HARMONY: ADAPTIVE SELF-GUIDANCE THROUGH HIERARCHICAL SUBNETWORK PROBING

Anonymous authors
Paper under double-blind review

ABSTRACT

Guidance improves diffusion generation, but its quality and cost depend on the auxiliary prediction used to construct the correction. We introduce HARMONY, which obtains complementary guidance signals from a diffusion transformer's own hierarchy and adapts their influence across time and space. An early readout of the residual stream supplies a weak prediction without repeating transformer blocks, while an optional block-dropped pass provides a second direction. Agreement between these probes controls when and where guidance is applied. A conditional error model relates population agreement to optimal shrinkage and explains the effects of unequal probe quality and correlated errors. Across the reported ImageNet, text-to-image, and video evaluations, HARMONY improves generation quality while its Lite variant requires only two additional readout-head evaluations. On DiT-XL/2, the full method reduces FID from 2.27 to 1.93 with increased recall. The framework connects the choice of a weak prediction with adaptive guidance strength through information already available during inference.

1 INTRODUCTION

Guidance turns a denoising prediction into a decision about what a generated image or video should emphasize. In modern diffusion and flow models, that decision shapes semantic alignment, detail, diversity, and the compute required at every sampling step (Dhariwal & Nichol, 2021; Ho & Salimans, 2022; Rombach et al., 2022; Esser et al., 2024; Wan et al., 2025). The familiar conditional–unconditional contrast is effective, yet increasing its strength can introduce saturation and artifacts or reduce diversity (Sadat et al., 2024; 2025a; Chung et al., 2025). Understanding when a correction is reliable is therefore as important as constructing a large correction.

A useful perspective is that guidance contrasts a model with a weaker prediction of the same underlying target. AutoGuidance demonstrated the value of a compatible weaker model (Karras et al., 2024); training-free alternatives create such contrasts through input blurring, attention perturbation, token perturbation, layer skipping, or stochastic block dropping (Hong et al., 2023; Ahn et al., 2024; Hong, 2024; Ifriqi et al., 2025; Rajabi et al., 2025; Stability AI, 2024; Chen et al., 2026; Gu & Hou, 2025). These mechanisms expose different errors and incur different costs. Guidance schedules, limited intervals, spatial adjustments, and feedback controllers further show that the useful correction varies along a trajectory and across the image (Wang et al., 2024; Kynkäänniemi et al., 2024; Shen et al., 2024; Koulischer et al., 2025). The unresolved design question is how to connect a cheap source of contrast with a sample-dependent assessment of its reliability.

A diffusion transformer offers such a source within its residual stream. Intermediate states have already been computed on the way to the final prediction, and applying the shared output head before the last few residual updates gives an early prediction without another block-stack evaluation (Peebles & Xie, 2023; Ma et al., 2024a; Esser et al., 2024). Its difference from the final prediction exposes the contribution of late refinement. This use of an intermediate representation differs from early-exit acceleration and feature caching: the full prediction is retained, and the early readout becomes a guidance probe rather than a replacement for it (Moon et al., 2024; Tang et al., 2023; Ma et al., 2024c;b; Chen et al., 2024).

HARMONY combines this depth probe with an optional block-dropping probe and uses their agreement to modulate the correction. Under an error-amplification model, a population correlation equals the geometric mean of two optimal shrinkage factors. This relation motivates agreement-based

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.

guidance while making unequal signal-to-noise ratios and shared errors explicit. A global gate controls the step, a token gate allocates its influence spatially, and an orthogonal projection suppresses the radial component of the update. The resulting hierarchy offers a Lite mode using only extra head evaluations and a fuller mode that spends an additional pass on a complementary direction.

Evaluation covers class-conditional ImageNet with DiT and SiT, text-to-image generation with SD3, SD3.5 and FLUX.1-dev, and video generation with Wan2.1. The reported comparisons jointly examine sample quality, diversity, prompt alignment, and compute. Ablations isolate the probe directions and gating components, while the architecture study examines how the depth contribution varies across transformer families.

Our contributions are:

- A hierarchical self-guidance framework coupling two probe directions with adaptive time and token weights.
- An early residual-stream readout that supplies a guidance direction without repeating transformer blocks.
- A shrinkage analysis of probe agreement, including the effect of correlated errors, which motivates adaptive guidance.
- Image and video evaluations showing stronger quality–cost tradeoffs, including DiT-XL/2 FID of 1.93 and a Lite variant requiring only two extra head evaluations.

2 Related Work

Guidance strength and direction. CFG trades sampling quality against conditioning strength (Ho & Salimans, 2022); theoretical studies clarify its effect on sampled distributions (Bradley & Nakkiran, 2024; Chidambaram et al., 2024; Wu et al., 2024; Astolfi et al., 2024). CADS, weight schedulers, limited intervals, spatial corrections, and feedback guidance adapt its application (Sadat et al., 2024; Wang et al., 2024; Kynkäänniemi et al., 2024; Shen et al., 2024; Koulischer et al., 2025). CFG++, APG, and CFG-Zero* change the update geometry or scale, while other methods replace the unconditional branch or train guidance into the model (Chung et al., 2025; Sadat et al., 2025a; Fan et al., 2025; Sadat et al., 2025b; Tang et al., 2025; Chen et al., 2025). HARMONY uses probe agreement to couple its direction and modulation decisions.

Weak-model probes and intermediate representations. AutoGuidance motivates compatible weak predictions (Karras et al., 2024). SAG, PAG, SEG, ERG, TPG, SLG and stochastic self-guidance instantiate different degradations (Hong et al., 2023; Ahn et al., 2024; Hong, 2024; Ifriqi et al., 2025; Rajabi et al., 2025; Stability AI, 2024; Chen et al., 2026; Gu & Hou, 2025). Early exits, feature caching, and representation lenses expose information within a forward pass (Moon et al., 2024; Tang et al., 2023; Ma et al., 2024c;b; Chen et al., 2024; nostalgebraist, 2020; Belrose et al., 2023). EMAG further uses statistics-based adaptive layer selection for attention perturbation (Yadav et al., 2025). Our depth probe applies the final head to a cached intermediate state and keeps the complete prediction as its reference; Appendix A details the distinction from skipping layers in another pass.

