PP-CFG: POLARITY-PRESERVING GUIDANCE FOR CONTROLLABLE FEW-STEP IMAGE AND VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-step diffusion distillation reduces sampling cost but often weakens adjustable classifier-free guidance (CFG) and runtime negative prompting. We introduce PP-CFG, an architecture that keeps positive and negative text conditions separate and combines their attention responses with opposite signs within a shared visual backbone. Its Scale-Time Polarity Mixer refines each condition using the guidance scale and denoising timestep, allowing prompt influence to vary throughout generation. The student retains a continuous scale input and selectable negative prompts with one visual backbone pass per step. We evaluate PP-CFG on Wan2.1 for video generation with distribution matching distillation (DMD), and on Qwen-Image with both latent consistency model (LCM) distillation and DMD. PP-CFG achieves 81.46 VBench in four steps and lower video teacher–student distances across guidance scales than step-distilled DMD. On NegElement-1K, our benchmark of 1,000 matched image and video prompt cases, it achieves higher target-removal rates than the corresponding teachers. These experiments support the method's generality across attention architectures, generation modalities, and distillation objectives while retaining competitive generation quality.

1 INTRODUCTION

Diffusion and flow-matching models produce high-quality images and videos (Wan et al., 2025; Wu et al., 2025), but repeated denoising evaluations make generation expensive. Few-step distillation reduces this cost by transferring a teacher's generation behavior to a student (Yin et al., 2024b;a). Users also need to adjust prompt adherence and suppress unwanted content as their prompts and preferences change. A practical student should retain both controls after distillation.

Classifier-free guidance (CFG) combines positive and negative predictions to control generation (Ho & Salimans, 2022). Fixed-guidance distillation absorbs guided predictions into student weights, typically losing runtime scale adjustment and negative prompting. Applying external CFG after distillation can degrade output quality (Fig. 1). Conditioning the student on guidance scale restores an adjustable input (Meng et al., 2023), but does not by itself retain a selectable negative prompt. Attention-based guidance provides a negative-prompt path (Chen et al., 2025; Guo & Du, 2025). Table 1 summarizes the considered configurations. The challenge is to retain both controls within one visual backbone pass while preserving the distinct roles of the two prompts.

We propose Polarity-Preserving CFG (PP-CFG), where *polarity* denotes the positive and negative roles of the text conditions. The method keeps these contexts separate and uses a Scale-Time Polarity Mixer to refine each one according to the requested guidance scale and denoising timestep. These refinements adjust how visual features attend to each prompt and what information the text provides at each denoising stage. Following Negative-Away Steer Attention (NASA) (Nguyen et al., 2025), polarity blocks subtract the negative attention response from the positive response to update one shared visual stream (Fig. 2). We jointly train the mixer and backbone using either LCM or DMD. At inference, users can choose both the guidance scale and negative prompt while the model uses one visual backbone pass per step.

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.

Setting	Adj.	Neg.	1-NFE
Fixed guidance	✗	✗	✓
LCM (SD)	✓	✗	✗
DMD (SD)	✓	✗	✗
LCM (CFG-E+SD)	✓	✗	✓
DMD (CFG-E+SD)	✓	✗	✓
PCM	✓	✓	✗
PP-CFG	✓	✓	✓

Table 1: Runtime controls. Adj.: adjustable scale; Neg.: negative prompts; 1-NFE: one visual backbone pass per step. SD: step distillation; CFG-E: CFG embedding; PCM: phased consistency model.

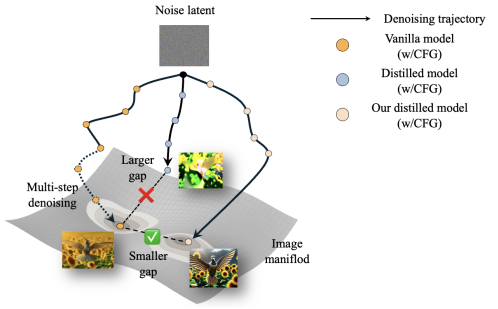

Figure 1: **CFG after distillation.** External CFG can degrade student outputs and increase their discrepancy from the teacher. PP-CFG learns scale-dependent guidance within a shared visual stream.

We evaluate guidance adjustment and generation quality, and introduce NegElement-1K to assess whether models suppress specified elements through negative prompting. The benchmark uses matched image and video prompt cases to examine this capability in both modalities. Our experiments show that PP-CFG retains runtime control while maintaining competitive generation quality. We summarize our contributions as follows:

- We introduce PP-CFG, which learns separate text transformations conditioned on scale and time to retain adjustable guidance and negative prompting within one visual backbone pass.
- We introduce NegElement-1K, a benchmark of 1,000 matched image and video prompt cases for evaluating target removal through negative prompting.
- We evaluate PP-CFG with DMD on Wan and with both LCM and DMD on Qwen-Image. These experiments support its generality across attention architectures, generation modalities, and distillation objectives.

2 Related Work

Diffusion distillation. Few-step distillation accelerates image and video diffusion models (Luo et al., 2023; Yin et al., 2024a; Wang et al., 2023). For video, DCM separates semantic and detail modeling (Lv et al., 2025), while Self-Forcing trains on inference rollouts (Huang et al., 2025). CausVid learns causal generators from bidirectional teachers (Yin et al., 2025); TMD matches teacher ODE transition kernels (Nie et al., 2026). ASD aligns generations across denoising step counts through adversarial self-distillation (Yang et al., 2025). AVD uses adaptive regression and temporal regularization to mitigate oversaturation and temporal collapse (You et al., 2026). These methods improve sampling efficiency and generation quality. Our focus is preserving guidance-scale and negative-prompt controls after distillation.

Guidance and negative prompting. CFG combines conditional and unconditional predictions using two model evaluations (Ho & Salimans, 2022). Guidance distillation replaces these evaluations with a scale-conditioned student (Meng et al., 2023). P-Guide steers the initial latent prior (Peng & Gao, 2026), while Internal Guidance combines intermediate and final transformer outputs (Zhou et al., 2026). Other work improves guidance through manifold constraints (Chung et al., 2024), projection (Sadat et al., 2025), or restricted intervals (Kynkäänniemi et al., 2024). For attention-space negative prompting, NASA subtracts responses to negative concepts (Nguyen et al., 2025), and Normalized Attention Guidance (NAG) controls the resulting feature magnitude (Chen et al., 2025). PP-CFG combines response subtraction with separate text refinements conditioned on scale and time, retaining both guidance-scale and negative-prompt inputs.

