: **The scale ladder behind Table 4.** Same five benchmarks and setting, with SWITCH ($K = 8$) swept over α and every imported operator at $\alpha = 1.5$. $\alpha = 1.4$ was run only on the two benchmarks where it was needed to settle a comparison. *Overall* is the unweighted mean of the five.

	VizWiz	O3	PhD-rat.	ReasonSeg	OCRBench	Overall
unguided, α=1	73.80	57.34	68.73	49.45	53.69	60.60
SWITCH, *scale ladder*						
$\alpha = 1.1$	76.61	61.64	69.59	50.69	53.54	62.41
$\alpha = 1.25$	78.22	65.27	70.19	51.21	53.91	63.76
$\alpha = 1.4$	–	64.78	72.44	–	–	–
$\alpha = 1.5$	78.06	63.61	72.78	51.86	53.77	64.02
$\alpha = 2$	75.31	56.52	73.57	51.23	50.02	61.33
imported operators, $\alpha = 1.5$						
APG, $\eta = 0.5$	77.59	63.15	72.18	51.65	54.11	63.74
APG, $\eta = 2$	78.43	64.13	72.84	51.67	53.82	64.18
CFG-Zero*	78.70	56.15	71.05	51.95	54.01	62.37
TCFG	78.11	65.61	71.18	51.35	53.87	64.02
EP-CFG	78.01	63.68	72.18	51.84	54.30	64.00

Exactness checks. On the stored trajectories of the five benchmarks, the reduced form $\ell_c - r\,\ell_u$ picks each operator's own token on 86,392 of 86,392 steps, over 7,414 documents. The code path with the operator switched off reproduces the SWITCH run on 1,131 of 1,131 VizWiz answers. EP-CFG is a positive rescale and so inert in exact arithmetic; in bf16 it changes 0.8% of VizWiz answers and 24% of PhD rationales, and it moves OCRBench v2 by $+0.53^*$ and PhD with a rationale by -0.60^*. Since an operator that cannot change a decision produces two significant results of opposite sign, differences below about 0.6 points in Table 4 are at the noise floor. APG's implied scale is also nearly constant across documents (a 10th to 90th percentile spread of 0.006 to 0.04 for $\eta = 2$), and the geometric quantity that sets it does not separate documents where guidance helped from those where it hurt (AUC 0.42 to 0.58), so it is not a per sample adaptive scale either.

A two-question screen for any guidance operator. The lemma and our failed attempts combine into a test that says, before any GPU time, whether an operator can change a decision at all. *Question 1: does the operator leave the span of the two streams?* If it does not, it is the guidance scale, and no experiment can show otherwise. This settles APG, CFG-Zero*, TCFG, EP-CFG and every composition of them. *Question 2: if it does leave the span, where in the 152k-dimensional logit space does its update go?* Three destinations are useless or harmful, and each is measurable in advance. (i) *The uniform shift.* Softmax ignores a constant added to every logit, and 60% to 97% of the guidance update is exactly that. It looks removable, but it carries the stop decision: removing it drops VizWiz from 77.77 to 19.61, because answers stop ending (3.4 to 24.3 steps) while 88.9% keep the same first word. (ii) *The tail.* Any weighting by p or $1/p$ moves the update onto tokens the model gives no mass, in a model that keeps 96.8% to 99.9% of its mass on five tokens. Moving

APG's split into the Fisher metric of the simplex does exactly this and loses 19 to 34 points at every η. (iii) *A metric that cannot be estimated.* Whitening the space needs a 3584×3584 covariance, whose measured rank is 380 to 599 at 300 to 600 samples; whitened APG has cosine 0.99999 with raw APG. The screen has a record. Question 2 rejected natural gradient guidance before any run, since all of its update landed on tokens with $p < 10^{-5}$; an earlier version that asked only whether the new update differs from the old one passed Fisher-metric APG, which then lost 34 points. The attention edit of PAI answers yes to Question 1, but Appendix A.6 shows that its new direction adds nothing once the scale is tuned.

A.5 Full benchmark results

Table 19: **Full benchmark results.** Accuracy (%) on ten benchmarks for every model, each with the four contrastive decoding baselines and ours. *Overall* is their unweighted mean and *Cost* is decoding time relative to unguided decoding. The shaded row of each block is the unguided model; best per column within a block in bold.