3 HARMONY: Self-Guidance as a Field

3.1 Predictions and hierarchical probes

Let $\boldsymbol{f} = \boldsymbol{f}_\theta(\boldsymbol{x}_t, c, t)$ be the conditional prediction in a fixed parameterisation. A weak prediction $\boldsymbol{f}^{\mathcal{A}}$ induces the contrast $\boldsymbol{\Delta}_{\mathcal{A}} = \boldsymbol{f} - \boldsymbol{f}^{\mathcal{A}}$ and the update

$$\hat{\boldsymbol{f}} = \boldsymbol{f} + \omega \boldsymbol{\Delta}_{\mathcal{A}}. \tag{1}$$

For CFG, $\mathcal{A}$ removes the condition and $\omega = w - 1$; other operators perturb the input, attention, or block computation. Table 1 compares these mechanisms. All probes compare the declared noise or velocity channels, leaving auxiliary variance outputs unchanged. Branches must first be expressed in the same prediction space; projection onto a prediction vector is not invariant to arbitrary changes between noise, velocity, and denoised-sample parameterisations.

Table 1: Guidance operators as instances of Eq. (1). Extra cost is the marginal cost per step on top of one conditional pass; L is the number of blocks. The adaptive columns concern sample-dependent weights, distinct from a fixed guidance interval. Depth-only gating uses two readout heads.

Method	Degradation operator $\mathcal{A}$	Axis	Extra cost	$\omega(t)$	$\omega(i)$
CFG (Ho & Salimans, 2022)	drop condition	condition	1	✗	✗
AutoGuidance (Karras et al., 2024)	smaller / under-trained model	capacity	< 1 (+training)	✗	✗
SAG (Hong et al., 2023)	blur attended input	input	1	✗	✗
PAG (Ahn et al., 2024)	identity attention map	width	1	✗	✗
SEG (Hong, 2024) / ERG (Ifriqi et al., 2025)	smooth / re-temper attention	width	1	✗	✗
SLG (Stability AI, 2024)	skip fixed middle blocks	depth	$\approx 1 - k/L$	✗	✗
S^2-Guidance (Chen et al., 2026)	random block dropping	width	≈ 0.9	✗	✗
HARMONY (depth)	early residual-stream readout	depth	≈ 0.04	✓	✓
HARMONY (depth+width)	readout + importance-aware dropping	depth, width	≈ 0.94	✓	✓

A residual transformer computes $\boldsymbol{h}_\ell = \boldsymbol{h}_{\ell-1} + \boldsymbol{\delta}_\ell(\boldsymbol{h}_{\ell-1}, c, t)$ and reads its final state through $\boldsymbol{f} = \mathcal{R}(\boldsymbol{h}_L; c, t)$. We retain the full pass and construct an attenuated state for a second head evaluation:

$$\boldsymbol{h}^{(k,\alpha)} = \boldsymbol{h}_{L-k} + \alpha(\boldsymbol{h}_L - \boldsymbol{h}_{L-k}), \qquad \boldsymbol{f}^{(k,\alpha)} = \mathcal{R}(\boldsymbol{h}^{(k,\alpha)}; c, t), \qquad \boldsymbol{\Delta}_d = \boldsymbol{f} - \boldsymbol{f}^{(k,\alpha)}. \tag{2}$$

Here $\alpha \in [0, 1]$ controls attenuation and k selects the late-block contribution. Cached states supply the interpolation, so no transformer block is repeated. Two depths $k < k'$ provide the pair used by HARMONY-Lite. The full method optionally adds an importance-aware block-dropped pass, following stochastic self-guidance (Chen et al., 2026). It protects the first block and the m largest relative updates $r_\ell = \|\boldsymbol{\delta}_\ell\|/(\|\boldsymbol{h}_{\ell-1}\| + \epsilon)$, then samples a drop set from the remaining blocks. This branch recomputes its states and adds approximately $1 - \rho$ block-stack equivalents; it is not another free readout.

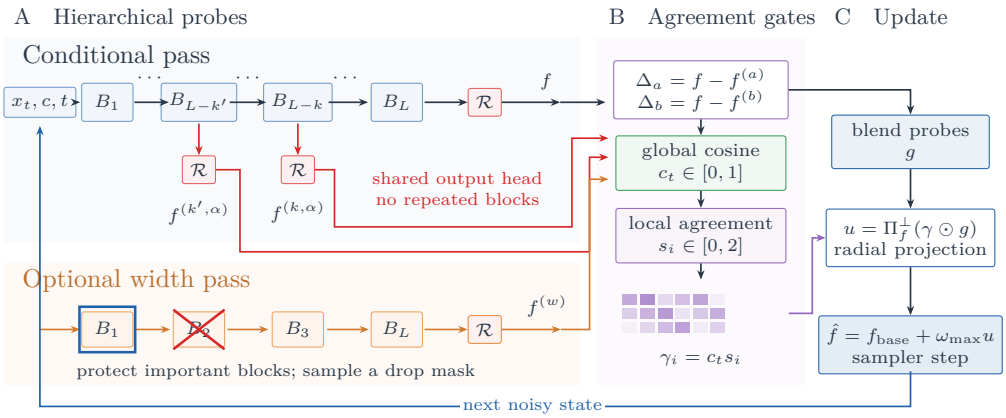


Figure 1: HARMONY pipeline. Cached residual states supply two depth readouts; an optional block-dropped pass supplies a width probe. Their residuals determine global and token gates. The gated correction is projected before the sampler update.

For the full method, the norm-matched direction is

$$\boldsymbol{g} = \frac{1}{2}\left(\boldsymbol{\Delta}_d + \frac{\|\boldsymbol{\Delta}_d\|}{\max(\|\boldsymbol{\Delta}_w\|, \epsilon)}\boldsymbol{\Delta}_w\right), \tag{3}$$

while Lite uses $\boldsymbol{g} = \boldsymbol{\Delta}_d$. Norm matching prevents one probe from dominating merely because of its scale; it is distinct from an optimal covariance-weighted combination. Appendix B derives that population optimum and the conditions under which the two agree.