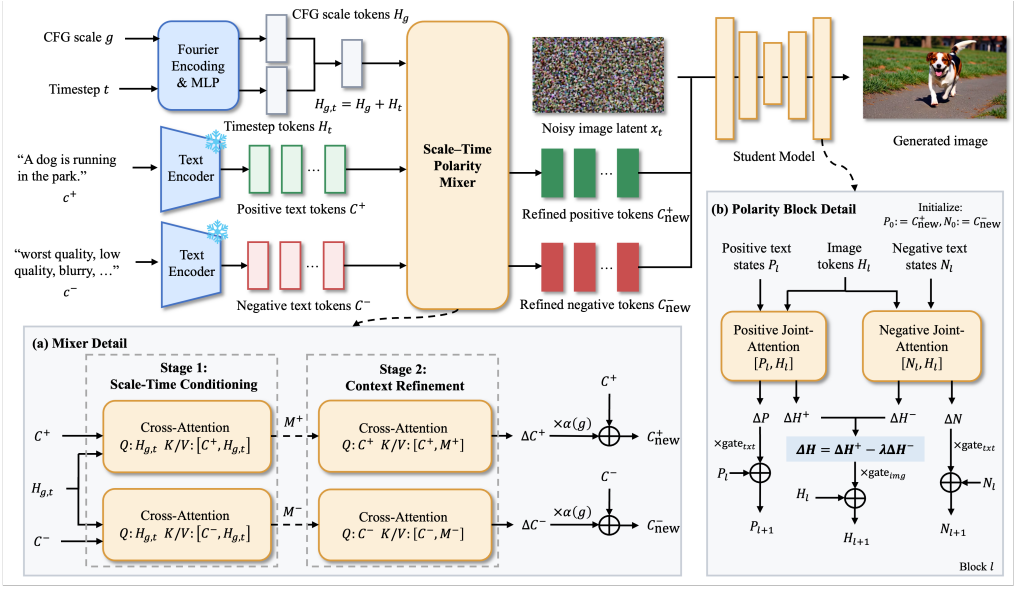

Figure 2: **PP-CFG on Qwen-Image.** (a) The mixer refines each text condition using guidance scale g and denoising time t; internal projections are omitted in the diagram. (b) Joint-attention paths update separate text states and combine their visual responses within one shared stream. Section 3.3 describes this instance and the Wan cross-attention implementation.

3 Method

3.1 Guidance Formulation and Overview

At each denoising step, the model receives a noisy latent x_t at time t, a positive prompt c^+ describing desired content, a negative prompt c^- naming unwanted content (or an empty prompt), and guidance scale g. Let $v_\theta(x_t, t, c)$ denote a backbone prediction under prompt c. Standard CFG combines two such predictions:

$$v_{\text{cfg}}(x_t, t, c^+, c^-; g) = v_\theta(x_t, t, c^-) + g\big(v_\theta(x_t, t, c^+) - v_\theta(x_t, t, c^-)\big). \tag{1}$$

The two evaluations let visual features evolve separately under each prompt. PP-CFG maintains one visual state: its mixer refines the two text conditions using (g, t), and polarity blocks combine their attention responses. This retains a separate negative-prompt input without two full backbone evaluations.

At $g = 1$, Eq. 1 reduces to the positive prediction. PP-CFG does not apply this equation at its output, so its endpoint behavior depends on the mixer and attention combination defined below.

3.2 Scale-Time Polarity Mixer

The mixer learns to adapt each text condition while the two attention paths share visual features. In attention from visual to text tokens, visual queries interact with text keys to determine attention weights, which are used to aggregate text values. Transforming the keys can express some matching variations otherwise produced by branch-specific query transformations; transforming the values also changes the information delivered to visual tokens. Conditioning these refinements on (g, t) allows each prompt's influence to vary with guidance strength and denoising stage.

A frozen text encoder maps $c^\pm$ to token sequences $C^\pm$. Sinusoidal embeddings followed by multilayer perceptrons, denoted E_g and E_t, produce scale and time tokens with matching lengths and widths:

$$H_g = E_g(g), \qquad H_t = E_t(t), \qquad H_{g,t} = H_g + H_t. \tag{2}$$

Their elementwise sum supplies both controls without increasing the number of conditioning tokens.

Two attention stages connect these controls to each prompt. Let $X^{\pm}$ and S be projections of $C^{\pm}$ and $H_{g,t}$ into a common mixer feature space; $[A, B]$ denotes concatenation along the token dimension. Each stage $F_i(Q, KV)$ applies attention followed by a feed-forward transformation, with its first input supplying queries and its second supplying keys and values:

$$M^{\pm} = F_1\big(S, [X^{\pm}, S]\big), \qquad U^{\pm} = F_2\big(X^{\pm}, [X^{\pm}, M^{\pm}]\big). \tag{3}$$

Stage 1 lets scale-time tokens read the corresponding prompt, producing a prompt-specific control representation $M^{\pm}$. Stage 2 passes that information back to the text tokens, producing $U^{\pm}$. The two stages have distinct parameters, each shared across polarities; neither path reads the other prompt. This preserves separate contexts before their visual responses are combined.

Separate output projections W^{+} and W^{-} map the refined features back to the text-context width. The gate $\alpha(g)$ controls the strength of this residual update:

$$\Delta C^{\pm} = W^{\pm} U^{\pm}, \qquad C^{\pm}_{\text{new}} = C^{\pm} + \alpha(g)\Delta C^{\pm}. \tag{4}$$

The scale tokens influence the content of the update, whereas $\alpha(g)$ scales its contribution. Distinct projections allow the two prompt roles to receive different updates despite sharing the attention stages. Initializing $W^{\pm}$ to zero makes the refined contexts equal to the original text representations at initialization. We use a square-root gate $\alpha(g) = \sqrt{(g - g_{\min})/(g_{\max} - g_{\min})}$, where $g_{\min} = 1$ and the configuration-specific $g_{\max}$ is reported in Table 4. This concave mapping makes the gate more responsive at low guidance scales while reducing its sensitivity as g increases.

3.3 Polarity Blocks across Backbones

Each polarity block reads the refined contexts from a shared visual state H_l, where l indexes backbone blocks. Let ΔH_l^{+} and ΔH_l^{-} denote the visual attention responses under the positive and negative conditions. Following NASA (Nguyen et al., 2025), the block combines them before updating the visual stream:

$$\Delta H_l = \Delta H_l^{+} - \lambda \Delta H_l^{-}. \tag{5}$$

Here λ weights the negative attention response; it has a different role from the runtime scale g supplied to the mixer. Subtraction acts on intermediate attention features, with the final guided prediction learned during distillation. The polarity mechanism is enabled by default with a fixed, configuration-specific λ reported in Table 4, but can be disabled at $g = 1$.

Wan cross-attention. Visual tokens supply shared queries, while C^{+}_{new} and C^{-}_{new} supply separate text keys and values. The two calculations share projection weights and are packed along the batch dimension. The text contexts are not updated by the backbone's cross-attention; their visual responses are combined using Eq. 5.

Qwen-Image joint attention. As illustrated in Fig. 2(b), initialize text states $P_0 = C^{+}_{\text{new}}$ and $N_0 = C^{-}_{\text{new}}$. At block l, separate joint-attention paths process $[P_l, H_l]$ and $[N_l, H_l]$. They update their text states independently and combine only their visual responses. The diagram's gate_{txt} and gate_{img} are the backbone's residual multipliers for text and image updates. Both implementations thus retain two prompt paths while advancing one visual stream.

3.4 Training and Inference

PP-CFG can be trained with either latent consistency model (LCM) distillation (Luo et al., 2023) or DMD (Yin et al., 2024b;a). LCM encourages consistent predictions at latent states linked by a teacher denoising trajectory; DMD matches the student's generated distribution to the guided teacher distribution using teacher and critic score estimates. We use DMD for Wan and evaluate both objectives on Qwen-Image, jointly training the student diffusion transformer (DiT) and mixer while keeping the text encoder frozen.