	OK-VQA	VizWiz	O3	O3-box	PhD	MMStar	TreeBench	BLINK IQ	V*	MME-RW	Overall	Cost
Qwen2.5-3B	40.84	67.75	62.88	62.91	73.09	56.47	40.25	23.33	**71.73**	38.04	53.73	1.00×
VCD (Leng et al., 2024)	43.48	66.43	59.25	60.26	74.43	55.26	36.30	24.00	68.59	39.45	52.75	1.96×
M3ID (Favero et al., 2024)	41.28	67.76	61.24	60.51	73.63	56.79	**40.99**	8.00	**71.73**	36.74	51.87	1.08×
PAI (Liu et al., 2024)	46.48	67.10	**64.30**	65.34	74.84	55.91	40.49	24.67	71.20	37.88	54.82	2.05×
SID (Huo et al., 2025)	47.92	70.07	62.49	**65.76**	76.14	**56.96**	39.26	24.00	71.20	42.63	55.64	2.00×
VLM-CFG (ours)	**51.96**	**72.34**	63.10	65.07	**77.12**	56.76	38.27	**30.67**	**71.73**	**44.50**	**57.15**	1.04×
Qwen2.5-7B	42.36	73.49	57.39	60.24	67.50	**62.43**	40.74	25.33	70.16	39.24	53.89	1.00×
VCD (Leng et al., 2024)	47.24	74.37	55.52	59.80	70.90	61.73	41.98	26.67	70.16	42.42	55.08	1.96×
M3ID (Favero et al., 2024)	42.16	75.06	55.61	57.73	67.51	58.01	37.04	25.33	70.68	46.17	53.53	1.08×
PAI (Liu et al., 2024)	50.68	75.81	**65.28**	**67.29**	70.49	62.26	41.48	28.00	70.16	41.84	57.33	2.05×
SID (Huo et al., 2025)	48.24	73.49	56.74	61.74	73.04	61.19	41.73	26.67	72.25	45.49	56.06	2.00×
VLM-CFG (ours)	**54.00**	**78.01**	63.94	65.86	**79.15**	62.17	**42.96**	**29.33**	**73.30**	**47.11**	**59.58**	1.04×
Qwen3-4B	48.32	**80.05**	51.04	52.65	74.47	**63.12**	43.21	28.67	72.25	45.34	55.91	1.00×
VCD (Leng et al., 2024)	47.84	76.82	51.48	53.90	74.38	57.96	**45.93**	31.33	75.39	46.59	56.16	1.96×
M3ID (Favero et al., 2024)	48.68	**80.05**	51.04	52.65	74.47	57.44	43.21	28.67	72.77	45.34	55.43	1.08×
PAI (Liu et al., 2024)	48.60	76.79	51.06	54.64	76.11	57.45	43.95	**32.00**	72.77	47.00	56.04	2.05×
SID (Huo et al., 2025)	48.92	78.19	**53.17**	**57.00**	76.16	57.12	43.70	30.00	**76.96**	**47.16**	56.84	2.00×
VLM-CFG (ours)	**49.88**	79.50	52.22	56.56	**77.12**	56.91	44.69	30.67	74.35	46.74	**56.86**	1.04×
Qwen3-8B	**52.96**	81.06	59.62	70.61	74.52	60.16	44.69	28.00	79.06	46.12	59.68	1.00×
VCD (Leng et al., 2024)	51.76	79.28	60.28	70.76	74.46	59.50	44.69	28.67	**81.15**	48.51	59.91	1.96×
M3ID (Favero et al., 2024)	52.68	81.09	59.63	70.49	74.65	60.20	45.68	26.67	78.53	46.22	59.58	1.08×
PAI (Liu et al., 2024)	51.72	79.71	62.09	71.08	76.22	**61.97**	44.44	30.00	79.58	47.37	60.42	2.05×
SID (Huo et al., 2025)	51.60	80.95	62.30	**72.64**	77.55	60.77	**46.42**	28.00	76.96	47.99	60.52	2.00×
VLM-CFG (ours)	52.36	**81.13**	**63.80**	72.43	**79.01**	61.16	45.19	**30.67**	79.58	**48.62**	**61.40**	1.04×
InternVL3.5-8B	40.68	62.46	60.42	63.15	72.34	66.05	41.98	**29.33**	73.30	50.55	56.03	1.00×
VCD (Leng et al., 2024)	40.36	62.25	58.76	61.53	73.61	65.22	42.96	**29.33**	72.25	49.92	55.62	1.96×
M3ID (Favero et al., 2024)	40.20	**62.55**	60.10	62.53	72.34	66.05	41.98	**29.33**	73.30	50.55	55.89	1.08×
PAI (Liu et al., 2024)	41.60	62.27	**62.35**	**65.08**	73.52	66.22	41.23	**29.33**	73.30	50.55	56.54	2.05×
SID (Huo et al., 2025)	43.24	62.39	58.67	61.81	75.55	66.33	**43.21**	**29.33**	**73.82**	**52.06**	**56.64**	2.00×
VLM-CFG (ours)	**45.40**	62.45	58.36	61.37	**76.68**	**67.05**	41.73	**29.33**	68.59	50.03	56.10	1.04×

