MASS FORCING: ON-POLICY MASS-AWARE DISTILLATION FOR AUTOREGRESSIVE VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Distribution Matching Distillation (DMD) has emerged as a promising paradigm for real-time video synthesis, distilling pretrained diffusion models into few-step generators. However, its reverse-KL objective is mode-seeking and can collapse the student's conditional video distribution onto a narrow subset of modes, leading to reduced sample diversity, diminished temporal dynamics, and color oversaturation. Although existing methods primarily mitigate this issue via coverage-oriented student initialization, we observe that such initial coverage is transient: across three different initializations, the fraction of dynamic videos drops sharply during subsequent DMD training, indicating that the unchanged mode-seeking objective, rather than the initialization, drives this degradation. Motivated by this training-time degradation, we propose Mass Forcing, a probability-mass-aware framework that incorporates coverage-oriented supervision throughout on-policy DMD training via complementary local and set-level objectives. First, at the local level, Mass Forcing estimates relative teacher-to-student density ratios across rollouts sharing a student-generated prefix through latent-space score integration and uses these estimates to rebalance conditional DMD updates, amplifying the training influence of rollouts underrepresented relative to the teacher. Second, recognizing that mismatched mode probabilities do not necessarily induce strong local score corrections, Mass Forcing aligns generated rollout sets with a reweighted reference distribution defined by the same mass estimates, providing cross-rollout supervision beyond individual gradient reweighting. Experiments on MovieGen and VBench demonstrate that Mass Forcing improves video fidelity and sustains temporal dynamics throughout training, achieving the highest Total scores among compared methods and a Dynamic Degree of 83% after 1,000 generator updates, versus 12-32% for the baselines, and introducing no additional inference-time computational overhead.

1 INTRODUCTION

Recent video diffusion models synthesize high-fidelity videos with fine visual detail and complex motion (Brooks et al., 2024; Yang et al., 2025; Polyak et al., 2024; Kong et al., 2024; Ma et al., 2025; Wan et al., 2025). Most of these models, however, are bidirectional: they jointly denoise all frames at every step, so no frame can be shown until the whole clip is generated, and this latency precludes real-time streaming and interaction. Autoregressive (AR) video models (Yan et al., 2021; Kondratyuk et al., 2024; Wang et al., 2024; Deng et al., 2025; Jin et al., 2025; Sand.ai et al., 2025) instead generate frames or chunks sequentially, each conditioned on the preceding context, which reduces time-to-first-frame and enables continuous, interactive generation. Yet AR diffusion models still require many denoising steps per chunk, keeping them far from real time.

Few-step AR distillation closes this gap by matching a causal student, initialized from a pretrained video diffusion model, to the teacher's output distribution via Distribution Matching Distillation (DMD) (Yin et al., 2024b;a; 2025); Self Forcing (Huang et al., 2025) performs this matching *on-policy*, on the student's own rollouts. However, DMD minimizes a *mode-seeking* reverse Kullback–Leibler (KL) divergence, which lets the student concentrate on a few teacher-supported modes, reducing sample diversity and temporal dynamics (Xu et al., 2025; You et al., 2026). Existing remedies for AR video mainly improve initialization, either with an AR teacher that bridges the bidirectional-to-causal gap, as in Causal Forcing (Zhu et al., 2026), or with a mode-covering start

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

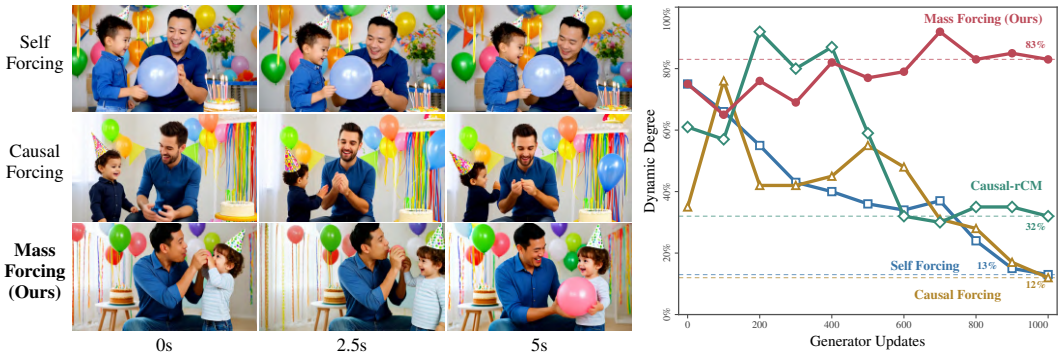


Figure 1: **Mass Forcing sustains temporal dynamics during DMD training. Left:** Frames generated for *"...a father blowing up a balloon for his excited child..."*. Self Forcing stays nearly static and Causal Forcing never inflates the balloon, whereas Mass Forcing shows the full inflation. **Right:** VBench Dynamic Degree over training. Despite different initializations, Self Forcing (Huang et al., 2025), Causal Forcing (Zhu et al., 2026), and Causal-rCM (Zheng et al., 2026a) decline sharply, while Mass Forcing stays between 65% and 92%. Dashed lines mark values at update 1,000; each point is evaluated on 100 MovieGen prompts with 10 seeds each.

from consistency distillation, as in Causal-rCM (Zheng et al., 2026a), leaving the subsequent mode-seeking objective unchanged.

Our key observation is that the mode coverage gained at initialization is transient under continued on-policy DMD. Figure 1 (right) tracks VBench Dynamic Degree (DD) (Huang et al., 2024), the fraction of generated videos classified as dynamic, during DMD training from three different initializations. Between 100 and 1,000 generator updates, DD falls from 66% to 13% for Self Forcing and from 76% to 12% for Causal Forcing, while Causal-rCM, despite its mode-covering initialization, drops from a peak of 92% to 32%. Different initializations thus change the early trajectory but not the late-training decline, suggesting that the objective, rather than the initialization, drives the collapse. This behavior is consistent with the well-known *zero-forcing* property of reverse KL (Bishop, 2007): because the divergence is evaluated only on student samples, a teacher-supported mode that the student stops generating incurs almost no penalty. We conjecture that motion-rich continuations are especially exposed in AR video generation: they spread probability across many diverse modes, whereas static continuations concentrate it, so on-policy training can drift toward static content at little cost. Meanwhile, Self Forcing's Aesthetic and Imaging Quality improve overall as training proceeds (Appendix D.4), revealing an early-stopping trade-off between visual quality and temporal dynamics. Resolving this dilemma requires supervision that actively preserves coverage throughout optimization, so that continued distillation improves visual quality without sacrificing dynamics.

Such supervision must look beyond individual samples: the DMD update checks whether each rollout is plausible under the teacher, but not whether the student generates it as often as the teacher would, so rollouts from underrepresented modes count no more than others and lost coverage is not actively restored. We therefore introduce *Mass Forcing*, a probability-mass-aware framework for on-policy video DMD. It samples multiple continuations from a shared student-generated prefix and estimates their relative teacher-to-student density ratios by integrating the score difference between the teacher and fake-score networks along latent-space paths. Weighting DMD by such ratios yields more mode-covering objectives (Xu et al., 2025); unlike prior work, which obtains the ratios from an auxiliary GAN discriminator in image distillation, our estimator reuses the networks already present in DMD.

The same estimates drive two complementary objectives. Locally, they reweight the conditional distribution-matching term of DMD toward underrepresented rollouts, while the classifier-free guidance (CFG) (Ho & Salimans, 2022) term stays uniformly weighted, so that guidance strength is unchanged. Such reweighting, however, can only amplify an existing score correction, which nearly vanishes when well-separated modes differ mainly in probability. At the set level, we therefore use maximum mean discrepancy (MMD) (Gretton et al., 2012; Li et al., 2015) to align the rollout set with a detached reference distribution reweighted by the same estimates, pulling rollouts toward those with higher estimated teacher support. Both objectives reuse the same on-policy rollouts, requiring no teacher-generated rollouts during training.