For multi-guidance DMD, we sample g uniformly from $[g_{\min}, g_{\max}]$ and collect the conditions as $\eta_g = (c^{+}, c^{-}, g)$. Let $G_\phi(z, \eta_g)$ denote the complete few-step student sampler with parameters ϕ and initial noise z. Its generated clean latent is $x_0^{\phi} = G_\phi(z, \eta_g)$. DMD perturbs this output at a sampled noise level τ to obtain $y_\tau = a_\tau x_0^{\phi} + \sigma_\tau \epsilon$, where ϵ is independent standard Gaussian noise and a_τ, σ_τ are the noising coefficients. Here τ is the time used to evaluate the distillation loss, whereas t is the student's denoising time; a_τ is unrelated to the mixer gate $\alpha(g)$.

Denote the noised student and guided teacher distributions by $q_\tau^\phi(\cdot \mid \eta_g)$ and $p_\tau^{\mathrm{T}}(\cdot \mid \eta_g)$. At a given guidance setting, DMD aims to reduce their Kullback–Leibler divergence D_{KL}:

$$\mathcal{L}_{\mathrm{DMD}}(\phi; \eta_g) = \mathbb{E}_\tau \left[D_{\mathrm{KL}} \left(q_\tau^\phi(\cdot \mid \eta_g) \parallel p_\tau^{\mathrm{T}}(\cdot \mid \eta_g) \right) \right]. \tag{6}$$

Training over a range of guidance strengths averages this objective over the sampled scales:

$$\mathcal{L}_{\text{MG-DMD}}(\phi) = \mathbb{E}_{g \sim \mathcal{U}(g_{\min}, g_{\max})} \left[\mathcal{L}_{\mathrm{DMD}}(\phi; \eta_g) \right]. \tag{7}$$

DMD optimizes this objective using a difference of scores, where a score is the gradient of a log density with respect to the noised latent y_τ. The teacher provides a score estimate s_{T}, while a critic with parameters ψ estimates the student score s_ψ. The Jacobian $J_\phi(y_\tau) = \partial y_\tau / \partial \phi$ maps this signal to student parameters, and $w(\tau)$ weights noise levels. The student update is

$$\nabla_\phi \mathcal{L}_{\text{MG-DMD}} \approx \mathbb{E}_{z,\tau,\epsilon,\eta_g} \left[w(\tau) J_\phi(y_\tau)^\top \left(s_\psi(y_\tau, \tau, \eta_g) - s_{\mathrm{T}}(y_\tau, \tau, \eta_g) \right) \right]. \tag{8}$$

The teacher and critic use the same g for their guided predictions, and the student receives it through the mixer. Appendix A gives the LCM objective and collects training configurations, including DMD weighting and the critic schedule.

At inference, each step refines $C^\pm$ using (g, t) and applies the polarity blocks to x_t. One visual backbone pass produces the student prediction, avoiding two complete output-space CFG evaluations. The mixer and dual attention paths still add computation beyond an unguided pass.

4 Experiments

4.1 Experimental Setup

Models and evaluation. We evaluate three combinations: Wan2.1-T2V-1.3B with DMD (Wan et al., 2025), and Qwen-Image with LCM or DMD (Wu et al., 2025). VBench measures video quality and semantic alignment, reported as Total, Quality, and Semantic scores (Huang et al., 2024). For images, GenEval tests compositional properties such as object count, color, and position (Ghosh et al., 2023); DPG-Bench evaluates adherence to detailed prompts (Hu et al., 2024). Table 2 gives sampling steps. Appendix A collects training configurations, and Appendix B collects evaluation settings.

Baselines. Video teacher and DMD, LCM, PCM, DCM, rCM, and AVD scores are from You et al. (2026). We run ASD (Yang et al., 2025) with its original configuration except for a chunk size of 1. We evaluate image students in our setup, including LCM and DMD variants, sCM (Lu & Song, 2025), and MeanFlow (Geng et al., 2025). SD denotes step distillation with external CFG; CFG-E+SD additionally embeds the guidance scale in the student. These configurations distinguish external guidance from a learned scale input. Our image variants are labeled Ours (LCM) and Ours (DMD); unless stated otherwise, subsequent image results use Ours (LCM). The step-distilled DMD in the video scale sweep is labeled DMD (Step Distill) in Fig. 5(b).

NegElement-1K. NegElement-1K contains 1,000 template-generated cases balanced across ten categories (100 each). Each case pairs an image prompt containing a target element with a negative prompt naming it, a video prompt for the same scene, and a target-free reference prompt. Annotations identify the target, three scene elements to preserve, and target position, size, and difficulty. We generate four images or one video per case. A Qwen3.5-VL judge checks target presence in each image or sampled video frames. *Removal success* is the fraction of samples judged target absent; preservation annotations do not enter this score. Failed or unparseable judgments count as failures. Appendix B collects the benchmark and judge protocol.

4.2 Main Quality Comparison

Table 2: Main generation results (0–100; higher is better). Bold marks the best listed student per metric. Teachers use two CFG evaluations per step. SD: step distillation; CFG-E: CFG embedding. The image teacher scores are reported by Wu et al. (2025); RL denotes their reinforcement-learning-enhanced model. Video result sources are specified above.

Video: Wan2.1-T2V-1.3B

Method	Steps	Total	Quality	Semantic
Teacher	50	80.13	84.11	64.20
DMD	4	80.66	**85.55**	61.10
LCM	6	72.12	77.62	50.15
PCM	6	73.70	79.36	51.40
DCM	6	73.92	79.27	52.16
rCM	4	80.15	84.80	61.55
AVD	4	81.35	85.27	65.65
ASD	4	80.79	81.01	**79.92**
Ours	4	**81.46**	84.69	68.53

Image: Qwen-Image

Method	Steps	GenEval	DPG
Teacher	50	$87.00/91.00^{\mathrm{RL}}$	88.32
sCM	8	60.00	85.54
LCM	8	61.85	80.69
LCM (SD)	8	81.94	87.02
LCM (CFG-E+SD)	8	81.64	86.90
DMD	8	79.08	86.40
DMD (SD)	8	80.72	86.71
DMD (CFG-E+SD)	8	80.59	86.68
MeanFlow	8	49.00	83.81
Ours (LCM)	8	**82.39**	**87.05**
Ours (DMD)	8	80.22	86.98

Table 2 shows that PP-CFG retains competitive generation quality. On video, it achieves the highest listed student VBench Total (81.46) in four steps, versus 80.66 for DMD, 81.35 for AVD, and 80.13 for the teacher. DMD leads Quality and ASD leads Semantic, so the gains are not uniform across submetrics. On images, Ours (LCM) obtains the best student GenEval and DPG-Bench scores, modestly above LCM (SD). Ours (DMD) improves DPG-Bench over DMD (SD) but not GenEval; the teacher remains higher on both image metrics.