Guidance against reasoning. PhD-CCS asks questions whose answers contradict common sense, so a model that answers from its language prior rather than from the image fails by construction. The obvious remedy is to let it think longer. Table 20 runs that comparison: the same models prompted to reason before answering, against the same models guided. Reasoning moves accuracy by at most 0.72 points and costs 0.24 on InternVL3.5-8B, while guidance adds 2.65 to 11.65. The two are not alternatives, since guidance acts on the logits and reasoning on the prompt, but on a benchmark built to punish the language prior, subtracting that prior directly is what helps.

A.6 Attention intervention on the conditional stream

The span lemma of Section 3.4 leaves one way to be more than a scale: add a direction that is not built from the two streams. A common way to do so in vision-language models is to strengthen the conditional stream's attention to the image and then apply guidance, as PAI (Liu et al., 2024) does. Its guided output is $\mathrm{CFG}(\alpha) + \alpha\,\delta$, where δ is the change the attention edit makes to ℓ_c, and δ lies almost entirely outside the span: it is 98% to 100% orthogonal to Δ on six benchmarks. The new direction does not pay. Against vanilla guidance tuned on the same benchmark, PAI is ahead by 0.21 points on VizWiz (not significant), behind by 4.79 points on MME-RealWorld lite, and on OpenImages it scores below unguided decoding (34.24 against 36.80). Adding three attention edits

Table 20: **Guidance against reasoning on a hallucination benchmark.** Accuracy (%) on PhD-CCS, the counter-common-sense split of PhD, for the unguided model, the same model prompted to reason before it answers, and VLM-CFG.

	Qwen2.5-3B	Qwen2.5-7B	Qwen3-4B	InternVL-3.5-8B
no guidance	73.09	67.50	74.47	72.34
reasoning	73.81	68.20	74.73	72.10
VLM-CFG	**77.12**	**79.15**	**77.12**	**76.68**

of our own to guidance gives the same picture: across 12 edit and benchmark pairs at matched decoding, none beats tuned guidance significantly and two lose significantly (Table 21). Strengthening attention in the conditional stream adds cost but no ability that the guidance scale does not already provide.

Table 21: **An attention edit on the conditional stream, against the guidance scale** (VizWiz, $n = 1131$, paired; $^*p < 0.05$). The edit is worth $+1.15$ points when guidance is off, and the gain is absorbed once α is tuned.

α	1.0 (off)	1.1	**1.2**	1.3	1.5	2.0
guidance alone	74.31	76.95	**78.53**	78.47	77.98	78.12
+ attention edit	75.28	77.57	78.95	78.66	77.89	78.58

The edits we built are $A_i \leftarrow f_\gamma(A_i)$ on the presoftmax scores of the image columns $i \in I$, with $f_\gamma(a) \in \{a(1+\gamma),\ a+\gamma,\ a+\gamma|a|\}$ (`mul`, `add`, `abs`); `abs` is the operator of PAI (Liu et al., 2024). For any edit producing z_p in place of z_c, the guided output is $z_u + \alpha\Delta + \alpha\delta$ with $\delta = z_p - z_c = k\Delta + \delta_\perp$, so the part along Δ is guidance at scale $\alpha(1+k)$. A single tuned α explains between -38% and $+1\%$ of the edit's effect on the output distribution, a per step oracle α at most 58%, and the orthogonal part 66% to 93%. The edit is orthogonal to guidance and dominated by it.

Two of the three operators are not well posed. Softmax is invariant to a constant added to a score row, so an operator that reads a quantity off the row must commute with that shift. Only `add` does. Shifting every edited row by $c \in [-40, 40]$ leaves the unedited model unchanged to 6×10^{-7} and changes none of 300 answers under `add`, but changes 6% to 18% of answers under `mul` and 20% to 55% under `abs`, with accuracy swings of up to 20.5 points.

A.7 Region knowledge: an upper bound

Every result in the main text uses no ground truth. Here we ask what guidance could gain if it knew *where* the answer is. With R the image tokens whose patch falls inside the annotated box or mask, the region can be used in the negative stream (black out or blur R in the negative image), or in the conditional stream (scale attention to R and to the rest of the image separately). Because every variant reads the dataset's own annotation, these are oracle variants: an upper bound on what region knowledge buys, not a deployable method.