Experiments on MovieGen (Polyak et al., 2024) and VBench (Huang et al., 2024) show that Mass Forcing achieves the highest Total and Quality scores among all compared methods, raising the Total score over the strongest baseline, Causal Forcing, from 82.51 to 83.38 on MovieGen and from 84.14 to 84.98 on VBench. Starting from the same ODE initialization as Self Forcing, Mass Forcing keeps DD between 65% and 92% throughout training and ends at 83% (Figure 1, right). All additional supervision is confined to training: the standard four-step AR generator and its IID Gaussian sampling are unchanged, and inference incurs no extra latency. Our main contributions are as follows:

- We show that the mode coverage provided by improved initialization is transient under on-policy DMD: across three initializations, temporal dynamics decline sharply as training continues, indicating that the objective, not the starting point, drives the collapse.
- We propose Mass Forcing, which estimates teacher-relative probability mass for rollouts sharing a student-generated prefix via score line integrals over existing DMD networks, and uses these estimates in two complementary objectives: mass-weighted conditional score distillation and mass-targeted rollout-set matching.
- Experiments on MovieGen and VBench show improved overall generation quality and sustained temporal dynamics over competitive few-step AR baselines, with no change to the inference pipeline or its cost.

2 METHOD

Both Mass Forcing objectives are driven by one quantity: the estimated teacher-relative mass of each continuation sampled from a shared student prefix (Figure 2). Section 2.2 estimates these masses, Section 2.3 uses them to reweight conditional score corrections, and Section 2.4 uses them to define a reference for rollout-set matching. Derivations and implementation details appear in Appendices B and C, respectively.

2.1 PRELIMINARIES

For a prompt c, an autoregressive generator represents a latent video $V = [x^{(1)}, \dots, x^{(B)}]$ as B chunks $x^{(b)} \in \mathbb{R}^d$ with distribution

$$Q_\theta(V \mid c) = \prod_{b=1}^{B} q_\theta(x^{(b)} \mid u^{(b)}, c), \qquad u^{(b)} = [x^{(1)}, \dots, x^{(b-1)}], \tag{1}$$

where brackets denote temporal concatenation and $u^{(1)} = \emptyset$. Each chunk is generated by $x^{(b)} = G_\theta(z^{(b)}; u^{(b)}, c)$, where $z^{(b)}$ denotes sampler randomness, including initial and intermediate Gaussian draws, and θ denotes generator parameters. Self Forcing (Huang et al., 2025) trains on these student-generated histories and applies holistic DMD to the resulting videos. Unlike this full-video data path, each Mass Forcing update uniformly samples a chunk index b, generates its prefix $u = u^{(b)}$ with the current student, detaches the prefix and its KV cache, and applies DMD to K independently sampled continuations of that chunk.

DMD matches the student to a frozen teacher through noise-averaged reverse KL (Yin et al., 2024b;a). For our conditional analysis, fix a sampled prefix u and write $x = G_\theta(z; u, c)$ and $y = \alpha_t x + \sigma_t \epsilon$ at noise level t, with $\epsilon \sim \mathcal{N}(0, I_d)$, $\alpha_t = 1 - \sigma_t$, and $0 < \sigma_t \leq 1$. Let $q_{\theta,t}, p_t$ be the noisy student and teacher continuation densities conditioned on (u, c), with scores $s_{q,t} = \nabla_y \log q_{\theta,t}$ and $s_{p,t} = \nabla_y \log p_t$. For $\mathcal{L}_{\mathrm{RKL}} = \mathbb{E}_t[\omega(t) D_{\mathrm{KL}}(q_{\theta,t} \| p_t)]$, the gradient is

$$\nabla_\theta \mathcal{L}_{\mathrm{RKL}} = \mathbb{E}_{t,z,\epsilon}\left[\omega(t)\alpha_t J_\theta^\top (s_{q,t}(y) - s_{p,t}(y))\right], \tag{2}$$

where $J_\theta = \partial G_\theta / \partial \theta$ and $\omega(t) \geq 0$ weights noise levels. We use loss-gradient signs and hold the prefix fixed within each update; Appendix B.1 relates this analysis to full-video DMD. Improved initialization and on-policy contexts leave this objective unchanged. Because Eq. (2) averages over student samples, each mode's sampling frequency follows the student's probability mass rather than the teacher's. Underrepresented teacher-supported modes therefore contribute fewer sampled correction terms, motivating explicit teacher-relative mass reweighting.

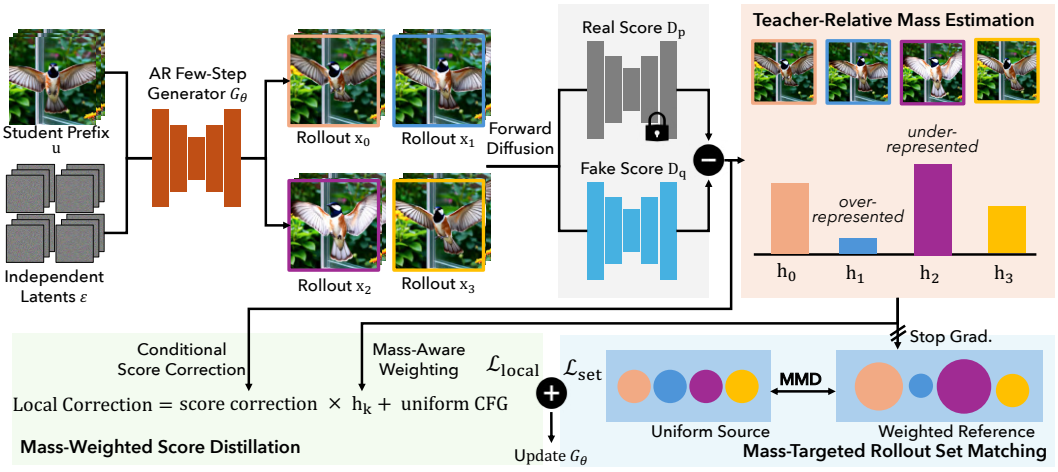


Figure 2: **Mass Forcing training.** From a shared student-generated prefix, Mass Forcing estimates teacher-relative weights over independently sampled rollouts. The weights modulate conditional DMD corrections while leaving CFG uniform and define a detached reference for rollout-set matching. Training combines $\mathcal{L}_{\text{local}}$ and $\mathcal{L}_{\text{set}}$ to update the generator.

2.2 Estimating Teacher-Relative Mass

Ideal mass weighting. The ratio $r_t(y) = p_t(y)/q_{\theta,t}(y)$ measures teacher support relative to student probability. The divergence-gradient formulation (Xu et al., 2025) motivates weighting Eq. (2) by

$$h_t^*(y) = 1 - \lambda + \lambda r_t(y), \qquad q_{\theta,t}(y)h_t^*(y) = (1-\lambda)q_{\theta,t}(y) + \lambda p_t(y), \tag{3}$$

where $\lambda \in [0,1]$ interpolates between reverse- and forward-KL gradients for exact scores and ratios. Figure 3(a,b) illustrates this change in the effective measure: underrepresented rollouts gain influence without changing the inference sampler.

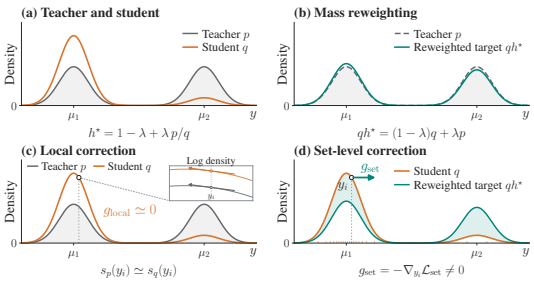


Figure 3: **Mass-aware local and set-level supervision.** (a,b) Reweighting adjusts the effective mass allocation. (c,d) Set-level matching complements weak local corrections.

Shared-prefix queries. We collect K detached rollouts $\bar{x}_k = G_\theta(z_k; u, c)$, $k = 1, \ldots, K$, using independent sampler randomness z_k. At a shared noise level, we form $\bar{y}_k = \alpha_t \bar{x}_k + \sigma_t \epsilon$ and $u_t = \alpha_t u + \sigma_t \epsilon_u$. The Gaussian corruptions ϵ and ϵ_u are independent of each other but fixed across the group; overbars mark detached samples. The bidirectional teacher and fake denoiser process $[u_t, y]$ and supply current-chunk clean-latent predictions $D_p(y)$ and $D_q(y)$, with (u_t, t, c) suppressed. The teacher is frozen, while the fake denoiser is trained on current student rollouts to track the evolving generator distribution. Their difference yields the estimated score field $\Delta_t(y) = \alpha_t(D_p(y) - D_q(y))/\sigma_t^2$. At fixed u_t, a joint score's current-chunk component is a conditional score, so this field estimates $\Delta_t^* = \nabla_y \log \tilde{r}_t$, where $\tilde{r}_t$ is the teacher-to-student ratio conditioned on (u_t, c). This is a noisy-prefix proxy for the ideal clean-prefix ratio; the residual parameterization and conditional-score identity are detailed in Appendix B.3.