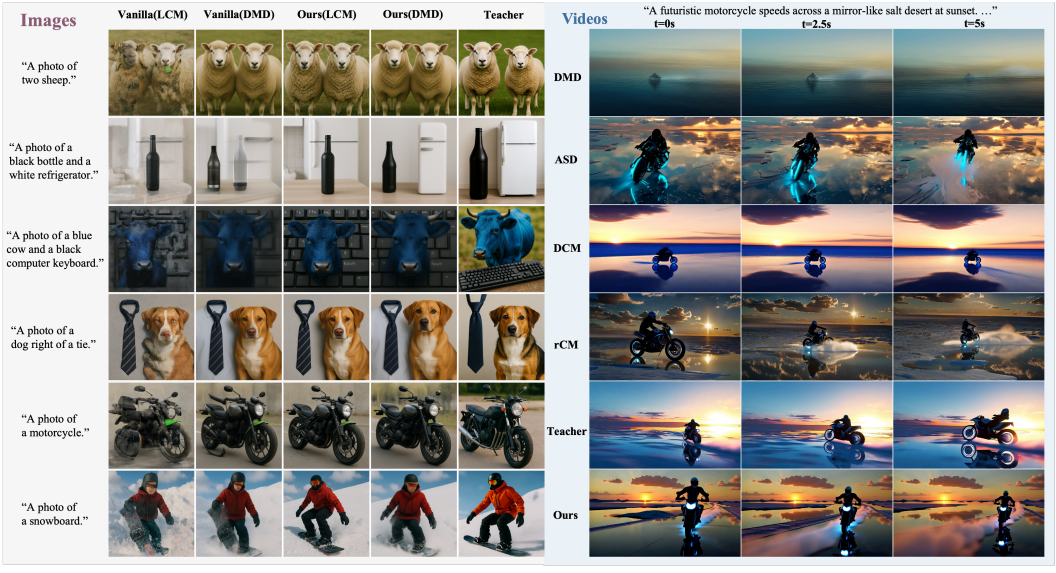

Figure 3: **Image and video generation quality.** Images compare vanilla LCM/DMD, their PP-CFG counterparts, and the Qwen-Image teacher under matched prompts. Videos compare PP-CFG, representative distillation baselines, and Wan2.1-T2V-1.3B, with three frames per video.

Figure 3 complements these metrics with matched-prompt examples. PP-CFG retains clear object attributes in the images and coherent motorcycle and scene structure across the sampled video frames.

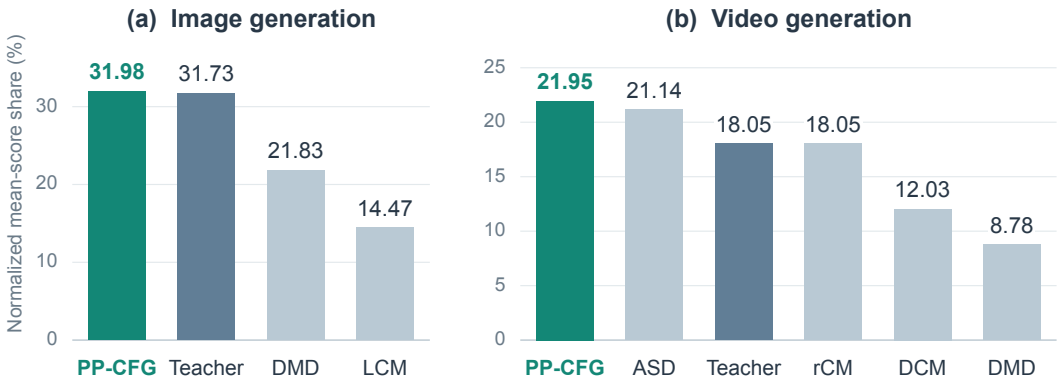

Figure 4: **User study.** Bars show each method's mean score divided by the sum of means within its modality (higher is better). PP-CFG is green; panels use different y-axis ranges.

User study. Ten participants evaluated visual quality and prompt alignment on a shared set of six video prompts and five image prompts. For video, each participant rated all six methods on every prompt using a 1–3 Likert scale (1 = poor, 3 = good); for image, each participant rated all four methods on every prompt using the same scale. Methods were presented in randomized order to each participant. Scores were first averaged across participants and prompts for each method, then converted to percentage shares by dividing each method's mean by the sum of means within the modality. Figure 4 reports these normalized mean-score shares: our image model receives 31.98% versus 31.73% for the teacher; our video model receives 21.95% versus 21.14% for ASD, indicating comparable aggregate ratings among the leading methods. The two panels use different comparison sets and denominators, so comparisons should stay within each modality.

4.3 CFG Adjustability

Figure 5 tests whether the distilled video model responds to changes in guidance scale. In panel (a), PP-CFG retains a recognizable miniature ship and surrounding scene as g increases from 1 to 9, while external CFG applied to step-distilled DMD produces over-exposure and loses detail. The teacher provides a reference for how generation changes with scale.

Panel (b) measures teacher–student alignment using cross-model Fréchet Video Distance (FVD): the Fréchet distance between their video feature distributions at the same guidance scale. We extract features with VideoMAE (Tong et al., 2022); lower values indicate closer distributions in this feature space. PP-CFG has lower distances at every plotted scale from 1 to 7. At scales 3 and 7, its distances are 276 and 373, compared with 387 and 605 for DMD. The smaller increase is consistent with the visual comparison, while the nonzero distances indicate that the teacher distribution is only approximated. Appendix B collects the feature and sampling details.

Panel (c) complements teacher alignment with VBench quality measurements. PP-CFG and the teacher generally improve in Total and Quality above $g = 1$ before leveling off, and their Semantic scores plateau at higher scales. Together, the curves and examples support a usable guidance-scale input in this video setting. Cross-model FVD uses the teacher, rather than real videos, as its reference; VBench supplies the complementary quality assessment. Appendix D provides per-channel latent statistics at $g = 7$. The next experiment separately evaluates negative-prompt control at a fixed scale.