The region signal is real. Boosting attention to the annotated region gains 2 to 5 points without guidance, while boosting a random region of the same size costs 3.6 to 12.0 points and boosting the background costs almost as much. The effect is grounding, not generic perturbation.

It does not survive prediction. Replacing the annotation with a region predicted by the model itself, with the same operator and strength, removes essentially the whole gain on VizWiz, TreeBench and V* (Table 22b). OpenImages is the exception: its objects are large and canonical, a predicted box matches the annotation, and most of the gain survives. Predicted regions work where the region is easy to predict, which is where grounding was least needed. Combining a predicted region with guidance, and localizing the negative stream with a predicted region, remain open.

Table 22: **Region knowledge as an oracle.** Accuracy (%), no guidance unless stated. **(a)** Region variants with the annotated region, with a random region and on the background. **(b)** The same attention operator with the annotated region or a region predicted by the model. Paired differences; $^*p < 0.05$.

(a) annotated regions

variant	VizWiz	TreeBench	V*	OpenImages
no guidance	73.80	40.49	70.16	36.33
region attention, `mul` $\gamma = 0.1$	76.30	42.72	73.82	38.00
random region, $\gamma = 0.5$	69.78	37.78	58.12	33.61
boost background	70.75	38.02	63.35	33.40
localized negative (region blacked out)	78.36	44.69	72.25	38.62
region attention + vanilla guidance	79.50	44.44	72.77	36.33
region attention + localized negative	77.31	45.93	73.82	40.50

(b) annotated against predicted region, gain over no guidance

	VizWiz	TreeBench	V*	OpenImages
`mul` $\gamma = 0.2$, annotated	$+2.04^*$	$+2.47$	$+3.66^*$	$+3.34^*$
`mul` $\gamma = 0.2$, predicted	$+0.04$	$+0.25$	-1.57	$+2.92^*$
`add` $\gamma = 3$, annotated	$+2.22^*$	$+1.98$	$+5.24^*$	$+5.64^*$
`add` $\gamma = 3$, predicted	-0.35	$+0.49$	-1.05	$+4.18^*$

B THE GUIDANCE FRAMEWORK IN DETAIL

B.1 WHAT DIFFUSION ASSUMES AND AUTOREGRESSIVE DECODING BREAKS.

The equation transfers; the machinery built around it rests on three assumptions that autoregressive decoding breaks. **(1) There is no trained null.** Diffusion drops the condition during training, so its unconditional prediction is trained; a vision-language model never does, though text-only pretraining makes the imageless prompt familiar, so b_q never vanishes. **(2) Every step is read once.** A diffusion sampler corrects small errors over many steps; an autoregressive model reads its guided vector once, through a linear head, into an argmax, and an early stop or a wrong first word cannot be undone. **(3) The state is the model's own transcript.** Diffusion evaluates both branches on the same noisy sample; here, the state is the prefix $y_{<t}$, a record of what the conditional has already decided about the image, so the reference must stand on *some* prefix and Equation 1 does not say which. The first two predict that most of the diffusion toolkit collapses to the guidance scale (Sections 3.2 and 3.4, Appendix A.1); the third opens an axis diffusion never had (Section 3.3).

B.2 ONE OPERATOR ON TWO FAMILIES

Let the model define, at each step, a transition kernel with natural parameter $\theta(s, c)$ in some exponential family, for state s and condition c. Guidance is the geometric tilt

$$\tilde{p}_\alpha(\cdot \mid s, c) \;\propto\; p(\cdot \mid s, c)^\alpha \, p(\cdot \mid s, \emptyset)^{1-\alpha} \;=\; p(\cdot \mid s, c) \, \exp\big((\alpha - 1)\,[\log p(\cdot \mid s, c) - \log p(\cdot \mid s, \emptyset)]\big). \tag{5}$$

Gaussian kernels. A diffusion reverse step is $\mathcal{N}(\mu(z, c), \sigma^2 I)$ with a variance that does not depend on c. Then

$$p_c^\alpha \, p_u^{1-\alpha} \;\propto\; \exp\Big(-\tfrac{1}{2\sigma^2}\big[\alpha\|z - \mu_c\|^2 + (1-\alpha)\|z - \mu_u\|^2\big]\Big) \;\propto\; \mathcal{N}\big(\alpha\mu_c + (1-\alpha)\mu_u,\; \sigma^2 I\big).$$

Because μ is affine in the noise prediction ϵ with a coefficient that does not depend on c, this is $\tilde{\epsilon} = \epsilon_u + \alpha(\epsilon_c - \epsilon_u)$, the update of Ho & Salimans (2022).