Relative log ratios. To recover relative mass from score differences, we integrate between rollouts. For medoid anchor a and straight path $\zeta_k(\xi) = a + \xi(\bar{y}_k - a)$, $\xi \in [0,1]$, the exact field satisfies

$$\log \tilde{r}_t(\bar{y}_k) - \log \tilde{r}_t(a) = \int_0^1 \Delta_t^*(\zeta_k(\xi))^\top (\bar{y}_k - a)\, \mathrm{d}\xi. \tag{4}$$

We approximate this latent-space integral using the learned field and three-point Simpson quadrature, obtaining ℓ_k. Shared corruption preserves the noise variance along each path.

Stabilized weights. With $\bar{\ell} = K^{-1}\sum_j \ell_j$, we form $v_k = \mathrm{clip}((\ell_k - \bar{\ell})/\tau, -\kappa, \kappa)$, where $\tau > 0$ is the temperature and $\kappa > 0$ bounds the scaled logits. Centering removes the common anchor offset, and the normalized weights are

$$\widehat{r}_k = K\frac{\exp(v_k)}{\sum_j \exp(v_j)}, \qquad h_k = 1 + \lambda_{\mathrm{eff}}(\widehat{r}_k - 1). \tag{5}$$

Here $\lambda_{\mathrm{eff}} \in [0, \lambda_{\max}]$ is a group-shared mixture coefficient incorporating the training ramp and reliability gates (Appendix C.2). The detached weights have mean one, reallocating rather than increasing the total rollout weight. h_k mirrors h_t^* in Eq. (3), with $\widehat{r}_k$ replacing r_t and λ_{eff} replacing λ. These finite-group estimates provide a coverage-oriented surrogate, not exact forward-KL optimization; Appendix B.4 specifies the stabilization and approximation boundaries.

2.3 Mass-Weighted Score Distillation

Since the mass estimates target the conditional teacher, we separate conditional distribution matching from classifier-free guidance (CFG) (Liu et al., 2026a) and weight only the former. Let $D_{\mathrm{cfg}} = D_{\mathrm{uncond}} + w_{\mathrm{cfg}}(D_p - D_{\mathrm{uncond}})$, where D_{uncond} omits text conditioning and w_{cfg} is the guidance scale. In clean-latent coordinates, we use the following normalized DMD surrogate:

$$g_k = \frac{1}{n_k}\left[h_k \underbrace{(D_{q,k} - D_{p,k})}_{\text{conditional matching}} + \underbrace{(D_{p,k} - D_{\mathrm{cfg},k})}_{\text{guidance}}\right], \tag{6}$$

where $D_{\bullet,k} = D_\bullet(\bar{y}_k)$ and $n_k > 0$ is a detached DMD normalizer. For $h_k = 1$, the numerator recovers $D_{q,k} - D_{\mathrm{cfg},k}$; otherwise, mass correction changes conditional matching without rescaling guidance. Since g_k is computed on detached rollouts, we regenerate each rollout using the same randomness z_k and prefix cache before updating θ, obtaining a differentiable copy x_k numerically equal to $\bar{x}_k$. We minimize the detached-target loss

$$\mathcal{L}_{\mathrm{local}} = \frac{1}{2Kd}\sum_{k=1}^{K} \|x_k - \mathrm{sg}(x_k - g_k)\|_2^2, \tag{7}$$

where sg stops gradients, giving $\nabla_{x_k}\mathcal{L}_{\mathrm{local}} = g_k/(Kd)$ (Appendix B.5).

2.4 Mass-Targeted Rollout-Set Matching

When teacher and student share well-separated component shapes but assign them different probabilities, their within-mode scores nearly coincide (Figure 3(c)), making the conditional correction at each rollout small. The line integral can still detect the imbalance by accumulating score differences along paths between modes. The weights h_k can thus be informative even when the conditional corrections they multiply are negligible. We therefore use these weights in a second objective that does not rely on local score differences. With $y_k = \alpha_t x_k + \sigma_t \epsilon$, the uniformly weighted source and reweighted reference are

$$\mu_\theta = \frac{1}{K}\sum_k \delta_{y_k}, \qquad \nu = \sum_k \pi_k \delta_{\bar{y}_k}, \qquad \pi_k = \mathrm{sg}(h_k/K), \tag{8}$$

where δ_y is a point mass. The measures initially share sample locations but assign different masses; reference locations and weights remain detached. We match them with maximum mean discrepancy (MMD) (Gretton et al., 2012; Li et al., 2015) and combine both objectives:

$$\mathcal{L}_{\mathrm{set}} = \mathrm{MMD}^2_{\mathcal{K}_\beta}(\mu_\theta, \nu), \tag{9}$$

$$\mathcal{L}_{\mathrm{MF}} = \mathcal{L}_{\mathrm{local}} + \gamma\mathcal{L}_{\mathrm{set}}, \tag{10}$$

where $\mathcal{K}_\beta$ is a Gaussian kernel with detached scale $\beta > 0$ and $\gamma \geq 0$ controls set-level supervision. At replay initialization, $y_k = \bar{y}_k$ and the source derivative is

$$\nabla_{y_i} \mathcal{L}_{\text{set}}|_{y=\bar{y}} = \frac{2}{K^2} \sum_j (1 - h_j) \nabla_1 \mathcal{K}_\beta(\bar{y}_i, \bar{y}_j), \tag{11}$$

where ∇_1 differentiates the first kernel argument, holding the references, weights, and kernel scale fixed. Under gradient descent, each Gaussian-kernel interaction attracts a rollout toward a reference with $h_j > 1$ or repels it from one with $h_j < 1$. Their sum can remain nonzero even where the local conditional score correction vanishes (Figure 3(d)); Appendix B.6 gives the derivation.

Training and inference. Generator updates hold prefixes, score predictions, weights, and references fixed, while the fake denoiser is trained separately on unweighted student clips (Appendix B.7). With $\lambda_{\text{eff}} = 0$ and $\gamma = 0$, this procedure reduces to unweighted conditional-chunk DMD, not the holistic Self Forcing data path. Algorithm 1 in Appendix C.4 summarizes the training loop, with one generator update followed by five fake-denoiser updates per iteration. Both objectives reuse the same on-policy rollouts without teacher-generated continuations. All added computation is training-only, preserving standard four-step AR inference with IID Gaussian noise.

3 Experiments

We evaluate whether coverage-oriented supervision improves video generation quality and sustains temporal dynamics during continued DMD training. We first compare Mass Forcing with existing video generators, then examine training trajectories and the contributions of local mass reweighting and set-level matching.

3.1 Experimental Setup

Model and training. We build on the Self Forcing implementation (Huang et al., 2025), using a causal Wan2.1-T2V-1.3B generator, a frozen bidirectional Wan2.1-T2V-14B teacher, and a bidirectional 1.3B residual network for fake-score estimation (Wan et al., 2025). Mass Forcing is applied during DMD training from the ODE-initialized generator, rather than as post-training of a converged DMD checkpoint. Following Self Forcing, we train on a filtered subset of VidProM (Wang & Yang, 2024) containing 248,221 text prompts. The default configuration uses $K = 4$ rollouts per prefix, temperature $\tau = 2$, logit bound $\kappa = 2$, and maximum mixture coefficient $\lambda_{\max} = 0.25$. The set-level reference uses the mixed weights h_k/K in Eq. (8). Optimizer settings, rollout sampling, and mass-activation rules are detailed in Appendix C.

Benchmarks and metrics. We evaluate on 1,003 MovieGen (Polyak et al., 2024) prompts and 946 VBench (Huang et al., 2024) prompts, with all videos generated at 480p and assessed using the standard VBench quality and semantic evaluators. Table 1 reports aggregate Quality, Semantic, and Total scores on the evaluator's normalized scale; Appendix D reports the individual dimension scores before aggregate-score normalization, multiplied by 100 for readability. MovieGen does not provide the metadata required for VBench semantic evaluation, so we construct it using Qwen3-VL-32B-Instruct. We exclude Spatial Relationship because no prompts satisfying VBench's criteria for this dimension were identified in the evaluated MovieGen set. All compared methods use the same metadata and eight-dimensional protocol, retaining VBench's per-dimension normalization and relative weights while renormalizing the retained weights to sum to one (Appendix D.1). We report Dynamic Degree (DD) separately to assess the fraction of videos classified as dynamic, alongside aggregate quality rather than treating motion alone as sufficient evidence of generation quality.