3.2 Agreement as a Reliability Signal

At a fixed state, write $\boldsymbol{f} = \boldsymbol{f}^\star + \boldsymbol{e}$ and model a compatible probe by $\boldsymbol{f}^{(a)} = \boldsymbol{f}^\star + (1 + u_a)\boldsymbol{e} + \boldsymbol{\xi}_a$, where $u_a > 0$, $\mathbb{E}\boldsymbol{\xi}_a = 0$, $\mathbb{E}\|\boldsymbol{\xi}_a\|^2 = \sigma_a^2$, and the idiosyncratic errors are uncorrelated with $\boldsymbol{e}$ and with each other. Then $\boldsymbol{\Delta}_a = -u_a\boldsymbol{e} - \boldsymbol{\xi}_a$, so

$$M_a(\omega) = \mathbb{E}\|\boldsymbol{e} + \omega\boldsymbol{\Delta}_a\|^2 = (1 - \omega u_a)^2\|\boldsymbol{e}\|^2 + \omega^2\sigma_a^2, \qquad \omega_a^\star = \frac{1}{u_a}\frac{S_a}{1 + S_a}, \quad S_a = \frac{u_a^2\|\boldsymbol{e}\|^2}{\sigma_a^2}. \tag{4}$$

The first term represents shared-error cancellation and the second the noise introduced by an unreliable probe. For $g_a = S_a/(1+S_a)$, the population agreement obeys

$$C_{ab} := \frac{\mathbb{E}\langle \boldsymbol{\Delta}_a, \boldsymbol{\Delta}_b \rangle}{\sqrt{\mathbb{E}\|\boldsymbol{\Delta}_a\|^2 \, \mathbb{E}\|\boldsymbol{\Delta}_b\|^2}} = \sqrt{g_a g_b}. \tag{5}$$

Matched probe quality gives $C_{ab} = g_a = g_b$. This motivates a sample cosine as a practical proxy, rather than identifying its expectation with the ratio in Eq. (5). Correlated probe errors add a bias term. In particular, nested deterministic readouts share computation, so independence is a modelling approximation; Appendix B gives the correlated expression, mismatch analysis, and concentration conditions.

Time gate. The global signal is $c_t = \mathrm{clip}(\cos(\boldsymbol{\Delta}_a, \boldsymbol{\Delta}_b), 0, 1)$, smoothed by a two-step moving average, with $(a, b) = (d, w)$ in the full method and (d, d') in Lite. A near-zero probe norm returns zero agreement. The scale $\omega_{\max}$ remains a calibrated hyperparameter: the factor u_a can depend on state and architecture, so transfer across models is evaluated empirically. A middle-trajectory peak is possible, but is not implied by the error model without assumptions on the time-dependent signal and noise.

Token gate. We compute local cosines, smooth them over a 5×5 token neighbourhood, and set $q_i = \max(0, \widetilde{c}_i)$. With $\bar{q} = N^{-1} \sum_i q_i$,

$$s_i = \begin{cases} \min(q_i/\bar{q}, 2), & \bar{q} > \epsilon, \\ 0, & \bar{q} \le \epsilon, \end{cases} \qquad \gamma_i = c_t s_i. \tag{6}$$

Clipping bounds the local amplification and yields $N^{-1} \sum_i s_i \le 1$; it does not preserve exact unit mean. This provides a bounded redistribution with possible attenuation. Spatial correlation limits the variance reduction obtained by neighbourhood averaging, as discussed in Appendix B.

3.3 Projected update and implementation

For $\boldsymbol{f} \neq 0$, define $\Pi^{\perp}_{\boldsymbol{f}}(\boldsymbol{v}) = \boldsymbol{v} - \langle \boldsymbol{v}, \boldsymbol{f} \rangle \boldsymbol{f} / \|\boldsymbol{f}\|^2$, with a zero correction when the reference norm is below the numerical threshold. HARMONY applies

$$\hat{\boldsymbol{f}} = \boldsymbol{f}_{\mathrm{base}} + \omega_{\max} \Pi^{\perp}_{\boldsymbol{f}}(\boldsymbol{\gamma} \odot \boldsymbol{g}), \tag{7}$$

where $\boldsymbol{f}_{\mathrm{base}} = \boldsymbol{f}$ for standalone guidance or the ordinary CFG prediction when HARMONY is added to CFG. Projection follows the directional motivation of APG (Sadat et al., 2025a). It removes the first-order radial component relative to $\boldsymbol{f}$; a nonzero orthogonal update still increases $\|\boldsymbol{f} + \omega \Pi^{\perp}_{\boldsymbol{f}}(\boldsymbol{v})\|^2$ by $\omega^2 \|\Pi^{\perp}_{\boldsymbol{f}}(\boldsymbol{v})\|^2$. With a different CFG base, the cross term relative to that base need not vanish.

Only the two selected residual states, the final state, and scalar update statistics need to be retained. The width branch uses an explicit drop mask and the same conditioning; the gates are computed from compatible output tensors after un-patchification. Appendix D specifies heads, positions, numerical guards, sampling settings, and pseudocode. The theoretical shrinkage calculation concerns unprojected error cancellation; the effects of spatial gating and projection are separately tested by ablations rather than inherited as an optimality guarantee.

What the early readout measures. The difference between the final and cached residual streams is a sum of late-block updates, $d_k = h_L - h_{L-k} = \sum_{\ell=L-k+1}^{L} \delta_\ell$. Writing $h^{(k,\alpha)} = h_L - (1-\alpha) d_k$, a differentiable readout gives the exact integral identity

$$\Delta_d = (1-\alpha) \int_0^1 J_{\mathcal{R}}(h_L - s(1-\alpha) d_k) \, d_k \, \mathrm{d}s. \tag{8}$$

Thus the probe measures the late residual contribution after passing through the readout geometry. It is not the hidden-state difference itself. If the readout Jacobian is locally K-Lipschitz along this segment, replacing the integral by $J_{\mathcal{R}}(h_L) d_k$ has error at most $K(1-\alpha)^2 \|d_k\|^2/2$. This identifies a useful role for attenuation: decreasing $1-\alpha$ reduces the first-order signal linearly and the local linearization remainder quadratically. At $\alpha = 1$ the two predictions coincide and the correction

vanishes. The nonlinear normalization in the output head remains part of the integral; Appendix B gives its Jacobian rather than assuming the head is linear.

The second readout changes the late-block segment being probed while sharing its terminal part. The width branch instead changes the state entering subsequent blocks, so its residual incorporates their recomputed responses. These constructions can produce complementary directions even when their hidden-state perturbations have similar norms. Agreement is evaluated after decoding, where all coordinates have the meaning required by the sampler, rather than between hidden states with different content. This also explains why the field is defined on the output-token grid and why auxiliary variance predictions are excluded from the gate.