Categorical kernels. With logits as the natural parameter, the tilt is $\tilde{\ell} = \ell_u + \alpha(\ell_c - \ell_u)$ followed by a softmax, which is Equation 1. The normaliser $Z = \sum_y p_c(y)^\alpha p_u(y)^{1-\alpha}$ is a finite sum, so $\tilde{p}$ is always a proper distribution. Ho & Salimans (2022) observe that in the Gaussian case the guided score is built from nonconservative fields and need not be the score of any distribution; in the categorical case that caveat does not arise.

B.3 THE BIAS DECOMPOSITION

For any reference q and the true marginal $\bar{p}(y \mid h) = \mathbb{E}_{v\sim p(v|h)}\, p(y \mid v, h)$, Bayes' rule gives $\log p(y \mid v, h) - \log \bar{p}(y \mid h) = \log p(v \mid y, h) - \log p(v \mid h) = \mathrm{pmi}(y; v \mid h)$. Adding and subtracting $\log \bar{p}$ gives Equation 3. The decomposition is exact for every q, and it assigns each method in the literature to one reference:

method	reference $q(y)$
VCD (Leng et al., 2024)	$p(y \mid h,$ noised $v)$
ICD (Wang et al., 2024d)	$p(y \mid h', v)$ with a distorted instruction
M3ID (Favero et al., 2024), VDD (Zhang et al., 2025b)	$p(y \mid h, \emptyset)$
autoguidance (Karras et al., 2024)	a weaker model, $p_{\text{weak}}(y \mid h, v)$
PAG, SEG (Ahn et al., 2024; Hong, 2024)	the same model with degraded attention to v
this paper	$p(y \mid \pi_t, \emptyset)$, with the prefix π_t chosen by the rule of Section 3.3

Why no budget theorem holds. Under teacher forcing it is tempting to sum the bracket over the answer and invoke the chain rule of mutual information, $\sum_t \mathrm{pmi}(y_t; v \mid y_{<t}) = \mathrm{pmi}(y_{1:T}; v)$, which would cap the total guidance an answer can receive. We state this only to rule it out. The identity needs the reference to be the marginal $\bar{p}$, and deleting the image from the prompt is not integrating over images. Measured, the cumulative teacher forced term on OCRBench v2 grows almost linearly, from 4.7 to 339 nats over 255 steps, with no saturation.

B.4 WHAT ITERATED GUIDANCE TARGETS AT THE SEQUENCE LEVEL

Per step tilting is not sequence level tilting. Under teacher forcing the image free factors do telescope, $\prod_t p(y_t \mid y_{<t}, \emptyset) = p(y_{1:T} \mid \emptyset)$, because this is the chain rule of the image free model itself and needs no marginal. The per step normaliser is a Rényi divergence, $Z_t = \exp\big((\alpha - 1)\, D_\alpha(p_c \,\|\, p_u)\big)$, so the distribution that iterated guidance samples is

$$\tilde{p}(y_{1:T} \mid v) \;\propto\; p(y_{1:T} \mid v)^{\alpha}\, p(y_{1:T} \mid \emptyset)^{1-\alpha}\; \exp\Big(-(\alpha - 1) \sum_t D_\alpha\big(p_c \,\|\, p_u\big)(y_{<t})\Big). \tag{6}$$

Iterated guidance is the sequence level tilt, penalised by the Rényi divergence accumulated along the path, so it is biased against paths that pass through states where the image matters most. This is the autoregressive counterpart of the observation of Bradley & Nakkiran (2024) that guidance does not sample the tilted distribution, here in closed form. It holds only for teacher forcing; for the prefix free reference the image free factors do not telescope and no sequence level target exists.

B.5 THE PLAUSIBILITY CONSTRAINT AS A TRUST REGION

The tilted distribution is the unique solution of $\tilde{p} = \arg\max_q \{\mathbb{E}_q[\ell_c - \ell_u] - (\alpha - 1)^{-1}\, \mathrm{KL}(q \,\|\, p_c)\}$, so guidance maximises the bracket inside a KL trust region around the conditional, with $\alpha - 1$ as an inverse temperature. The adaptive plausibility constraint of contrastive decoding (Li et al., 2023a; Leng et al., 2024) restricts the choice to tokens with $p_c(y) \geq \beta \max p_c$: the same trust region, implemented as a hard support constraint instead of a penalty. Appendix A.2 shows what that hard version does to the answers.