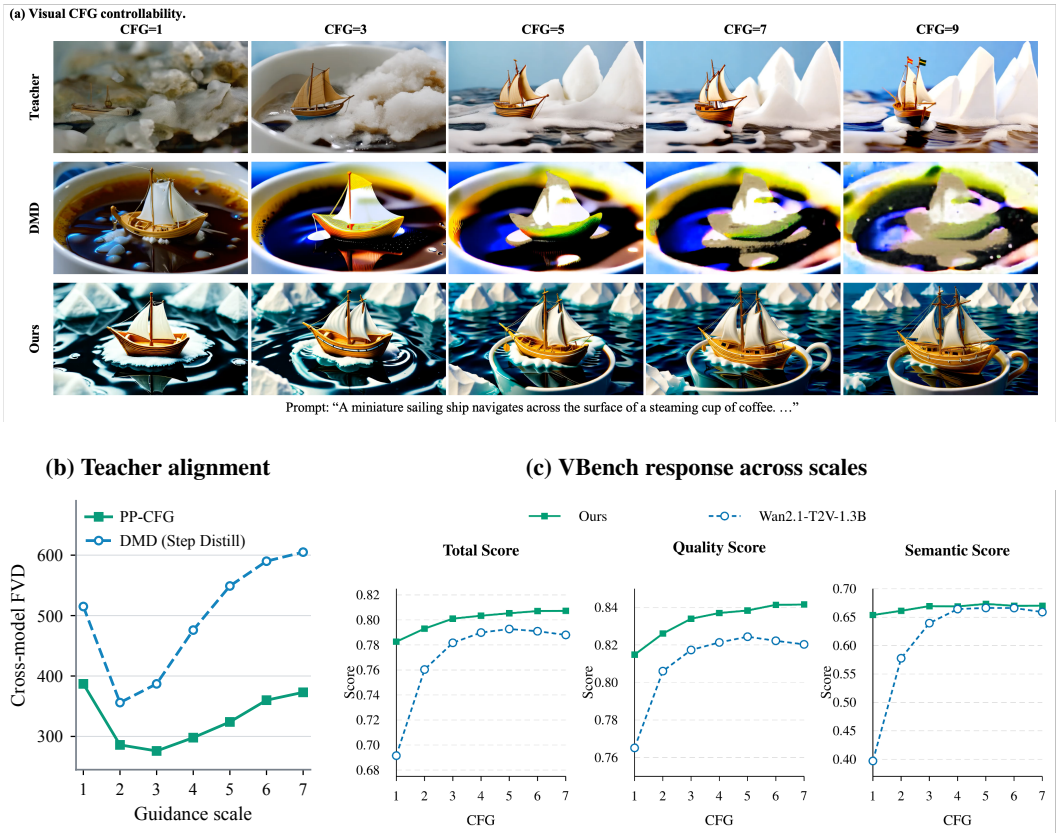

Figure 5: **CFG control and video teacher–student alignment.** (a) First frames across scales; the step-distilled DMD baseline uses external CFG. (b) Reported matched-scale FVD values in VideoMAE feature space (lower is better). (c) VBench responses of PP-CFG and the Wan2.1 teacher on a 0–1 scale. DMD is omitted from (c) because of the degradation in (a).

4.4 Negative-Prompt Control

Figure 6 evaluates target removal on NegElement-1K at CFG 4. Our 8-step image model reaches 12.08% success versus 7.35% for the 8-step Qwen-Image teacher, a gain of 4.73 percentage points over 4,000 samples per method. Our 4-step video model reaches 22.00% versus 10.20% for the 50-step Wan teacher, a gain of 11.80 points over 1,000 samples per method. Both teachers use external CFG; the video comparison is not step-matched.

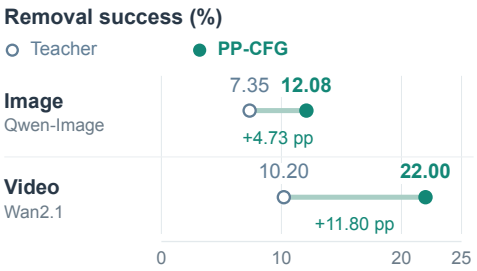

Figure 6: **Removal success.** NegElement-1K rates at CFG 4, judged by Qwen3.5-VL; failed judgments count as failures. Differences are in percentage points (pp).

PP-CFG achieves higher target-absence rates under this negative-prompt protocol in both modalities. The rates compare each model's generated outputs with the requested target; they do not isolate the effect of enabling a negative prompt within the same model.

Figure 7 shows target absence in the sunglasses and umbrella cases and semantic substitution in the floor-lamp and paper-cup cases. These selected outputs illustrate different responses to a specified negative concept.



Figure 7: **Runtime negative-prompt examples.** Columns 1/3 show images; 2/4 show video frames. Targets appear in red in positive prompts, with negative prompts above. Teacher and Ours denote the corresponding models for each modality. Examples show target absence or semantic substitution; green boxes mark substituted objects.

4.5 Ablations

Table 3: VBench ablations on video generation (higher is better). Each variant changes the Scale-Time Polarity Mixer as indicated.

Variant	Change to the mixer	Total	Quality	Semantic
Concat-t tokens	Concatenate scale and time tokens	80.20	84.20	64.20
w/o timestep	Remove time-token injection	80.12	84.14	64.05
w/o mixer	Remove the entire mixer	78.44	81.85	64.78
Full	Retain all components	**81.46**	**84.69**	**68.53**

Table 3 examines generation quality under changes to the mixer. The full design scores highest on all three VBench metrics. Removing time conditioning lowers these scores, while removing the entire mixer further reduces Total and Quality. Concatenating scale and time tokens also underperforms their elementwise sum. These results support the combined design for video quality; the control behavior is evaluated separately by the scale sweep and NegElement-1K.

5 Conclusion

PP-CFG retains adjustable guidance and negative-prompt inputs in few-step generators by refining the two text conditions separately and combining their attention responses within one visual stream. Experiments on Wan2.1 and Qwen-Image support its generality across cross-attention and joint attention, video and image generation, and LCM and DMD distillation. The method achieves competitive generation quality, lower video teacher–student distances across guidance scales than step-distilled DMD, and higher target-removal rates on NegElement-1K than the corresponding teachers. These results support preserving distinct prompt roles as an architectural approach to retaining runtime controls after distillation.

AI USE STATEMENT

We used generative AI tools to assist with language editing and polishing, literature retrieval and discovery, and writing research code. The authors reviewed the AI-assisted material incorporated into this work and take responsibility for the final manuscript, citations, code, and results.

ETHICS STATEMENT

All participants provided informed consent for the user study and received no compensation. No institutional ethics approval or formal exemption was obtained.

REPRODUCIBILITY STATEMENT

Section 3 describes the conditioning architecture and its implementations in Wan and Qwen-Image. Section 3.4 gives the DMD objective and update, Appendix A collects the implementation and training configurations, and Appendix B describes evaluation protocols. The user study is reported in Section 4.2. We plan to publicly release the code, model checkpoints, NegElement-1K prompts, and evaluation scripts.

REFERENCES

Dar-Yen Chen, Hmrishav Bandyopadhyay, Kai Zou, and Yi-Zhe Song. Normalized attention guidance: Universal negative guidance for diffusion models. *arXiv preprint arXiv:2505.21179*, 2025.

Jiaxiang Cheng, Bing Ma, Xuhua Ren, Hongyi Henry Jin, Kai Yu, Peng Zhang, Wenyue Li, Yuan Zhou, Tianxiang Zheng, and Qinglin Lu. Phased one-step adversarial equilibrium for video diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 3237–3245, 2026.

Hyungjin Chung, Jeongsol Kim, Geon Yeong Park, Hyelin Nam, and Jong Chul Ye. Cfg++: Manifold-constrained classifier free guidance for diffusion models. *arXiv preprint arXiv:2406.08070*, 2024.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J. Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. *arXiv preprint arXiv:2505.13447*, 2025.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. *Advances in Neural Information Processing Systems*, 36: 52132–52152, 2023.

Wenqi Guo and Shan Du. Vsf: Simple, efficient, and effective negative guidance in few-step image generation models by value sign flip. *arXiv preprint arXiv:2508.10931*, 2025.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. ELLA: Equip diffusion models with LLM for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

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. *Advances in Neural Information Processing Systems*, 37:122458–122483, 2024.

Cheng Lu and Yang Song. Simplifying, stabilizing and scaling continuous-time consistency models. In *International Conference on Learning Representations*, 2025.

Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023.

Zhengyao Lv, Chenyang Si, Tianlin Pan, Zhaoxi Chen, Kwan-Yee K Wong, Yu Qiao, and Ziwei Liu. DCM: Dual-expert consistency model for efficient and high-quality video generation. *arXiv preprint arXiv:2506.03123*, 2025.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14297–14306, 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*, pp. 18004–18013, 2025.