Comparisons and inference. The comparisons cover bidirectional generators, AR generators, and distilled AR generators, as grouped in Table 1. The closest baselines are CausVid, Self Forcing, Causal Forcing, and Causal-rCM, which allow us to compare with distribution matching, on-policy training, and improved initialization strategies. For each baseline, we use its officially released checkpoint and native inference pipeline. Mass Forcing retains four-step AR generation with IID Gaussian noise; neither score estimation nor rollout-set matching is used at inference. Throughput (FPS) and latency are measured on a single NVIDIA H100 GPU with batch size 1 and averaged after

Table 1: Quantitative comparisons with existing methods. Our method consistently improves upon the Self-Forcing baseline across all evaluation metrics while maintaining real-time inference speed. Best results are in **Bold** and second best are underlined.

Model	Throughput ↑ (FPS)	Latency ↓ (s)	MovieGen Evaluation ↑				VBench Evaluation ↑			
			Total	Quality	Semantic	Dynamic	Total	Quality	Semantic	Dynamic
Bidirectional Video Diffusion Models										
LTX-Video (HaCohen et al., 2024)	8.98	13.5	80.40	82.44	72.21	31.90	81.71	83.78	73.42	43.06
Wan2.1 (Wan et al., 2025)	0.24	330	81.47	82.73	76.41	59.02	83.95	85.17	79.04	62.50
AVD (You et al., 2026)	4.61	17.55	80.17	81.54	74.69	61.11	82.07	83.51	76.32	73.61
Autoregressive Video Diffusion Models										
NOVA (Deng et al., 2025)	0.88	4.1	77.32	78.36	73.14	50.45	79.19	79.96	76.11	44.44
Pyramid Flow (Jin et al., 2025)	6.70	2.5	80.11	81.07	76.24	34.40	82.05	83.51	76.21	59.72
SkyReels-V2 (Chen et al., 2025)	0.49	112	80.67	82.51	73.33	50.25	82.73	84.86	74.19	63.89
MAGI-1 (Sand.ai et al., 2025)	0.19	282	81.04	81.81	77.95	22.13	80.42	81.38	76.59	16.67
Distilled Autoregressive Video Models										
CausVid (Yin et al., 2025)	17.0	0.69	80.50	81.72	75.62	57.33	82.56	83.92	77.11	65.28
Self Forcing (Huang et al., 2025)	17.0	0.69	81.75	82.77	77.70	63.41	83.49	84.51	79.40	61.11
Causal Forcing (Zhu et al., 2026)	17.0	0.69	82.51	83.28	**79.42**	77.57	84.14	85.22	79.83	86.11
Causal-rCM (Zheng et al., 2026a)	17.0	0.69	81.68	82.80	77.21	48.26	83.85	85.10	78.84	69.44
Mass Forcing (Ours)	17.0	0.69	**83.38**	**84.53**	78.78	**83.35**	**84.98**	**86.17**	**80.23**	**88.89**

model loading, excluding VAE decoding. For AR methods, the first chunk is excluded from timing, so the reported latency measures subsequent-chunk generation rather than time to first frame.

3.2 Evaluation Results

Quantitative comparison. Mass Forcing achieves the highest Total and Quality scores among the compared methods on both benchmarks (Table 1). On MovieGen, it obtains 83.38 Total and 84.53 Quality, exceeding the strongest baseline, Causal Forcing, by 0.87 and 1.25 points, respectively. On VBench, it obtains 84.98 Total, 86.17 Quality, and 80.23 Semantic, improving over Causal Forcing by 0.84, 0.95, and 0.40 points. Relative to Self Forcing, the Quality gains are 1.76 points on MovieGen and 1.66 points on VBench. These improvements do not imply dominance on every metric: MovieGen Semantic remains below Causal Forcing (78.78 versus 79.42), while the overall Total score improves.

Quality and dynamics. The dimension-level results show that the aggregate gains coexist with substantial motion rather than depending on nearly static outputs. Compared with Self Forcing, Mass Forcing increases DD from 63.41% to 83.35% on MovieGen and from 61.11% to 88.89% on VBench, while improving Aesthetic Quality from 63.88 to 65.92 and from 65.90 to 67.08, respectively. Mass Forcing also exceeds Causal Forcing in DD on both benchmarks while achieving higher aggregate Quality. Some individual metrics also trade off: for example, VBench Imaging Quality is lower than Self Forcing's (66.93 versus 69.53). The complete results in Tables 4–7 make these differences explicit.

Qualitative comparison. Figure 4 complements Figure 1 with two further comparisons of Self Forcing, Causal Forcing, and Mass Forcing at matched timestamps. In Figure 4(a), Mass Forcing depicts the rabbit eating the moon, whereas Self Forcing and Causal Forcing do not visibly realize the requested eating action. Figure 4(b) highlights scene-level movement: Mass Forcing produces a changing view of the frozen, CN-tower-inspired landscape while preserving layered ice detail. Together, these examples illustrate prompt-aligned action and scene movement alongside visual detail, complementing the aggregate quality and dynamics measurements. Additional five-frame comparisons with full prompts appear in Appendix E.

3.3 Behavior During Continued Training

Temporal dynamics. Figure 1 compares DD over 1,000 generator updates using 100 MovieGen prompts with 10 seeds per prompt at each evaluation point. Between updates 100 and 1,000, Self

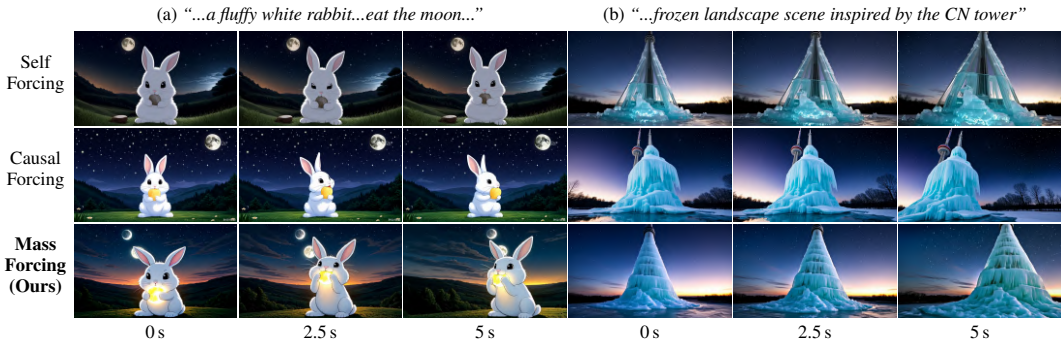


Figure 4: **Qualitative comparison with Self Forcing and Causal Forcing.** Two showcase prompts illustrate character-pose changes and scene evolution. Each group displays frames at 0, 2.5, and 5 seconds, with methods aligned by row. Mass Forcing preserves detailed appearance as the character pose and viewpoint change.

Forcing declines from 66% to 13% DD and Causal Forcing from 76% to 12%, whereas Mass Forcing remains between 65% and 92% and finishes at 83%. The trajectories are not monotonic, but they distinguish a sustained dynamic regime from the pronounced late-training degradation of the baselines. Beyond these training trajectories, the full-benchmark evaluations in Table 1 show that Mass Forcing improves DD over Self Forcing from 63.41% to 83.35% on MovieGen and from 61.11% to 88.89% on VBench. Figures 6 and 7 in Appendix E provide further qualitative evidence that Mass Forcing produces more dynamic videos than state-of-the-art baselines.

3.4 Ablation Studies

Table 2 examines guidance decoupling, set-level supervision, and mass-weighting design choices on MovieGen and VBench, using the Ours results from Table 1 as the reference. Dimension-level results and video counts appear in Appendix D.3, with a visual comparison in Appendix F.

Table 2: Ablation results on MovieGen and VBench.

Variant	MovieGen ↑		VBench ↑	
	Qual.	Sem.	Qual.	Sem.
Ours	**84.53**	**78.78**	**86.17**	**80.23**
w/o CFG decoupling	83.42	75.24	84.77	80.16
w/o MMD	83.35	<u>78.59</u>	85.19	78.27
Shuffled-ratio control	83.22	77.48	84.68	79.89
$K = 2$	83.43	77.90	85.29	77.48
Endpoint-only	83.59	77.63	<u>85.34</u>	<u>80.18</u>
$\lambda_{\max} = 0.5$	83.22	76.97	84.25	80.00
$\lambda_{\max} = 1.0$	<u>83.67</u>	78.38	85.21	79.17

Qual.: Quality; Sem.: Semantic.
w/o CFG decoupling also omits MMD.