B.6 THE DESIGN SPACE

Table 23 arranges the choices a guided vision-language decoder has to make into four axes. The first three are where the diffusion literature has spent its effort, and the first two differences of Section B.1 predict that most of that effort does not carry over (Section 3). The fourth has no diffusion counterpart; the third difference predicts that it matters, and that it is invisible wherever the answer is a single token (Section 3.3).

B.7 EXPERIMENTAL PROTOCOL

All results use Qwen2.5-VL-7B-Instruct (Bai et al., 2025b) in bf16 with greedy decoding. Unless stated, the negative stream deletes the image from the prompt, the blend acts on the final hidden

Table 23: **The design space of guided decoding.** Four axes, in the order of Section 3: what diffusion does on each, what autoregressive decoding changes (the differences of Section B.1), and what we therefore predict before measuring. Three of the axes are where the diffusion toolkit lives; where the reference stands exists only in an autoregressive model. The last column points to the experiment that tests the prediction.

axis	diffusion practice	what changes	prediction	tested
what the negative stream sees	dropped prompt, weaker model, perturbed attention	no trained null; read once	image deletion is the limit of every corruption; a weaker model is too close or too far, and weakening its access to the image reaches deletion but cannot pass it	§4.1.2
where the reference stands	not a choice: both branches share one state	the state is the answer so far	the choice matters, and is invisible on single token answers	§4.1.1
the operator on the two streams	APG, CFG-Zero*, TCFG, rescaling	the same	every such operator is plain guidance at another scale	§4.1.3
where the equation acts	noise or velocity prediction	one linear read into an argmax	the final hidden state equals the logits; earlier sites are not guidance	App. A.1

state (Appendix A.1), and $\alpha = 1.5$. Every method is compared with vanilla guidance at its own best scale, never at a shared default. Every guided variant is compared with the $\alpha = 1$ run of its own implementation, which differs from unguided decoding on about 15% of documents through rounding order alone. Between two guided variants, 75% to 96% of answers are identical, so unpaired standard errors overstate the uncertainty by a factor of two to three; we report paired bootstrap intervals (10k resamples) or exact McNemar tests, and * marks an interval that excludes zero. Run to run jitter of greedy decoding is about 0.35 points on VizWiz. The benchmarks span answer length, from a single token to several hundred, because the reference rules can only differ once there is an answer prefix, so they separate along answer length.

B.8 Benchmarks

benchmark	answer	metric	n
V* (Wu & Xie, 2024)	option letter	accuracy	191
TreeBench (Wang et al., 2026)	option letter	accuracy	405
MME-RealWorld lite (Zhang et al., 2025a)	option letter	accuracy	1919
MMStar (Chen et al., 2024a)	option letter	accuracy	1500
BLINK IQ (Fu et al., 2024)	option letter	accuracy	150
VizWiz grounding (Gurari et al., 2018; Chen et al., 2022)	short answer	VQA accuracy	1131
OK-VQA lite (Marino et al., 2019)	short answer	exact match	500
OpenImages (Kuznetsova et al., 2020)	short answer	accuracy	958
PhD (Liu et al., 2025)	yes or no, or rationale then answer	accuracy	1506
HallusionBench (Guan et al., 2024)	yes or no	accuracy	360
O3 (Dahou et al., 2025)	short list	F1	2001
O3-box (Dahou et al., 2025)	short list with boxes	F1	2001
RefChartQA (Vogel et al., 2025)	short answer	accuracy	6223
ReasonSeg (Lai et al., 2024)	box coordinates	IoU	776
OCRBench v2 (Fu et al., 2025)	free text	accuracy	2000
DetailCaps (Dong et al., 2024)	long description	CAPTURE	500
AMBER (Wang et al., 2023)	long description	Cover	1004
MathVerse (Zhang et al., 2024b)	chain of thought	accuracy	2130

C Extended related work

C.1 Classifier-free guidance in generation

A cheap substitute that became the standard. Guidance in diffusion models began with classifier guidance (Dhariwal & Nichol, 2021), which pushes the sampler along the gradient of a separate classifier trained on noisy images, trading diversity for fidelity. Ho & Salimans (2022) asked whether that classifier was needed at all. A single network trained with the condition randomly dropped