Weili Nie, Julius Berner, Nanye Ma, Chao Liu, Saining Xie, and Arash Vahdat. Transition matching distillation for fast video generation. *arXiv preprint arXiv:2601.09881*, 2026.

Xin Peng and Ang Gao. P-Guide: Parameter-efficient prior steering for single-pass CFG inference. *arXiv preprint arXiv:2605.06124*, 2026.

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

Shitong Shao, Hongwei Yi, Hanzhong Guo, Tian Ye, Daquan Zhou, Michael Lingelbach, Zhiqiang Xu, and Zeke Xie. Magicdistillation: Weak-to-strong video distillation for large-scale few-step synthesis. *arXiv preprint arXiv:2503.13319*, 2025.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, 2022.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Fu-Yun Wang, Zhaoyang Huang, Alexander William Bergman, Dazhong Shen, Peng Gao, Michael Lingelbach, Keqiang Sun, Weikang Bian, Guanglu Song, Yu Liu, Xiaogang Wang, and Hongsheng Li. Phased consistency models. *arXiv preprint arXiv:2405.18407*, 2024.

Xiang Wang, Shiwei Zhang, Han Zhang, Yu Liu, Yingya Zhang, Changxin Gao, and Nong Sang. Videolcm: Video latent consistency model. *arXiv preprint arXiv:2312.09109*, 2023.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Yongqi Yang, Huayang Huang, Xu Peng, Xiaobin Hu, Donghao Luo, Jiangning Zhang, Chengjie Wang, and Yu Wu. Towards one-step causal video generation via adversarial self-distillation. *arXiv preprint arXiv:2511.01419*, 2025.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. *Advances in neural information processing systems*, 37:47455–47487, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6613–6623, 2024b.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22963–22974, 2025.

Yuyang You, Yongzhi Li, Jiahui Li, Yadong Mu, Quan Chen, and Peng Jiang. Adaptive video distillation: Mitigating oversaturation and temporal collapse in few-step generation. *arXiv preprint arXiv:2603.21864*, 2026.

Haiyu Zhang, Xinyuan Chen, Yaohui Wang, Xihui Liu, Yunhong Wang, and Yu Qiao. Accvideo: Accelerating video diffusion model with synthetic dataset. *arXiv preprint arXiv:2503.19462*, 2025.

Kaiwen Zheng, Yuji Wang, Qianli Ma, Huayu Chen, Jintao Zhang, Yogesh Balaji, Jianfei Chen, Ming-Yu Liu, Jun Zhu, and Qinsheng Zhang. Large scale diffusion distillation via score-regularized continuous-time consistency. *arXiv preprint arXiv:2510.08431*, 2025.

Xingyu Zhou, Qifan Li, Xiaobin Hu, Hai Chen, and Shuhang Gu. Guiding a diffusion transformer with the internal dynamics of itself. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11536–11545, 2026.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. *arXiv preprint arXiv:2602.02214*, 2026.

A Implementation Details

All three configurations train the student backbone and Scale-Time Polarity Mixer jointly with a frozen text encoder. The two mixer stages share weights across polarities, with separate zero-initialized output projections. Wan uses text cross-attention; Qwen-Image updates separate text states through joint attention. Table 4 collects the run-specific settings.

Table 4: Implementation and training configurations. All models use AdamW with a generator learning rate of 10^{-6} and 50 warm-up steps, gradient checkpointing, bf16 mixed precision, and FSDP full sharding. The Scale-Time Polarity Mixer uses 4 tokens, hidden dim 768, 4 heads, and $\sqrt{\cdot}$ alpha gating with zero-initialized output projections.

Setting	Wan–DMD	Qwen–LCM	Qwen–DMD
Architecture			
Backbone	Wan-1.3B (30 layers)	Qwen-Image	Qwen-Image
Attention type	Cross	Joint	Joint
Backbone dim / heads	1536 / 12	3072 / 24	3072 / 24
Mixer tokens / dim / heads	4 / 768 / 4	4 / 768 / 4	4 / 768 / 4
Optimization			
Optimizer	AdamW (β=0, 0.999)	AdamW (β=0.9, 0.999)	AdamW (β=0, 0.999)
Generator LR	10^{-6}	10^{-6}	10^{-6}
Critic LR	4×10^{-7}	—	4×10^{-7}
Batch size (per GPU $\times$ GPUs)	1×8	1×8	1×8
Max grad norm	10.0	1.0	1.0
Precision	bf16	bf16	bf16
Distillation			
Objective	DMD	LCM (consistency)	DMD
Training steps	2 000	1 000	1 000
Critic update ratio	5:1	—	5:1
EMA decay / start step	0.99 / 200	—	—
DMD grad normalize	✓	—	✓
Teacher solver	—	Euler (50 steps)	—
LCM distance d	—	Pseudo-Huber (c=10^{-3})	—
Guidance & negative prompts			
CFG scale range $[g_{\min}, g_{\max}]$	[1, 9]	[1, 5]	[1, 5]
Negative prompt	pool of precomputed embeds	empty string	empty string
Neg. subtraction λ	1	0.001	0.001
CFG norm rescale	—	✓	✓
Data & evaluation			
Training data	internal captioned video pairs	internal-Image-10k	internal-Image-10k
Resolution	480×832, 81 f	512×512	512×512
Sampling steps	4	8	8

LCM training. LCM encourages consistent clean-latent predictions along a guided teacher trajectory (Luo et al., 2023). Let $f_\phi(x_t, t, \eta_g)$ be the student's estimate under $\eta_g = (c^+, c^-, g)$ from Section 3.4. For noise times $t' < t$, a teacher solver follows the guided probability-flow ordinary differential equation (PF-ODE) to construct $\hat{x}_{t'} = \Psi_{\mathrm{T}}(x_t, t, t'; \eta_g)$. The target network f_{ϕ^-} uses an exponential moving average (EMA) of student parameters. The consistency objective is

$$\mathcal{L}_{\mathrm{LCM}}(\phi) = \mathbb{E}_{x_t,t,t',\eta_g}\left[d\big(f_\phi(x_t, t, \eta_g), \mathrm{sg}\big[f_{\phi^-}(\hat{x}_{t'}, t', \eta_g)\big]\big)\right], \tag{9}$$

where d measures the discrepancy between latent predictions and sg stops gradients through the target. Both predictions use the same prompts and guidance scale. Matching them teaches the student to reach a consistent endpoint from different noise levels. PP-CFG supplies these conditions through the mixer and polarity blocks while retaining the consistency objective.

Step distillation. Step distillation(SD) trains the student to reproduce the teacher's guided output without explicitly distilling the guidance mechanism itself. Concretely, both the teacher and the student apply CFG during training: the student's CFG prediction is supervised to match the teacher's CFG prediction across the consistency trajectory.

CFG-Embedding step distillation. This variant additionally distils the guidance mechanism into a single forward pass. A lightweight CFG-embedding network, whose architecture mirrors the timestep embedding, maps the scalar guidance scale to a vector that is added to the timestep-embedding output. The final linear layer is zero-initialised so that the embedding has no effect at the start of training. During training, the student receives only the positive prompt and performs a single forward pass conditioned on guidance scale; its output is supervised to match the teacher's CFG prediction, thereby distilling both the denoising trajectory and the guidance signal simultaneously. At inference the student requires only one forward pass while retaining the ability to modulate generation quality via guidance scale.