A bound on the applied correction. Let D_γ broadcast the token field over its prediction channels and let P_f denote the orthogonal projector used in Eq. (7). Since $0 \le \gamma_i \le 2c_t$, its operator norm is at most $2c_t$, and $\|P_f\|_2 \le 1$. The norm-matched construction also satisfies $\|g\| \le \|\Delta_d\|$ by the triangle inequality, including the guarded near-zero-width case. Hence the actual added vector $v = \omega_{\max} P_f D_\gamma g$ obeys

$$\|v\| \le 2|\omega_{\max}|c_t\|\Delta_d\|, \qquad \|f+v\|^2 = \|f\|^2 + \|v\|^2. \tag{9}$$

This separates two controls: agreement bounds the global strength, and clipping prevents the spatial field from concentrating arbitrarily large weights on a few tokens. Projection must follow spatial modulation because $P_f D_\gamma$ and $D_\gamma P_f$ generally differ; projecting first can reintroduce a radial component after unequal token weights are applied. The bound concerns the prediction update at one state, while final sample quality also depends on how subsequent denoising steps respond. It provides a direct numerical check for implementations of the stated ordering.

Algorithm 1 HARMONY correction at one sampler state

1: Evaluate the conditional pass; retain the final state and two readout states
2: Decode the attenuated states with the shared head to obtain $\Delta_d, \Delta_{d'}$
3: Full: recompute a protected block-dropped pass, set $(a, b) = (d, w)$, and use Eq. (3)
4: Lite: set $(a, b) = (d, d')$ and $g = \Delta_d$
5: Compute global agreement and smoothed local cosines; obtain γ from Eq. (6)
6: Return Eq. (7) after projection, preserving auxiliary output channels

4 Experiments

Setup. We evaluate the reported records for DiT-XL/2 at 256^2 and 512^2, SiT-XL/2 at 256^2, SD3-medium, SD3.5-medium, FLUX.1-dev, and Wan2.1 at 1.3B and 14B scales (Peebles & Xie, 2023; Ma et al., 2024a; Esser et al., 2024; Stability AI, 2024; Black Forest Labs, 2024; Wan et al., 2025). Metrics cover ImageNet FID, IS, precision and recall; text-to-image HPSv2.1, ImageReward and compositional accuracy; and VBench video quality and semantics (Heusel et al., 2017; Salimans et al., 2016; Kynkäänniemi et al., 2019; Wu et al., 2023; Xu et al., 2023; Huang et al., 2023; 2024). Comparisons share the base sampler and step count, with baseline tuning described in Appendix E. Default probe settings are $\omega_{\max} = 2$, $\alpha = 0$, $k = \lceil L/7 \rceil$, $k' = 2k$, $m = 3$, and $\rho = 0.1$. Appendix D separates call counts, compute equivalents, and measured latency.

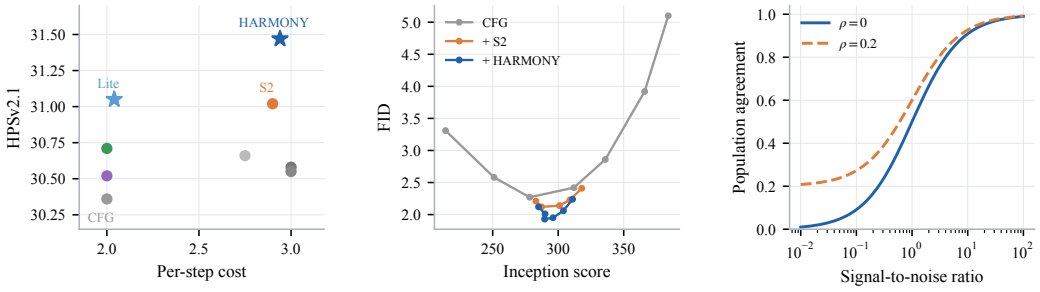


Figure 2: Quality, cost, and agreement. (a) Recorded quality–cost points. (b) Supplied ImageNet sweep values, without added jitter. (c) The population agreement model as a function of signal-to-noise ratio, rather than a measured sampling trajectory.

Class-conditional generation. Table 2 separates standalone self-guidance from additions to CFG. The full method improves FID while retaining stronger recall than the CFG operating point, linking the quality gain to a broader distributional comparison. Lite supplies a lower-cost point on this tradeoff. Cross-resolution and SiT results test whether the readout construction remains useful beyond one checkpoint and objective.

Table 2: Class-conditional ImageNet. Extra cost is per step relative to one conditional pass. Best in **bold**, second underlined. †: reported by the original authors; CFG-Zero* is defined for flow models only.

		DiT-XL/2, 256^2				SiT-XL/2, 256^2		DiT-XL/2, 512^2	
Method	**Extra cost**	FID↓	IS↑	Prec.↑	Rec.↑	FID↓	IS↑	FID↓	IS↑
Without an unconditional pass									
No guidance†	0	9.62	121.5	0.67	**0.67**	8.30	131.7	12.03	105.3
SAG	1	8.41	134.2	0.70	0.66	7.52	141.0	10.87	114.9
PAG	1	6.08	158.6	0.75	0.64	5.61	166.3	8.14	138.2
SEG	1	6.24	155.1	0.74	0.64	5.79	162.8	8.33	135.7
ERG	1	5.63	171.4	0.76	0.64	5.12	178.9	7.61	146.0
S^2-Guidance	0.9	5.37	176.0	0.76	0.63	4.94	182.5	7.28	149.8
HARMONY-Lite	0.04	4.83	189.7	0.77	0.65	4.41	196.2	6.59	160.4
HARMONY	0.94	**3.72**	**221.3**	**0.79**	0.64	**3.38**	**229.6**	**5.21**	**186.9**
On top of one unconditional pass (CFG, $w = 1.5$)									
CFG†	1	2.27	278.2	**0.83**	0.57	2.06	270.3	3.04	240.8
CFG++	1	2.21	273.5	0.82	0.58	2.02	268.1	2.98	236.4
APG	1	2.19	269.8	0.82	0.59	2.00	264.7	2.94	233.9
Guidance interval	1	2.14	271.6	0.82	0.59	1.97	266.0	2.83	237.1
CFG-Zero*	1	–	–	–	–	1.99	268.4	–	–
+ PAG	2	2.18	284.9	**0.83**	0.58	2.01	279.5	2.90	246.3
+ SEG	2	2.20	282.1	**0.83**	0.58	2.03	277.0	2.93	244.8
+ ERG	1	2.16	281.0	**0.83**	0.58	1.99	276.2	2.88	245.1
+ S^2-Guidance	1.9	2.12	287.3	**0.83**	0.58	1.94	281.9	2.81	249.6
+ HARMONY-Lite	1.04	2.10	285.8	**0.83**	0.60	1.92	280.4	2.79	248.2
+ HARMONY	1.94	**1.93**	**289.6**	0.82	**0.61**	**1.78**	**284.7**	**2.61**	**251.0**