Guidance decoupling. The w/o CFG decoupling variant omits both decoupling and MMD; w/o MMD retains decoupling. With MMD absent in both, the decoupled variant has higher DD: 65.00% versus 61.81% on MovieGen and 77.78% versus 65.28% on VBench. Quality changes little on MovieGen (83.35 versus 83.42) and improves on VBench (85.19 versus 84.77). This comparison supports decoupling for sustaining dynamics; the gap to Ours reflects both components.

Set-level supervision. The w/o MMD variant removes the set-level term by setting $\gamma = 0$, retaining local mass-weighted score distillation. Removing MMD reduces Quality from 84.53 to 83.35 on MovieGen and from 86.17 to 85.19 on VBench. DD also decreases from 83.35% to 65.00% and from 88.89% to 77.78%, respectively, supporting the contribution of set-level matching to both quality and dynamics. The loss of dynamics despite retaining local mass reweighting supports the need for interactions across rollouts, rather than relying solely on rescaled pointwise corrections.

Weight assignment. The shuffled-ratio control permutes the ratio weights among rollouts within each prefix group, preserving the group weight distribution while breaking the correspondence between a rollout and its estimated teacher-relative support. Quality decreases to 83.22 on MovieGen and 84.68 on VBench, drops of 1.31 and 1.49 points from the reference. MovieGen Semantic also

decreases from 78.78 to 77.48, supporting the importance of rollout-specific weight assignment for both quality and semantic alignment. DD also falls to 59.62% and 69.44%, linking correct weight assignment to sustained motion rather than only improvements in aggregate quality. Since shuffling preserves the weight values, this comparison suggests that the benefit depends on which rollouts receive additional influence, not just the presence of nonuniform weights.

Rollout count and integration. Reducing the rollout count from $K = 4$ to $K = 2$ yields Quality scores of 83.43 on MovieGen and 85.29 on VBench, retaining much of the reference configuration's quality while trailing the larger rollout group. The $K = 2$ model retains high DD on VBench (87.50%) but has lower Aesthetic Quality (66.13 versus 67.08), suggesting that additional rollout alternatives benefit visual quality beyond preserving motion. Replacing Simpson integration with an endpoint-only estimate yields Quality scores of 83.59 and 85.34, respectively, also below the reference. Endpoint-only weighting discards information accumulated between the anchor and each rollout; its lower scores favor using the path, rather than only its endpoint, to compare relative mass.

Mixture strength. Increasing $\lambda_{\max}$ from 0.25 to 0.5 gives Quality scores of 83.22 on MovieGen and 84.25 on VBench; setting it to 1.0 gives 83.67 and 85.21, respectively. Both settings remain below the default in Quality and Semantic on both benchmarks, favoring the moderate mixture strength used in our default configuration. Stronger reweighting therefore does not translate into larger gains: both higher coefficients also lower DD, suggesting that preserving a uniform student component helps balance motion and quality under finite-group mass estimates.

4 Related Work

Bidirectional video generation. Bidirectional video diffusion models jointly denoise entire clips to capture visual detail and temporal coherence (Yang et al., 2025; Kong et al., 2024; Wan et al., 2025). Full temporal context helps coordinate appearance and motion across frames, but repeated clip-wide denoising makes high-resolution generation computationally expensive. LTX-Video addresses this cost through latent compression (HaCohen et al., 2024), while rCM and Adaptive Video Distillation combine distillation with regularization to balance efficiency and generation quality (Zheng et al., 2026b; You et al., 2026). However, clip-level processing still limits low-latency streaming.

Autoregressive video generation and distillation. Autoregressive models support streaming by conditioning generation on preceding context, but AR diffusion models retain iterative denoising costs (Deng et al., 2025; Jin et al., 2025; Chen et al., 2025; Sand.ai et al., 2025). CausVid enables few-step generation by matching a causal student's distribution to a bidirectional teacher's through DMD (Yin et al., 2025). Self Forcing reduces exposure bias using student-generated histories (Huang et al., 2025), while Causal Forcing improves ODE initialization with an AR teacher (Zhu et al., 2026) and Causal-rCM broadens initial coverage through consistency distillation (Zheng et al., 2026a). Further extensions improve long-video training and context handling (Yang et al., 2026; Liu et al., 2026b; Chen et al., 2026). These DMD-based pipelines retain the mode-seeking reverse-KL objective; Mass Forcing instead introduces coverage-oriented supervision throughout on-policy optimization. See Appendix A for an extended discussion.

5 Conclusions

We introduced Mass Forcing, a probability-mass-aware framework that provides coverage-oriented supervision throughout on-policy DMD training for few-step autoregressive video generation. Rather than relying on initialization alone, Mass Forcing makes continued distillation responsive to teacher-relative mass imbalances through complementary local and set-level objectives. Estimated density ratios rebalance conditional score corrections while keeping CFG uniformly weighted, and define a rollout-set reference that provides distributional feedback beyond pointwise gradient reweighting. Experiments on MovieGen and VBench demonstrate improved overall generation quality, while training trajectories show sustained temporal dynamics during continued optimization. These benefits require no changes to the standard four-step autoregressive inference pipeline and incur no additional inference-time computational overhead.

AI USE STATEMENT

In this work, we used generative AI tools to polish the writing of the paper, that is, to improve the grammar, wording, flow, and readability of text written by the authors, including rephrasing and condensing existing paragraphs and figure captions. Additionally, we used Qwen3-VL-32B-Instruct to construct the semantic evaluation metadata for the MovieGen prompts, as described in Appendix D.1. The authors manually reviewed every AI-assisted edit, checking that the revised text preserves the intended technical meaning and that all claims, equations, and reported numbers remain consistent with our derivations and experimental results. We take responsibility for the final content of this work, including AI-assisted text and evaluation metadata.

ETHICS STATEMENT

Mass Forcing aims to sustain temporal dynamics and diversity in real-time autoregressive video generation. As with any capable video generator, models trained with our method could be misused to create misleading, deceptive, or harmful content, including non-consensual depictions of real people, and real-time generation lowers the cost of producing such content at scale. Our method is a training-time distillation objective that relies on a pretrained teacher and student-generated videos rather than additional real video data; the distilled student may therefore inherit the biases and safety limitations of the teacher. We distill only from publicly released models and use them in accordance with their licenses. All evaluations use publicly available benchmark prompts (MovieGen and VBench), and no human subjects, crowdsourced annotations, or personal data were involved. We encourage any deployment of such models to include safeguards such as content filtering, provenance metadata, and watermarking of generated videos.

REPRODUCIBILITY STATEMENT

We have made several efforts to ensure reproducibility. Section 2 specifies the Mass Forcing objectives, and Algorithm 1 in Appendix C.4 gives the complete training loop, including rollout sampling, mass estimation, the local and set-level losses, and generator and critic updates. Appendix B provides the mathematical derivations and states their underlying assumptions. Appendix C lists the models and their initialization, the prefix and continuation sampling scheme, score-query timesteps, mass-estimation hyperparameters and activation gates, optimizer settings, batch sizes, training duration, and random seeds. Section 3.1 and Appendix D describe the evaluation benchmarks and metrics and report per-dimension results. All models, codebases, and benchmarks we build on (Wan2.1, Self Forcing, MovieGen, and VBench) are publicly available. We will release our code and trained checkpoints upon publication.

REFERENCES

Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer, 2007.

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, 2023.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *NeurIPS*, 2024a.

Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, et al. Skyreels-v2: Infinite-length film generative model. In *arXiv*, 2025.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *CVPR*, 2024b.

Shuo Chen, Cong Wei, Sun Sun, Ping Nie, Kai Zhou, Ge Zhang, Ming-Hsuan Yang, and Wenhu Chen. Context forcing: Consistent autoregressive video generation with long context. In *ICML*, 2026.

Haoge Deng, Ting Pan, Haiwen Diao, Zhengxiong Luo, Yufeng Cui, Huchuan Lu, Shiguang Shan, Yonggang Qi, and Xinlong Wang. Autoregressive video generation without vector quantization. In *ICLR*, 2025.

Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *JMLR*, 2012.

Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. In *arXiv*, 2024.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *arXiv*, 2022.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *NeurIPS*, 2022.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *ICLR*, 2023.

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

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 *CVPR*, 2024.

Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. In *ICLR*, 2025.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, et al. Videopoet: A large language model for zero-shot video generation. In *ICML*, 2024.

Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. In *arXiv*, 2024.

Yujia Li, Kevin Swersky, and Rich Zemel. Generative moment matching networks. In *ICML*, 2015.

Shanchuan Lin, Ceyuan Yang, Hao He, Jianwen Jiang, Yuxi Ren, Xin Xia, Yang Zhao, Xuefeng Xiao, and Lu Jiang. Autoregressive adversarial post-training for real-time interactive video generation. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *NeurIPS*, 2025.

Dongyang Liu, Peng Gao, David Liu, Ruoyi Du, Zhen Li, Qilong Wu, Xin Jin, Sihan Cao, Shifeng Zhang, Steven HOI, et al. Decoupled dmd: Cfg augmentation as the spear, distribution matching as the shield. In *ICLR*, 2026a.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. In *ICLR*, 2026b.

Guoqing Ma, Haoyang Huang, Kun Yan, Liangyu Chen, Nan Duan, Shengming Yin, Changyi Wan, Ranchen Ming, Xiaoniu Song, Xing Chen, et al. Step-video-t2v technical report: The practice, challenges, and future of video foundation model. In *arXiv*, 2025.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. In *arXiv*, 2024.

Sand.ai, Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, WQ Zhang, Weifeng Luo, et al. Magi-1: Autoregressive video generation at scale. In *arXiv*, 2025.

Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. In *ICLR*, 2023.

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. In *arXiv*, 2025.

Wenhao Wang and Yi Yang. VidProM: A million-scale real prompt-gallery dataset for text-to-video diffusion models. In *NeurIPS*, 2024.

Yuqing Wang, Tianwei Xiong, Daquan Zhou, Zhijie Lin, Yang Zhao, Bingyi Kang, Jiashi Feng, and Xihui Liu. Loong: Generating minute-level long videos with autoregressive language models. In *arXiv*, 2024.

Yilun Xu, Weili Nie, and Arash Vahdat. One-step diffusion models with f-divergence distribution matching. In *arXiv*, 2025.

Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers. In *arXiv*, 2021.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Ying-Cong Chen, Yao Lu, Song Han, and Yukang Chen. Longlive: Real-time interactive long video generation. In *ICLR*, 2026.

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *ICLR*, 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. In *NeurIPS*, 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 *CVPR*, 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 *CVPR*, 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. In *CVPR*, 2026.

Kaiwen Zheng, Guande He, Min Zhao, Jintao Zhang, Huayu Chen, Jianfei Chen, Chen-Hsuan Lin, Ming-Yu Liu, Jun Zhu, and Qianli Ma. Causal-rcm: A unified teacher-forcing and self-forcing open recipe for autoregressive diffusion distillation in streaming video generation and interactive world models. In *arXiv*, 2026a.

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. In *ICLR*, 2026b.

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. In *ICML*, 2026.

APPENDIX CONTENTS

A Extended Related Work **14**

A.1 Bidirectional Video Diffusion Models 14

A.2 Autoregressive Video Generation and Distillation 14

B Mathematical Derivations of Mass Forcing **15**

B.1 Conditioning, Notation, and Scope 15

B.2 From Distribution Matching to Mass-Weighted Gradients 16

B.3 From Denoising Predictions to Conditional Density Ratios 17

B.4 Path Integration, Shared Noise, and Normalized Weights 18

B.5 Target-Aligned Local Score Distillation 19

B.6 Why Rollout-Set Matching Complements Local Corrections 20

B.7 Combined Generator Gradient and Fake-Score Training 22

C Implementation Details **24**

C.1 Models and On-Policy Rollout Sampling 24

C.2 Mass Estimation and Activation 25

C.3 Replay, Optimization, and Evaluation 25

C.4 Training Algorithm 27

D Detailed Experimental Results **28**

D.1 MovieGen 28

D.2 VBench 32

D.3 Ablation Studies 33

D.4 Visual Quality During DMD Training 35

E Extended Qualitative Comparisons **36**

F Qualitative Ablation **38**

A Extended Related Work

A.1 Bidirectional Video Diffusion Models

Bidirectional video diffusion models synthesize clips via joint spatiotemporal denoising, leveraging full attention across frames to capture appearance and temporal coherence. Early work extends image diffusion to video through joint image–video training (Ho et al., 2022) or temporal layers added to pretrained image models (Blattmann et al., 2023). Subsequent work improves visual fidelity by fine-tuning spatial modules with high-quality images while preserving motion learned from videos (Chen et al., 2024b). Transformer-based models, including CogVideoX (Yang et al., 2025), Sora (Brooks et al., 2024), HunyuanVideo (Kong et al., 2024), and Wan (Wan et al., 2025), scale this formulation to improve text-conditioned synthesis. To reduce sampling cost, LTX-Video (HaCohen et al., 2024) lowers per-step computation through spatiotemporal latent compression, while video distillation reduces the number of denoising steps. To balance fidelity and diversity in few-step generation, rCM (Zheng et al., 2026b) combines consistency distillation with score regularization, and Adaptive Video Distillation (You et al., 2026) uses adaptive regression and temporal regularization to mitigate oversaturation and motion degradation. Despite these efficiency gains, bidirectional models still jointly denoise entire clips, incurring latency and memory costs that increase with clip length and limit their suitability for low-latency interactive and streaming applications.

A.2 Autoregressive Video Generation and Distillation

Autoregressive video pretraining models future frames or chunks conditioned on preceding context. Early methods use discrete tokens for sequence modeling (Hong et al., 2023; Kondratyuk et al., 2024) or masked prediction with causal video tokenization (Villegas et al., 2023). In continuous latent spaces, NOVA (Deng et al., 2025) combines frame-wise autoregression with spatial set-wise prediction, while Pyramid Flow (Jin et al., 2025) improves efficiency through pyramidal denoising and history compression. Diffusion Forcing (Chen et al., 2024a) introduces independent per-token noise levels, while SkyReels-V2 and MAGI-1 (Chen et al., 2025; Sand.ai et al., 2025) scale temporally structured denoising to long videos. However, these AR diffusion models still require many denoising steps per frame or chunk, limiting real-time generation.

Few-step AR distillation seeks to reduce denoising cost while preserving generation quality under autoregressive rollout. AAPT (Lin et al., 2025) explores adversarial student-forcing training for one-step streaming generation, whereas CausVid (Yin et al., 2025) uses DMD (Yin et al., 2024b;a) to match a causal student's output distribution to that of a bidirectional teacher. Within DMD-based pipelines, subsequent work improves both causal initialization and training under generated histories. Causal Forcing (Zhu et al., 2026) addresses the bidirectional-to-causal architectural gap by using an AR teacher for ODE initialization, while Causal-rCM (Zheng et al., 2026a) uses teacher-forcing consistency distillation to improve initial mode coverage. During DMD training, Self Forcing (Huang et al., 2025) reduces exposure bias by conditioning rollouts on student-generated histories. For longer generation, LongLive (Yang et al., 2026) introduces streaming training and cache management, Rolling Forcing (Liu et al., 2026b) uses joint rolling-window denoising, and Context Forcing (Chen et al., 2026) provides long-context teacher supervision. Despite improved initialization and temporal context handling, these DMD-based approaches retain reverse-KL distribution matching and its mode-seeking tendency. In contrast, Mass Forcing modifies the training objective to explicitly promote mode coverage throughout on-policy DMD optimization.

B MATHEMATICAL DERIVATIONS OF MASS FORCING

This appendix follows the complete mathematical chain underlying Mass Forcing: autoregressive conditioning, distribution-matching gradients, denoising-based score estimation, relative mass integration, and the local and set-level generator updates. We distinguish exact population identities from the finite-particle, learned-score surrogate used in training. All parameter gradients use loss-gradient signs; gradient descent follows their negative.

B.1 CONDITIONING, NOTATION, AND SCOPE

Autoregressive sampling and training histories. In Eq. (1), B chunks of d scalar latent entries form the vectorized video $V \in \mathbb{R}^{Bd}$. Following CausVid and Self Forcing (Yin et al., 2025; Huang et al., 2025), each chunk alternates denoising and re-noising at decreasing noise levels, then enters the clean history represented by a key–value (KV) cache. The sampler randomness $z^{(b)}$ includes both initial and intermediate re-noising draws. Teacher Forcing conditions on clean data histories, while Diffusion Forcing uses noise-corrupted data histories with independently sampled noise levels. Self Forcing instead constructs these histories from its own generated chunks, reducing exposure bias and enabling holistic distribution matching on the completed rollout.