B Evaluation Protocols

Generation quality and baselines. VBench, GenEval, and DPG-Bench results in Table 2 are reported on a 0–100 scale; Fig. 5(c) uses 0–1. Video teacher and baseline results, except ASD and PP-CFG, are quoted from You et al. (2026); ASD is our run with chunk size 1. The image teacher scores match Tables 3–4 of Wu et al. (2025). Our local evaluations follow the corresponding cited benchmark protocols and use the same evaluation settings as the quoted results.

Text-to-Video (T2V) Evaluation. We evaluate T2V generation quality using **VBench v1** Huang et al. (2024), which measures 16 complementary dimensions. All video baseline scores are quoted from You et al. (2026), **except** ASD, which we reproduce locally with `chunk=1`. For the VBench results in Table 2, methods that support CFG are evaluated at $g = 9$.

Our T2V model is distilled from Wan2.1-T2V-1.3B via DMD and evaluated at the 2 k-step checkpoint. We use separate learning rates for the generator (1×10^{-6}) and the critic (4×10^{-7}), both optimised with AdamW ($\beta_1 = 0.0,\ \beta_2 = 0.999$, weight decay 0.01). The denoising step list is $[1000,\ 750,\ 500,\ 250]$ with timestep shift 5.0. We apply EMA with a decay of 0.99 starting at step 200, and update the generator every 5 critic steps. The maximum gradient norm is 10.0. All training is conducted in bf16 mixed precision with gradient checkpointing. For the Scale-Token Polarity Mixer, we use 4 scale tokens with a hidden dimension of 768, 4 attention heads, and α mode $\sqrt{\cdot}$. Videos are generated at 480×832 resolution with 21 latent frames.

Text-to-Image (T2I) Evaluation. Image generation quality is evaluated on **GenEval** Ghosh et al. (2023) (553 prompts covering single-object, two-object, counting, colour, position, and colour-attribute categories) and **DPG-Bench** Hu et al. (2024) (1 065 unique prompts with detailed paragraph descriptions). For each prompt we generate 4 images with seed 42 at resolution 512×512. For the GenEval and DPG-Bench results in Table 2, methods that support CFG are evaluated at $g = 4$. Teacher scores are quoted from Wu et al. (2025).

Cross-model FVD. The two sets compared in Fig. 5(b) are teacher and student videos generated at the same guidance scale, using VideoMAE features. Only matched-scale distances are reported. We use the SSv2-finetuned VideoMAE ViT-g hybrid checkpoint `vit_g_hybrid_pt_1200e_ssv2_ft.pth`. Its final spatiotemporal tokens are mean-pooled and normalized to give one 1,408-dimensional feature per video. We sample one contiguous 16-frame window per video (sampling seed 20260917), divide pixel values by 255, resize the short side to 128 pixels, center-crop to 128×128, and bilinearly resize to 224×224 without channel normalization. At each $g = 1,\ldots,7$, we evaluate 1,000 videos per model under the same VBench prompt conditions; generation seeds are unavailable, so matching is by condition only. FVD uses float64 feature statistics, maximum-likelihood covariances, and the real part of SciPy's covariance-product matrix square root, without regularization or symmetrization. These covariances are rank-deficient because $N = 1{,}000 < 1{,}408$.

NegElement-1K. Each model generates 4,000 images or 1,000 videos from the 1,000 cases at guidance scale 4. Target-absence judgments are aggregated over generated samples, with failed judgments counted as failures.

Benchmark construction. NegElement-1K contains 1,000 negative-element removal cases generated deterministically by the template script `build_neg_element_1k.py`. The benchmark is balanced across ten categories (100 cases each): TABLETOP, KITCHEN-FOOD, INDOOR-FURNITURE, OFFICE-STUDY, URBAN-STREET, PARK-NATURE, VEHICLE-TRANSPORT, SPORTS-TOYS, PERSON-ACCESSORY, and TEXT-LOGO-WATERMARK. For the first eight (object) categories, each case is formed by crossing one of ten scenes with one of ten target objects; a spatial relation is chosen round-robin from ten options (e.g. *to the left of*, *behind*, *on top of*). The PERSON-ACCESSORY category crosses ten portrait scenes with ten wearable items, and TEXT-LOGO-WATERMARK crosses ten product surfaces with ten text/logo targets. Difficulty labels (`easy`/`medium`/`hard`; 433/397/170) and target-size labels (`small`/`medium`/`large`; 320/540/140) are assigned by rule.

Prompt templates. Each case stores four prompts:

- **positive_prompt** (contains the target): "`{scene}, with {target_present} {relation} {anchor}, {style}.`"
- **negative_prompt** (for classifier-free guidance): "`{target}, {generic}`"
- **target_free_prompt** (target-absent reference): "`{scene}, with empty space {relation} {anchor}, {style}.`"
- **video_prompt**: "`A short static video shot of {scene}, with {target_present} {relation} {anchor}, {video_motion}.`"

Each case additionally records a `preserve` list of scene elements that must remain intact and a `target_free_prompt` that describes the same composition with the target replaced by empty space, serving as a target-free reference.

Annotated example. Case `neg_element_000000` (category TABLETOP, difficulty `easy`, target size `medium`):

- **target**: *red apple*
- **positive_prompt**: "A clean wooden table holding a white ceramic plate and a clear glass of water, with one red apple to the left of the white ceramic plate, realistic lighting, clean composition."
- **negative_prompt**: "red apple, apple"
- **target_free_prompt**: "... with empty space to the left of the white ceramic plate ..."
- **remove**: [*red apple*]; **preserve**: [*wooden table*, *white ceramic plate*, *clear glass of water*]

Generation protocol. Each model generates at guidance scale 4. For images, 4,000 samples are produced (4 per case $\times$ 1,000 cases). For video, 1,000 samples are produced (1 per case).

VLM judge. Target-absence judgments are obtained from **Qwen3.5-VL-397B-A17B** (FP8 quantisation; model identifier `Qwen/Qwen3.5-397B-A17B-FP8`). Decoding uses greedy sampling (`temperature = 0`).

Image judge prompt:

```
You are a strict visual evaluator.
Task: Determine whether the target object is visible in the
image.
Target object(s): {target}
Rules:
- Answer target_present=true if the target object, a synonym,
or a visually similar replacement is visible.
- Answer target_present=false only if the target object is
absent.
- If uncertain, choose target_present=true.
Return strictly valid JSON:
{"target_present": true, "confidence": 0.0, "reason":
"short reason"}
```

The video judge prompt is identical except it asks whether the target "appears in any sampled frame of the video."

Frame sampling (video): Up to 161 frames are decoded; 8 frames are selected at equally-spaced indices. Each frame is resized so that the longer side is at most 768 px and encoded as a JPEG data-URL (quality = 88). Images are resized to a maximum side of 1,024 px (JPEG quality = 90).

Output parsing: A regex `\{.*\}` extracts JSON from the model response. The boolean field `target_present` determines the verdict.

Decision aggregation: Each sample is judged independently. A case is counted as a *removal success* at the sample level when `target_present = false`. For case-level reporting, a case is successful when $\geq$50% of its samples are successes (majority vote with threshold 0.5).