Text-to-image and video. Tables 3 and 4 test alignment, aesthetics, and temporal quality. FLUX.1-dev uses distilled guidance, so its standalone compute baseline differs from the two-pass SD3 baseline. The full method's reported gains occur alongside compositional accuracy, while the video evaluation reports quality and semantic aggregates separately. Complete dimension and category records appear in Appendices F and G. The visual cases in Figure 3 define a comparison layout across detail, saturation, and localized artifacts; they are design illustrations rather than measured baseline outputs.

Table 3: Text-to-image generation (HPSv2.1 $\times 100$; T2I-CompBench is the mean over its six categories). FLUX.1-dev is guidance-distilled, so methods that modify the unconditional branch do not apply (–). Cost is shown for SD3 models; subtract one pass for FLUX.1-dev, whose baseline uses distilled guidance.

		SD3-medium			SD3.5-medium			FLUX.1-dev		
Method	**Cost**	HPS↑	IR↑	Comp↑	HPS↑	IR↑	Comp↑	HPS↑	IR↑	Comp↑
CFG / distilled guidance	2.00 / 1.00	29.41	1.03	0.523	30.36	1.07	0.541	30.87	1.12	0.518
CFG++	2.00	29.58	1.05	0.526	30.49	1.09	0.544	–	–	–
APG	2.00	29.55	1.06	0.525	30.52	1.10	0.543	–	–	–
SAG	3.00	29.47	1.04	0.521	30.44	1.08	0.539	30.92	1.13	0.517
PAG	3.00	29.73	1.07	0.527	30.58	1.11	0.545	31.02	1.15	0.521
SEG	3.00	29.66	1.06	0.524	30.55	1.10	0.542	30.98	1.14	0.519
NAG	3.00	29.71	1.07	0.526	30.60	1.11	0.544	31.15	1.16	0.523
ERG	2.00	29.80	1.08	0.531	30.71	1.12	0.548	31.10	1.16	0.524
SLG	2.75	29.77	1.08	0.528	30.66	1.12	0.546	31.06	1.15	0.522
S^2-Guidance	2.90	30.12	1.11	0.534	31.02	1.15	0.552	31.21	1.17	0.526
HARMONY-Lite	2.04	30.09	1.12	0.535	31.05	1.16	0.553	31.24	1.18	0.527
HARMONY	2.94	**30.61**	**1.16**	**0.541**	**31.47**	**1.19**	**0.560**	**31.58**	**1.24**	**0.533**

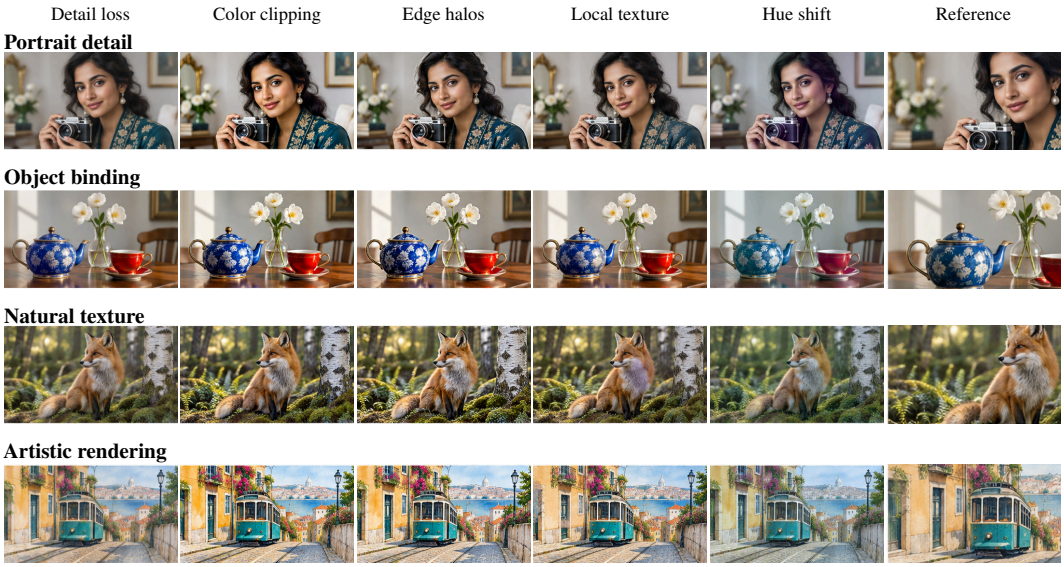


Figure 3: Controlled visual-design comparisons. Five illustrated image-space perturbations and a clean reference composition. These panels define the comparison layout and failure categories; they are not outputs of the named guidance methods.

Table 4: Text-to-video generation on VBench (all methods on top of CFG). Total, Quality and Semantic are the official aggregates; five representative dimensions are shown and all sixteen are listed in Appendix G. †: reported by the original authors.