yields both a conditional prediction ϵ_c and an unconditional one $\epsilon_\emptyset$, and extrapolating from the second towards the first, $\epsilon_\emptyset + \alpha\,(\epsilon_c - \epsilon_\emptyset)$, reproduces the effect of the classifier gradient. The stated goal was modest: remove an auxiliary model. The consequences were not. GLIDE (Nichol et al., 2022) found that human raters preferred classifier-free guidance over CLIP guidance, and from then on it was the default. Latent Diffusion (Rombach et al., 2022) made guidance with a dropped text prompt the standard open recipe, and DALL·E 2 (Ramesh et al., 2022) and Imagen (Saharia et al., 2022) built on it, the latter adding dynamic thresholding to tame the saturation that large scales cause. The rule then survived every change of model family. It moved to autoregressive and masked token generators by mixing logits instead of scores (Gafni et al., 2022; Yu et al., 2022; Chang et al., 2023; Sun et al., 2024a; Tian et al., 2024), to rectified flow (Esser et al., 2024), to video (Polyak et al., 2024), and to audio (Liu et al., 2023; Copet et al., 2023). Scale did not retire it either. The largest current generators still run two passes per step (Esser et al., 2024; Polyak et al., 2024), or pay to distill the guided model into one network (Meng et al., 2023; Black Forest Labs et al., 2025). Unified models that both read and produce images make the point most sharply: they decode with guidance when they generate an image and without it when they answer a question about one (Chameleon Team, 2024; Wang et al., 2024c; Xie et al., 2025; Chen et al., 2025e).

What guidance does, and how to repair it. A second body of work asks what the extrapolation actually computes. It does not sample the sharpened distribution $p(x)\,p(c \mid x)^\alpha$ it is often said to target; it behaves like a predictor followed by a corrector (Bradley & Nakkiran, 2024), pushes samples towards the extremes of the class support (Chidambaram et al., 2024), and raises classification confidence at the price of diversity (Wu et al., 2024), although some of these distortions fade in high dimensions (Lehman Pavasovic et al., 2025). A third body repairs the failures that large scales produce, namely oversaturation, lost diversity, and samples that leave the data manifold. It rescales the guided prediction (Lin et al., 2024), anneals the condition (Sadat et al., 2024), applies guidance only in a middle interval of noise levels (Kynkäänniemi et al., 2024), or schedules its strength over time (Wang et al., 2024b). Others change the geometry of the update itself: a correction derived from the sampling dynamics (Zheng & Lan, 2024), a manifold constraint (Chung et al., 2025), a projection that shrinks the component parallel to the prediction (Sadat et al., 2025a), or an optimised scale for flow matching (Fan et al., 2025). Nearly all of these corrections assume a continuous sample space in which a guided prediction can leave a data manifold and be brought back. Whether they carry any meaning when the output is a distribution over a discrete vocabulary, read once per token, has to our knowledge not been tested.

What the negative branch should see. The most productive insight of this literature is that the subtracted term does not have to be unconditional. Self-attention guidance (Hong et al., 2023) and perturbed-attention guidance (Ahn et al., 2024) contrast the model with a copy of itself whose attention is degraded, and later work smooths the attention energy (Hong, 2024), skips layers (Hyung et al., 2025), or selects which heads to perturb (Ahn et al., 2025). Autoguidance (Karras et al., 2024) contrasts against a smaller or undertrained copy and argues that much of the benefit of classifier-free guidance comes from the quality gap between the two predictions rather than from the condition. The negative branch can also be an independent random condition (Sadat et al., 2025b) or a negative prompt (Ban et al., 2024). Guidance terms can be summed and negated to compose or remove concepts (Liu et al., 2022; Brack et al., 2023; Schramowski et al., 2023; Gandikota et al., 2023), split into separate scales for an image and a text condition (Brooks et al., 2023), adapted for inversion (Mokady et al., 2023), and turned into a training signal for 3D generation, where the guidance term alone turns out to drive score distillation (Poole et al., 2023; Yu et al., 2024). Even the work that removes the cost of guidance keeps its effect, by folding the guided distribution into training (Chen et al., 2025c;b; Tang et al., 2025). Taken together, the generation community treats classifier-free guidance as a design space with four axes: what the negative branch sees, how the two predictions are combined, how the strength evolves along the trajectory, and where in the network the equation acts (Chen et al., 2025a; Nguyen et al., 2025; Guo & Du, 2026). **Visual understanding has no counterpart. No widely used vision-language model decodes with guidance on its image, although answering a question about an image is conditional autoregressive generation in exactly the sense in which a generator over image tokens draws an image from a caption.**

C.2 Contrastive decoding

The same equation, under other names. Language modelling reached the contrast of two predictions on its own. Contrastive decoding (Li et al., 2023a) subtracts the log probabilities of a small

amateur model from those of a large expert, and restricts the choice to tokens the expert already finds plausible. DoLa (Chuang et al., 2024) contrasts a late layer of a model with an early one, context-aware decoding (Shi et al., 2024) contrasts a prediction made with a document against one made without it, and Sanchez et al. (2024) apply classifier-free guidance to the prompt of a language model and report gains equivalent to doubling its parameter count. In the terms of Section C.1 these are one method with three choices of negative branch: a weaker model, which is autoguidance before autoguidance; a weakened copy of the same model, as in perturbation guidance; and a dropped condition, which is classifier-free guidance itself.