Failure handling: If JSON parsing fails, the sample defaults to `target_present = true` (conservative; counts as removal failure). API-level exceptions are similarly treated as failures with the reason field recording the error.

C Efficiency and Resource Reporting

PP-CFG advances one visual stream per denoising step, with additional mixer and positive/negative attention computation. Counting backbone evaluations therefore describes the execution structure but does not quantify wall-clock or memory cost. All experiments run on 8 commercially available GPUs with bf16 precision.

Training cost. **Wan–DMD** trains for 2 000 steps at $\sim$45 s/step using 81-frame videos at 480×832 resolution, totaling $\sim$25 h wall-clock ($\sim$200 GPU-hours); each step includes teacher, critic, and generator forward and backward passes, data loading, and gradient synchronization across 8 GPUs. **Qwen–LCM** trains for 1 000 steps at $\sim$4 s/step on 512×512 images, completing in $\sim$1.2 h ($\sim$10 GPU-hours). **Qwen–DMD** trains for 1 000 steps at $\sim$5 s/step, completing in $\sim$1.4 h ($\sim$11 GPU-hours); the slightly higher per-step cost reflects the additional critic forward pass. Per-step times include all forward and backward passes (teacher, critic where applicable, and student), data loading, and gradient synchronization; checkpoint saving is excluded.

Inference latency and memory. End-to-end inference (text encoding, denoising loop, and VAE decoding) is measured on a single GPU with batch size 1. For **text-to-image** (Qwen-Image, 512×512): our 8-step PP-CFG model generates an image in **2.9 s**, while the 50-step teacher with true CFG requires **21.2 s**—a $\mathbf{7.3}\times$ speed-up. Peak GPU memory is 55.5 GB for both configurations (dominated by the text encoder and VAE). For **text-to-video** (Wan 1.3B, $81 \times 480 \times 832$): our 4-step model produces an 81-frame clip in **15.0 s** versus **280.7 s** for the 50-step teacher with CFG—an $\mathbf{18.7}\times$ speed-up; peak memory is 20.7 GB.

PP-CFG per-step overhead. The polarity mixer adds $\sim$23 M parameters to Wan 1.3B (1.7% of the backbone) and $\sim$37 M to Qwen-Image (0.6%). Measured per-step forward latency increases by 71 ms on Wan and 137 ms on Qwen-Image;

D Additional Distribution Analysis

Figure 8 complements the matched-scale FVD analysis in Fig. 5(b). At CFG 7, PP-CFG's per-channel latent distributions generally resemble the teacher's modes and concentration more closely than DMD's, which are broader or shifted in several channels. These marginal statistics describe each channel separately and do not characterize the full joint latent distribution.

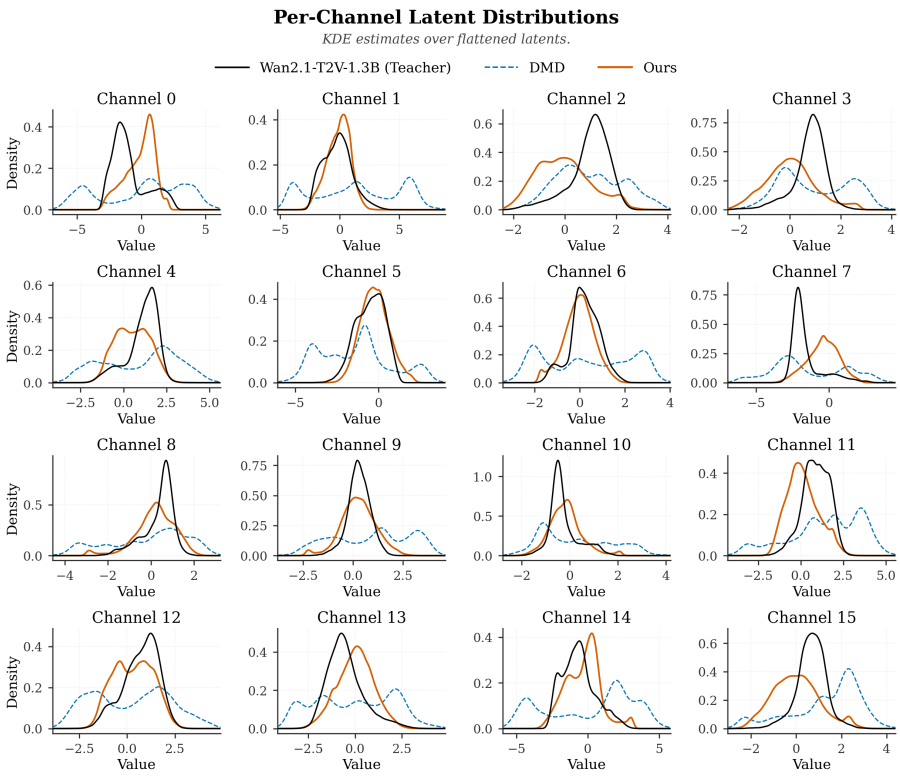

Figure 8: **Per-channel latent distributions at CFG 7.** For each of the 16 latent channels, we flatten the values and estimate densities using Gaussian kernels. The Wan2.1 teacher, DMD, and PP-CFG distributions provide a marginal comparison complementary to cross-model FVD.

E Full Runtime-Capability Comparison

Table 5 expands the grouped settings in Table 1. Entries describe the listed configurations: a cross indicates that a capability is not provided in that configuration, rather than an impossibility for the underlying method. Fixed-guidance students absorb guided predictions into their weights; the SD variants apply external CFG, and CFG-E+SD variants learn a scale input. An adjustable scale alone does not provide a separate runtime negative-prompt input.

Table 5: Runtime capabilities of the considered distilled-generator configurations. Adj.: adjustable guidance scale; Neg.: runtime negative prompts; 1-NFE: one visual backbone evaluation per denoising step.

Method	**Adj.**	**Neg.**	**1-NFE**
POSE (Cheng et al., 2026)	✗	✗	✓
DCM (Lv et al., 2025)	✗	✗	✓
MagicDistillation (Shao et al., 2025)	✗	✗	✓
Self-Forcing (Huang et al., 2025)	✗	✗	✓
Causal Forcing (Zhu et al., 2026)	✗	✗	✓
TMD (Nie et al., 2026)	✗	✗	✓
rCM (Zheng et al., 2025)	✗	✗	✓
AccVideo (Zhang et al., 2025)	✗	✗	✓
Standard DMD (Yin et al., 2024a)	✗	✗	✓
LCM (Luo et al., 2023)	✗	✗	✓
sCM (Lu & Song, 2025)	✗	✗	✓
AVD (You et al., 2026)	✗	✗	✓
ASD (Yang et al., 2025)	✗	✗	✓
LCM (SD)	✓	✗	✗
DMD (SD)	✓	✗	✗
LCM (CFG-E+SD)	✓	✗	✓
DMD (CFG-E+SD)	✓	✗	✓
PCM (Wang et al., 2024)	✓	✓	✗
PP-CFG	✓	✓	✓

F LIMITATIONS

Experiments on Wan and Qwen-Image support PP-CFG's generality across cross-attention and joint attention, video and image generation, and LCM and DMD distillation. This evidence comes from the evaluated model and training combinations. Extending the same design to additional model families and generation tasks is a natural direction for future work.