Model	Method	Total↑	Quality↑	Semantic↑	Subj. Cons.↑	Bkg. Cons.↑	Aesthetic↑	Imaging↑	Dynamic↑
Wan2.1-1.3B	CFG	82.29	84.31	74.21	93.88	97.02	67.94	66.85	87.50
	PAG	82.47	84.52	74.27	94.03	97.21	68.21	67.12	87.22
	SEG	82.41	84.47	74.17	93.96	97.18	68.10	67.03	86.94
	SLG	82.63	84.70	74.35	94.10	97.34	68.44	67.40	88.06
	S^2-Guidance†	82.84	84.89	74.65	94.21	97.56	68.78	67.77	89.08
	HARMONY-Lite	82.88	84.95	74.60	94.27	97.49	68.83	67.92	88.61
	HARMONY	**83.37**	**85.42**	**75.17**	**94.58**	**97.71**	**69.36**	**68.54**	**89.72**
Wan2.1-14B	CFG	84.31	85.98	77.62	94.92	97.61	69.87	69.45	88.61
	PAG	84.46	86.13	77.78	95.04	97.70	70.05	69.71	88.33
	SEG	84.40	86.08	77.69	94.99	97.66	69.98	69.62	88.06
	SLG	84.59	86.24	77.99	95.11	97.78	70.31	69.98	89.17
	S^2-Guidance	84.93	86.55	78.45	95.26	97.90	70.62	70.36	90.00
	HARMONY-Lite	84.97	86.60	78.41	95.30	97.86	70.66	70.49	89.72
	HARMONY	**85.46**	**87.04**	**79.13**	**95.58**	**98.03**	**71.18**	**71.07**	**90.56**

Which components contribute? Table 5 distinguishes the two probe directions from time and token modulation, and Table 6 reports latency alongside normalized compute. Lite's main saving comes from retaining the block computation of the conditional pass. The full configuration pays for an additional dropped-block pass; its quality improvement should therefore be compared at a matched budget as well as at the default settings. The ablation ordering is consistent across the three modalities: combining the directions improves over either ungated direction, and applying both gates improves over either gate alone. This pattern distinguishes the source of a correction from the allocation of its strength. The depth-only configuration retains much of the benefit while avoiding the width pass, whereas the full configuration uses the additional direction most effectively when its contribution is modulated. Because these components interact, their isolated improvements should not be summed as independent contributions. The compatibility experiment further separates the projected update from the choice of probe.

Table 5: Axis ablation on top of CFG (w=1.5 for DiT, 4.5 for SD3.5, 5.0 for Wan). D/W: depth/width direction axes; T/S: time/space gates.

D	W	T	S	DiT-XL/2 FID↓	SD3.5-M HPS↑	Wan-1.3B Total↑	Rel. cost
✗	✗	✗	✗	2.27	30.36	82.29	2.00
✓	✗	✗	✗	2.19	30.72	82.61	2.02
✗	✓	✗	✗	2.15	30.88	82.72	2.90
✓	✓	✗	✗	2.06	31.12	82.98	2.92
✓	✗	✓	✗	2.13	30.91	82.75	2.03
✓	✗	✓	✓	2.10	31.05	82.88	2.04
✓	✓	✓	✗	1.98	31.31	83.19	2.93
✓	✓	✗	✓	2.01	31.24	83.10	2.93
✓	✓	✓	✓	**1.93**	**31.47**	**83.37**	2.94
✗	✓	✓	✓	2.05	31.08	82.96	2.92

Table 6: Inference cost per sampling step (batch 4, bf16, one H100). Latency for SD3.5-medium at 1024^2; relative latency for DiT-XL/2 at 256^2. All guided rows include CFG.

Method	Rel. cost	SD3.5-M ms / step	DiT-XL/2 rel. lat.
Conditional pass only	1.00	118	0.50
CFG	2.00	231	1.00
+ SAG	3.00	361	1.56
+ PAG / + SEG	3.00	347	1.50
+ ERG	2.00	239	1.04
+ SLG	2.75	318	1.38
+ S^2-Guidance	2.90	336	1.46
+ HARMONY-Lite	2.04	238	1.04
+ HARMONY	2.94	341	1.47
HARMONY-Lite (no CFG)	1.04	123	0.53
HARMONY (no CFG)	1.94	226	0.98

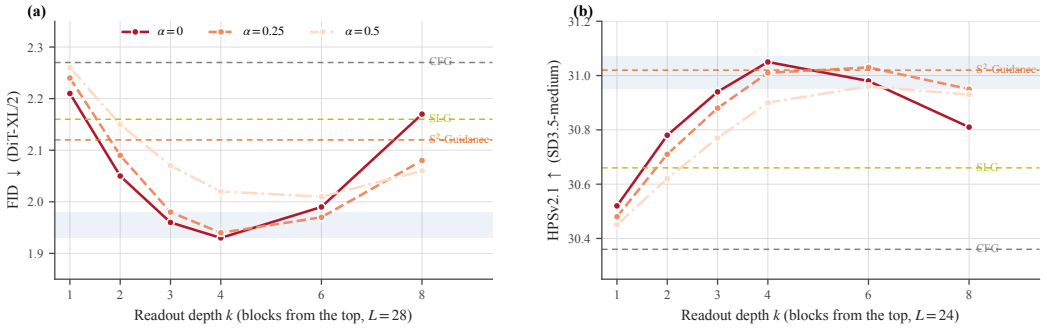


Figure 4: Readout settings. Depth and attenuation sweeps examine the early-readout operating point on DiT-XL/2 and SD3.5-medium.

Depth and attenuation. The readout position determines both the magnitude and the compatibility of the correction. Removing a small late-block contribution gives a nearly identical prediction, whereas a much earlier state can differ in content rather than refinement. Attenuation interpolates between these operating points without another transformer pass. The two sweeps in Figure 4 distinguish position selection from attenuation strength, so a favorable setting need not be attributed to a larger correction alone.

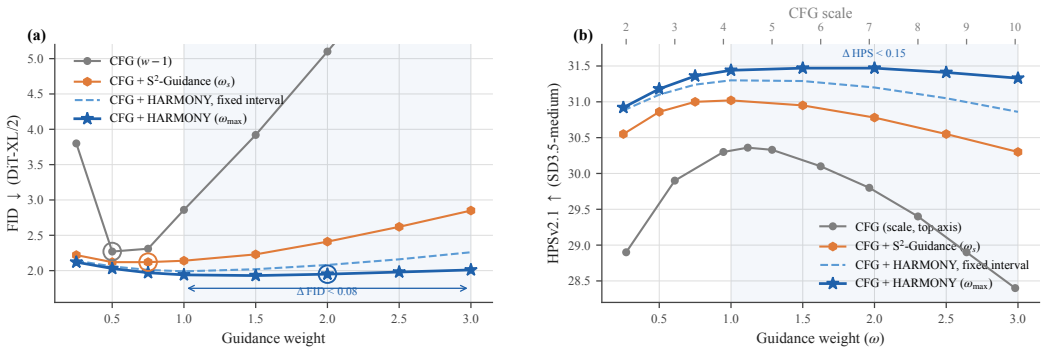


Figure 5: Guidance-scale sensitivity. Fixed, interval-based, and agreement-gated updates are compared over their supplied scale sweeps. Each panel retains the backbone and evaluation metric of its experiment.