Full-video DMD objective. Corrupting the full video as $V_t = \alpha_t V + \sigma_t \epsilon_V$, with $\epsilon_V \sim \mathcal{N}(0, I_{Bd})$, induces the noisy student distribution $Q_{\theta,t}$. Let P_t denote the corresponding noisy teacher distribution; both are conditioned on prompt c, and I_{Bd} is the full-video identity matrix. The holistic Self Forcing objective is

$$\mathcal{L}_{\mathrm{DMD}}^{\mathrm{video}} = \mathbb{E}_t[\omega(t) D_{\mathrm{KL}}(Q_{\theta,t} \| P_t)], \tag{B.1}$$

where the frozen teacher provides its score and a separate fake denoiser estimates the student score from generated clips. ODE and consistency-based initialization improve the starting generator (Yin et al., 2025; Zhu et al., 2026; Zheng et al., 2026a), but do not change this subsequent reverse-KL objective.

Fixed-prefix analysis. For a selected student-generated prefix u and text prompt c, let $x = G_\theta(z; u, c) \in \mathbb{R}^d$ and $y = \alpha_t x + \sigma_t \epsilon$, where z contains the generator-rollout randomness and $\epsilon \sim \mathcal{N}(0, I_d)$ is independent score-query corruption. Unless specified otherwise, (u, c, t) are fixed, the teacher distribution does not depend on θ, and $\sigma_t > 0$. For the population gradient derivations, assume differentiable densities and generator maps, valid interchange of differentiation and integration, and vanishing boundary terms in integration by parts. Gaussian corruption supplies positive smooth densities even when the clean generator law is not described by a smooth density; the remaining regularity and integrability conditions are separate assumptions.

Full-video versus fixed-prefix objectives. Let $P(V \mid c)$ be the clean teacher video distribution and let $p(x^{(b)} \mid u^{(b)}, c)$ denote its chain-rule conditionals; this factorization does not require the teacher network itself to be autoregressive. When the clean-video KL is finite, its chain-rule decomposition is

$$\begin{aligned} D_{\mathrm{KL}}(Q_\theta \| P) &= \mathbb{E}_{V \sim Q_\theta}\left[\sum_{b=1}^{B} \log \frac{q_\theta(x^{(b)} \mid u^{(b)}, c)}{p(x^{(b)} \mid u^{(b)}, c)}\right] \\ &= \sum_{b=1}^{B} \mathbb{E}_{u^{(b)} \sim Q_\theta^{(<b)}}\left[D_{\mathrm{KL}}\Big(q_\theta(\cdot \mid u^{(b)}, c) \| p(\cdot \mid u^{(b)}, c)\Big)\right], \end{aligned} \tag{B.2}$$

where $Q_\theta^{(<b)}$ is the student's prefix marginal and prompt conditioning is suppressed in the outer densities. Differentiating this full objective also differentiates the prefix distribution. Our conditional analysis instead holds the sampled prefix fixed, matching the detached-prefix update in the implementation. The same KL chain rule applies to noisy full videos, but then the conditionals use noisy prefixes; they are not the clean-prefix conditionals obtained by noising only the next chunk.

Exact fields and learned estimates. The scores $s_{q,t}$ and $s_{p,t}$ refer to population densities in the ideal conditional formulation. The denoisers D_q, D_p and residual R_ϕ are network predictions. Likewise, Δ_t^* denotes an exact conditional score-difference field, whereas Δ_t is its learned estimate.

Overbars on $\bar{x}_k, \bar{y}_k$ mark detached reference samples; the unbarred x_k, y_k are differentiable replays with the same forward values at the start of a generator update.

B.2 From Distribution Matching to Mass-Weighted Gradients

A transport identity for implicit generators. At fixed (u, c, t), abbreviate $q = q_{\theta,t}$, $p = p_t$, and $r = p/q$. The noisy sample is $y_\theta = \alpha_t G_\theta(z; u, c) + \sigma_t \epsilon$, with $\partial y_\theta / \partial \theta = \alpha_t J_\theta$. Define the conditional mean Jacobian

$$\mathcal{A}_{\theta,t}(y) = \mathbb{E}[\alpha_t J_\theta \mid y_\theta = y]. \tag{B.3}$$

It has one column per generator parameter. For any smooth compactly supported test function ψ, differentiation through the sample and integration by parts give

$$\begin{aligned} \nabla_\theta \int q(y)\psi(y)\,\mathrm{d}y &= \mathbb{E}[\alpha_t J_\theta^\top \nabla_y \psi(y_\theta)] \\ &= \int q(y)\mathcal{A}_{\theta,t}(y)^\top \nabla_y \psi(y)\,\mathrm{d}y \\ &= -\int \psi(y)\,\mathrm{div}_y[q(y)\mathcal{A}_{\theta,t}(y)]\,\mathrm{d}y. \end{aligned} \tag{B.4}$$

Thus the density derivative satisfies

$$\nabla_\theta q(y) = -\,\mathrm{div}_y[q(y)\mathcal{A}_{\theta,t}(y)], \tag{B.5}$$

where divergence is applied columnwise and returned as a parameter-gradient vector. This argument does not require an invertible generator or a one-to-one map from z to y.

Reverse KL. The conditional DMD objective is

$$\mathcal{L}_{\mathrm{RKL}} = \mathbb{E}_t[\omega(t) D_{\mathrm{KL}}(q_{\theta,t} \| p_t)]\,. \tag{B.6}$$

At a fixed noise level, the teacher is constant with respect to θ and

$$\begin{aligned} \nabla_\theta D_{\mathrm{KL}}(q\|p) &= \int \left(\log\frac{q}{p} + 1\right) \nabla_\theta q\,\mathrm{d}y \\ &= -\int \left(\log\frac{q}{p} + 1\right) \mathrm{div}_y(q\mathcal{A}_{\theta,t})\,\mathrm{d}y \\ &= \int q\mathcal{A}_{\theta,t}^\top \nabla_y \log\frac{q}{p}\,\mathrm{d}y \\ &= \mathbb{E}_{z,\epsilon}\left[\alpha_t J_\theta^\top (s_{q,t} - s_{p,t})\right]. \end{aligned} \tag{B.7}$$

The constant term vanishes because $\int \nabla_\theta q\,\mathrm{d}y = 0$. Averaging over t with $\omega(t)$ recovers Eq. (2), the standard score-difference gradient underlying DMD (Yin et al., 2024b;a).

Forward KL. Keeping the same generator distribution and fixed teacher,

$$\begin{aligned} \nabla_\theta D_{\mathrm{KL}}(p\|q) &= -\int \frac{p}{q}\nabla_\theta q\,\mathrm{d}y \\ &= \int r\,\mathrm{div}_y(q\mathcal{A}_{\theta,t})\,\mathrm{d}y \\ &= -\int q\mathcal{A}_{\theta,t}^\top \nabla_y r\,\mathrm{d}y. \end{aligned} \tag{B.8}$$

Since

$$\nabla_y r = r\nabla_y \log(p/q) = r(s_{p,t} - s_{q,t}), \tag{B.9}$$

the forward-KL gradient is

$$\nabla_\theta D_{\mathrm{KL}}(p\|q) = \mathbb{E}_{z,\epsilon}\left[\alpha_t J_\theta^\top r_t(y)(s_{q,t}(y) - s_{p,t}(y))\right]. \tag{B.10}$$

Relative to reverse KL, the correction is weighted by the teacher-to-student density ratio, consistent with the general divergence-gradient analysis of f-Distill (Xu et al., 2025).

Mixed divergence and effective mass. For a fixed $\lambda \in [0,1]$, define

$$\mathcal{L}_{\text{mix}} = \mathbb{E}_t\big[\omega(t)\big((1-\lambda)D_{\text{KL}}(q_{\theta,t}\|p_t) + \lambda D_{\text{KL}}(p_t\|q_{\theta,t})\big)\big],$$
$$\nabla_\theta \mathcal{L}_{\text{mix}} = \mathbb{E}_{t,z,\epsilon}\big[\omega(t)\alpha_t J_\theta^\top h_t^*(y)(s_{q,t}(y) - s_{p,t}(y))\big], \tag{B.11}$$

with $h_t^* = 1 - \lambda + \lambda r_t$. Multiplying the sampling density by this weight gives

$$q_{\theta,t} h_t^* = (1-\lambda)q_{\theta,t} + \lambda p_t, \qquad \mathbb{E}_{q_{\theta,t}}[h_t^*] = 1. \tag{B.12}$$

This is an identity for the effective measure in the gradient estimate, not a change to the generator's inference sampler. The derivation already accounts for the dependence of q and r on θ. Once this gradient field has been evaluated, differentiating through its score or ratio factors again would not reproduce the derived first-order gradient; the implementation instead injects the detached field through a surrogate loss.