Visual contrastive decoding. VCD (Leng et al., 2024) brought the contrast to vision-language models. It contrasts the prediction on the clean image with one on a copy corrupted by Gaussian noise, keeps the plausibility constraint of Li et al. (2023a), and targets object hallucination. Its authors note in passing that the objective echoes classifier-free guidance in image generation. The line that followed took the equation and left the design space behind. Dozens of methods differ mainly in what the contrast stream sees: a distorted instruction (Wang et al., 2024d), no image at all (Favero et al., 2024; Zhang et al., 2025b), a transformed or shuffled image (Woo et al., 2024; Xia et al., 2026), a pruned or amplified set of visual tokens (Huo et al., 2025; Zhu et al., 2025; Yin et al., 2025a), a different layer (Tong et al., 2025), or a different attention pattern (An et al., 2025; Chen et al., 2024b; Gong et al., 2024; Kan et al., 2024; Cao et al., 2025; Jo et al., 2026). Others intervene inside the network, penalising excessive attention to summary tokens during beam search (Huang et al., 2024) or amplifying attention to image tokens (Liu et al., 2024). The newest entrants have started to call themselves guidance (Cao et al., 2025; Jo et al., 2026), and a few works borrow classifier-free guidance by name for a specific control task (Zhang et al., 2024c). Seen through Section C.1, each of these methods fixes one point in the guidance design space: a negative branch chosen by hand, a single constant strength, and a rule, never stated, for what the negative branch has seen of the answer so far. In every method we examined that rule is the same: the negative stream is fed the tokens chosen by the guided one.

The line deserves credit for making hallucination a measurable target and for producing interventions that need no training. Its reach, however, is narrow. The methods are evaluated almost entirely on hallucination benchmarks such as POPE (Li et al., 2023b), CHAIR (Rohrbach et al., 2018), the hallucination subset of MME, and MMHal (Sun et al., 2024b), and rarely on general question answering, reasoning, reading, or grounding. Where general benchmarks have been checked, the effect of VCD is small and changes sign from one model to the next: on LLaVA-1.5 it improves POPE while lowering MathVista and HallusionBench (Chen et al., 2025d). None of the recent open model releases we know of, including Qwen2.5-VL (Bai et al., 2025b), Qwen3-VL (Bai et al., 2025a) and InternVL3.5 (Wang et al., 2025), decodes with contrastive decoding or even mentions it, in contrast to image generation, where guidance ships with every release. Reproduction studies now question even the hallucination gains. Yin et al. (2025b) trace them to two side effects, a crude shift of the output distribution in a single direction and a plausibility constraint that turns sampling into greedy search, and show that baselines built only from these side effects match the reported gains. Bendre et al. (2026) reproduce these findings on LLaVA-1.5 and Qwen2.5-VL and conclude that the gains do not translate into stronger visual grounding. **Contrastive decoding has stayed a hallucination patch: narrowly evaluated, absent from the models people deploy, and now doubted by the studies that reproduce it.**

Where this paper stands. We read these critiques as evidence about one point in the design space, not about guidance on the image as such. The implementations they test fix every axis of Section C.1 at a single value, and their evaluation looks only where a gain is hardest to separate from a bias towards one answer. Unlike the contrastive decoding line, which builds each contrast stream as a remedy for hallucination, we treat that stream as the unconditional branch of classifier-free guidance and bring the full design space of Section C.1 into vision-language decoding. In contrast to diffusion, where both branches follow the same trajectory, autoregressive decoding raises a question with no diffusion analogue: which past the unconditional branch should be conditioned on. The nearest work keeps separate caches for the two branches in visual autoregressive generation (Xu et al., 2026), rebuilds the unconditional input at every step for masked diffusion language models (Li et al., 2025), or builds the negative input for language reasoning by hiding the tokens the model is already sure of (Tang et al., 2026). None studies it for vision-language models with the image dropped, and M3ID, which also drops the image, leaves it unexamined. Different from the reproduction studies, which ask whether contrastive decoding reduces hallucination, we ask whether guidance on the

image improves the answers of a vision-language model in general, and we evaluate it on general benchmarks rather than on hallucination suites alone.