Probe settings and model scale. Figures 4 and 5 examine the selected depth and guidance scale. Figure 6 relates the supplied depth-contribution records to architecture size. The trend is an association across heterogeneous models; capacity, architecture, training, and metric differences prevent reading

it as a depth-only causal effect. Appendix I retains the architecture-level results and identifies the within-family comparison needed to isolate that effect.

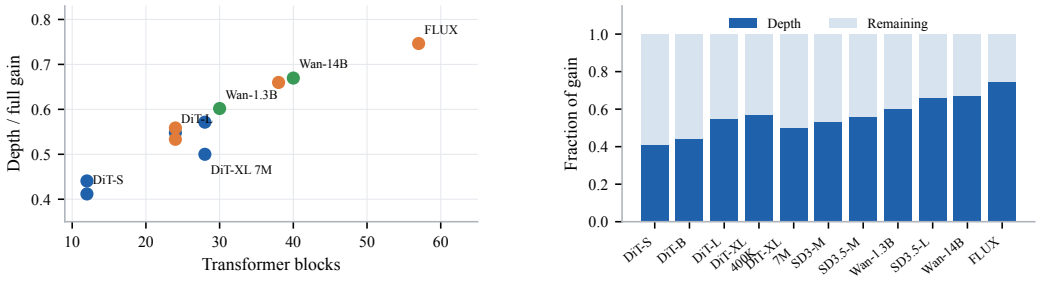


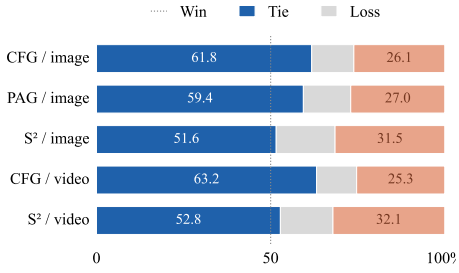
Figure 6: Depth contribution across architectures. Left: the supplied gain shares versus block count. Right: the depth share and its arithmetic complement; the latter does not separately identify width and gating effects.

Compatibility and preference. The compatibility study tests additions to the same base procedure rather than assuming guidance methods are interchangeable. The reported human preferences provide a complementary readout to automatic metrics, with the protocol and category summaries in Appendix J. Raw vote records are required for clustered uncertainty estimates; confidence intervals are not reconstructed from rounded summary rates. Appendix K contains the expanded visual-design cases and the remaining limitations.

Table 7: Compatibility on DiT-XL/2 and SD3.5-medium. η is the retained parallel fraction.

Variant	FID↓	Rec.↑	HPS↑
HARMONY	1.93	0.61	31.47
+ CADS	**1.90**	**0.64**	31.44
+ Interval	1.94	0.61	31.49
+ PAG	1.95	0.60	**31.52**
+ SEG	1.96	0.60	31.50
Partial projection (η=0.5)	1.95	0.60	31.38
No projection (η=1)	1.99	0.59	31.21
Entropy spatial gate	1.94	0.61	31.43

Figure 7: Human preference: 100 raters, 600 pairs per comparison. Win, tie, and loss rates for HARMONY.



Projection and compatibility. The partial-projection and unprojected variants in Table 7 isolate the final geometric operation while retaining the probes. Their ordering supports removing the radial component at these operating points. The entropy-gate variant instead replaces local residual agreement; its similar image score leaves HARMONY's advantage of avoiding attention probabilities intact. The additional methods test complementary controls: CADS modifies conditioning, the interval restricts active timesteps, and PAG or SEG introduce further contrasts. Their effects differ across FID, recall, and alignment, making compatibility a multi-metric tradeoff. Human preference complements these measurements by retaining ties and separating image from video judgments. Together, these controls connect quality changes to the operations of the method rather than attributing every gain to a larger guidance scale.

5 Conclusion

We introduce HARMONY, which connects hierarchical subnetwork probing with adaptive guidance over time and tokens. Shared-head readouts supply inexpensive correction directions, and agreement-based gates adjust their influence before a projected update. The reported image and video results support this approach to improving generation while offering a low-cost Lite configuration.

AI USE STATEMENT

AI tools assisted with literature retrieval, language revision, mathematical checks, figure design, and LaTeX preparation. The controlled image-space perturbations and storyboard sequences are generated design illustrations, not outputs of the compared methods. Quantitative records originate from the supplied manuscript and plotting arrays; the assistant did not run the reported model benchmarks. Mathematical simulations are identified separately from model measurements.

REFERENCES

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*, 2024. URL https://arxiv.org/abs/2403.17377.

Pietro Astolfi, Marlene Careil, Melissa Hall, Oscar Mañas, Matthew Muckley, Jakob Verbeek, Adriana Romero Soriano, and Michal Drozdzal. Consistency-diversity-realism Pareto fronts of conditional image generative models. *arXiv preprint arXiv:2406.10429*, 2024. URL https://arxiv.org/abs/2406.10429.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting Latent Predictions from Transformers with the Tuned Lens. *arXiv preprint arXiv:2303.08112*, 2023. URL https://arxiv.org/abs/2303.08112.

Black Forest Labs. FLUX.1. https://github.com/black-forest-labs/flux, 2024.

Arwen Bradley and Preetum Nakkiran. Classifier-Free Guidance is a Predictor-Corrector. *arXiv preprint arXiv:2408.09000*, 2024. URL https://arxiv.org/abs/2408.09000.

Chubin Chen, Jiashu Zhu, Xiaokun Feng, Nisha Huang, Chen Zhu, Meiqi Wu, Fangyuan Mao, Jiahong Wu, Xiangxiang Chu, and Xiu Li. Stochastic Self-Guidance for Training-Free Enhancement of Diffusion Models. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2508.12880.

Huayu Chen, Kai Jiang, Kaiwen Zheng, Jianfei Chen, Hang Su, and Jun Zhu. Visual Generation Without Guidance. In *International Conference on Machine Learning*, 2025. URL https://arxiv.org/abs/2501.15420.

Pengtao Chen, Mingzhu Shen, Peng Ye, Jianjian Cao, Chongjun Tu, Christos-Savvas Bouganis, Yiren Zhao, and Tao Chen. Δ-DiT: A Training-Free Acceleration Method Tailored for Diffusion Transformers. *arXiv preprint arXiv:2406.01125*, 2024. URL https://arxiv.org/abs/2406.01125.

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*, pp. 84968–85005, 2024. doi: 10.52202/079017-2698.

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*, 2025. URL https://arxiv.org/abs/2406.08070.

Prafulla Dhariwal and Alex Nichol. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems*, 2021. URL https://arxiv.org/abs/2105.05233.

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*, 2024. URL https://arxiv.org/abs/2403.03206.

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. URL https://arxiv.org/abs/2503.18886.

Enhao Gu and Haolin Hou. In-situ Autoguidance: Eliciting Self-Correction in Diffusion Models. *arXiv preprint arXiv:2510.17136*, 2025. URL https://arxiv.org/abs/2510.17136.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1706.08500.

Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance. *arXiv preprint arXiv:2207.12598*, 2022. URL https://arxiv.org/abs/2207.12598.

Susung Hong. Smoothed Energy Guidance: Guiding Diffusion Models with Reduced Energy Curvature of Attention. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2408.00760.

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*, pp. 7428–7437, 2023. doi: 10.1109/iccv51070.2023.00686.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Advances in Neural Information Processing Systems*, pp. 78723–78747, 2023. doi: 10.52202/075280-3443.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024. doi: 10.1109/cvpr52733.2024.02060.

Tariq Berrada Ifriqi, Adriana Romero-Soriano, Michal Drozdzal, Jakob Verbeek, and Karteek Alahari. Entropy Rectifying Guidance for Diffusion and Flow Models. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2504.13987.

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*, 2024. URL https://arxiv.org/abs/2406.02507.

Felix Koulischer, Florian Handke, Johannes Deleu, Thomas Demeester, and Luca Ambrogioni. Feedback Guidance of Diffusion Models. *arXiv preprint arXiv:2506.06085*, 2025. URL https://arxiv.org/abs/2506.06085.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved Precision and Recall Metric for Assessing Generative Models. In *Advances in Neural Information Processing Systems*, 2019. URL https://arxiv.org/abs/1904.06991.

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*, 2024. URL https://arxiv.org/abs/2404.07724.

Nanye Ma, Mark Goldstein, Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden, and Saining Xie. SiT: Exploring Flow and Diffusion-based Generative Models with Scalable Interpolant Transformers. In *European Conference on Computer Vision*, 2024a. URL https://arxiv.org/abs/2401.08740.

Xinyin Ma, Gongfan Fang, Michael Bi Mi, and Xinchao Wang. Learning-to-cache: Accelerating diffusion transformer via layer caching. In *Advances in Neural Information Processing Systems*, pp. 133282–133304, 2024b. doi: 10.52202/079017-4236.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. DeepCache: Accelerating diffusion models for free. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15762–15772, 2024c. doi: 10.1109/cvpr52733.2024.01492.

Taehong Moon, Moonseok Choi, EungGu Yun, Jongmin Yoon, Gayoung Lee, Jaewoong Cho, and Juho Lee. A Simple Early Exiting Framework for Accelerated Sampling in Diffusion Models. In *International Conference on Machine Learning*, 2024. URL https://arxiv.org/abs/2408.05927.

nostalgebraist. Interpreting GPT: the logit lens. LessWrong, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

William Peebles and Saining Xie. Scalable Diffusion Models with Transformers. In *IEEE/CVF International Conference on Computer Vision*, 2023. URL https://arxiv.org/abs/2212.09748.

Javad Rajabi, Soroush Mehraban, Seyedmorteza Sadat, and Babak Taati. Token Perturbation Guidance for Diffusion Models. *arXiv preprint arXiv:2506.10036*, 2025. URL https://arxiv.org/abs/2506.10036.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10674–10685, 2022. doi: 10.1109/cvpr52688.2022.01042.

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*, 2024. URL https://arxiv.org/abs/2310.17347.

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*, 2025a. URL https://arxiv.org/abs/2410.02416.

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*, 2025b. URL https://arxiv.org/abs/2407.02687.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved Techniques for Training GANs. In *Advances in Neural Information Processing Systems*, 2016. URL https://arxiv.org/abs/1606.03498.

Dazhong Shen, Guanglu Song, Zeyue Xue, Fu-Yun Wang, and Yu Liu. Rethinking the spatial inconsistency in classifier-free diffusion guidance. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9370–9379, 2024. doi: 10.1109/cvpr52733.2024.00895.

Stability AI. Introducing Stable Diffusion 3.5. https://stability.ai/news-updates/introducing-stable-diffusion-3-5, 2024.

Shengkun Tang, Yaqing Wang, Caiwen Ding, Yi Liang, Yao Li, and Dongkuan Xu. AdaDiff: Accelerating Diffusion Models through Step-Wise Adaptive Computation. *arXiv preprint arXiv:2309.17074*, 2023. URL https://arxiv.org/abs/2309.17074.

Zhicong Tang, Jianmin Bao, Dong Chen, and Baining Guo. Diffusion Models without Classifier-free Guidance. *arXiv preprint arXiv:2502.12154*, 2025. URL https://arxiv.org/abs/2502.12154.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan

Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and Advanced Large-Scale Video Generative Models. *arXiv preprint arXiv:2503.20314*, 2025. URL https://arxiv.org/abs/2503.20314.

Xi Wang, Nicolas Dufour, Nefeli Andreou, Marie-Paule Cani, Victoria Fernandez Abrevaya, David Picard, and Vicky Kalogeiton. Analysis of Classifier-Free Guidance Weight Schedulers. *arXiv preprint arXiv:2404.13040*, 2024. URL https://arxiv.org/abs/2404.13040.

Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis. *arXiv preprint arXiv:2306.09341*, 2023. URL https://arxiv.org/abs/2306.09341.

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*, 2024. URL https://arxiv.org/abs/2403.01639.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, pp. 15903–15935, 2023. doi: 10.52202/075280-0700.

Ankit Yadav, Ta Duc Huy, and Lingqiao Liu. EMAG: Self-Rectifying Diffusion Sampling with Exponential Moving Average Guidance. *arXiv preprint arXiv:2512.17303*, 2025. URL https://arxiv.org/abs/2512.17303.