B.3 From Denoising Predictions to Conditional Density Ratios

Shared-prefix query construction. For K detached continuations $\bar{x}_k$ generated with independent rollout noises, the score inputs are

$$\bar{y}_k = \alpha_t \bar{x}_k + \sigma_t \epsilon, \qquad u_t = \alpha_t u + \sigma_t \epsilon_u. \tag{B.13}$$

The current-chunk corruption $\epsilon \sim \mathcal{N}(0, I_d)$ is shared across the group; ϵ_u is an independent standard Gaussian tensor with the prefix's shape and is also held fixed across queries. This sharing applies only to score-query corruption, not to the independent generator noises z_k. The networks receive the joint clip $[u_t, \bar{y}_k]$ and return clean-latent predictions; only the continuation block enters the generator correction.

Gaussian denoising identity. For a fixed conditioning state, write the noisy student density as

$$q_{\theta,t}(y) = \int \mathcal{N}(y; \alpha_t x, \sigma_t^2 I_d)\, q_\theta(\mathrm{d}x), \tag{B.14}$$

where $q_\theta(\mathrm{d}x)$ denotes the clean student probability measure. Differentiating the Gaussian kernel yields

$$\nabla_y \log q_{\theta,t}(y) = \frac{1}{q_{\theta,t}(y)} \int \frac{\alpha_t x - y}{\sigma_t^2} \mathcal{N}(y; \alpha_t x, \sigma_t^2 I_d)\, q_\theta(\mathrm{d}x)$$
$$= \frac{\alpha_t \mathbb{E}_q[x \mid y] - y}{\sigma_t^2}. \tag{B.15}$$

Let $D_q^*(y) = \mathbb{E}_q[x \mid y]$ and $D_p^*(y) = \mathbb{E}_p[x \mid y]$ be the population-optimal clean-latent denoisers under the same conditioning and corruption. Applying Eq. (B.15) to both distributions cancels the common $-y/\sigma_t^2$ term:

$$s_{p,t}(y) - s_{q,t}(y) = \frac{\alpha_t}{\sigma_t^2}\big(D_p^*(y) - D_q^*(y)\big). \tag{B.16}$$

Replacing the population denoisers by their network estimates gives the practical field in Eq. (B.19).

Bidirectional joint queries. For the jointly corrupted prefix and continuation, let $p_t^{\text{joint}}(u_t, y \mid c)$ and $q_t^{\text{joint}}(u_t, y \mid c)$ denote the teacher and student densities, and let $p_t^{\text{pre}}, q_t^{\text{pre}}$ be their prefix marginals. The conditional densities satisfy

$$\tilde{p}_t(y \mid u_t, c) = \frac{p_t^{\text{joint}}(u_t, y \mid c)}{p_t^{\text{pre}}(u_t \mid c)}, \qquad \tilde{q}_t(y \mid u_t, c) = \frac{q_t^{\text{joint}}(u_t, y \mid c)}{q_t^{\text{pre}}(u_t \mid c)}. \tag{B.17}$$

At fixed u_t, the prefix marginals are constant with respect to y, so

$$\nabla_y \log p_t^{\text{joint}}(u_t, y \mid c) = \nabla_y \log \tilde{p}_t(y \mid u_t, c),$$
$$\Delta_t^*(y) = \nabla_y \log \frac{\tilde{p}_t(y \mid u_t, c)}{\tilde{q}_t(y \mid u_t, c)} = \nabla_y \log \tilde{r}_t(y \mid u_t, c). \tag{B.18}$$

Thus the current-chunk component of a bidirectional joint score supplies a conditional score at the fixed noisy prefix; a separate causal teacher is not needed for this identity. The shared prefix also ensures that prefix-marginal density factors cancel when comparing continuations within a group. This result does not identify conditioning on u_t with conditioning on the exact clean u. For the empty prefix, the same expressions reduce to prompt-conditioned continuation densities.

Residual parameterization. The implementation uses $D_q = D_p + R_\phi$, where D_p is the frozen conditional teacher prediction and R_ϕ is a residual estimator with trainable parameters ϕ. Consequently,

$$\Delta_t(y) = \frac{\alpha_t}{\sigma_t^2}(D_p - D_q) = -\frac{\alpha_t}{\sigma_t^2} R_\phi(y). \tag{B.19}$$

The cancellation is algebraic at every queried point, including integration midpoints, and is specific to the conditional teacher used as the residual baseline. Whether this estimated field matches Δ_t^* depends on denoiser calibration, including at interpolated inputs.

B.4 Path Integration, Shared Noise, and Normalized Weights

Recovering relative log ratios. All rollouts in a group share (u_t, t, c). Choose the medoid anchor $a = \bar{y}_{k^*}$ with $k^* = \arg\min_j \sum_k \|\bar{y}_j - \bar{y}_k\|_2$, and use the path $\zeta_k(\xi) = a + \xi(\bar{y}_k - a)$, $\xi \in [0, 1]$. For the exact field $\Delta_t^* = \nabla_y \log \tilde{r}_t$, the chain rule gives

$$\frac{\mathrm{d}}{\mathrm{d}\xi} \log \tilde{r}_t(\zeta_k(\xi)) = \Delta_t^*(\zeta_k(\xi))^\top (\bar{y}_k - a). \tag{B.20}$$

Integrating from 0 to 1 yields

$$\int_0^1 \Delta_t^*(\zeta_k(\xi))^\top (\bar{y}_k - a)\, \mathrm{d}\xi = \log \tilde{r}_t(\bar{y}_k) - \log \tilde{r}_t(a). \tag{B.21}$$

The common anchor leaves only an additive constant shared by the group, so absolute density normalizers are unnecessary for relative weighting. This is integration in latent space, not integration over the diffusion time coordinate.

Three-point quadrature. Replacing the exact field by Δ_t and applying Simpson's rule to the scalar path integrand gives

$$\ell_k = \frac{1}{6}\left\langle \Delta_t(a) + 4\Delta_t\left(\frac{a + \bar{y}_k}{2}\right) + \Delta_t(\bar{y}_k),\ \bar{y}_k - a \right\rangle. \tag{B.22}$$

The anchor rollout has zero displacement and hence $\ell_{k^*} = 0$. The coefficients integrate a cubic scalar integrand exactly, but a neural score field need not vary cubically along the path. Score-estimation error and quadrature error are distinct: even an exact score field can require finer quadrature when it changes sharply between modes. A learned residual field also need not be conservative, in which case different paths or anchors can yield different estimates rather than one globally consistent log-density ratio.

Why corruption noise is shared. Write $a = \alpha_t \bar{x}_{k^*} + \sigma_t \epsilon$ and $\bar{y}_k = \alpha_t \bar{x}_k + \sigma_t \epsilon$. Then every path point is

$$\zeta_k(\xi) = \alpha_t\big[(1-\xi)\bar{x}_{k^*} + \xi \bar{x}_k\big] + \sigma_t \epsilon, \qquad \bar{y}_k - a = \alpha_t(\bar{x}_k - \bar{x}_{k^*}). \tag{B.23}$$

The Gaussian corruption variance is therefore $\sigma_t^2 I_d$ throughout the path. For fixed endpoint indices with independent noises ϵ_{k^*} and ϵ_k, interpolation would instead give

$$\mathrm{Cov}\big[\sigma_t\big((1-\xi)\epsilon_{k^*} + \xi\epsilon_k\big)\big] = \sigma_t^2\big((1-\xi)^2 + \xi^2\big) I_d, \tag{B.24}$$

which equals $\sigma_t^2 I_d/2$ at the midpoint despite querying the networks at noise level t. Shared corruption removes this variance mismatch, although the interpolated clean content can still be outside the rollout distribution. At $\alpha_t = 0$, the shared-noise particles coincide, all path displacements vanish, and the estimator carries no continuation-specific mass information.

Anchor cancellation and ideal importance weighting. To isolate normalization, first consider exact ratios $r = p/q$ and ideal IID reference particles from that same proposal q. With $\ell_k^* = \log r(\bar{y}_k) - \log r(a)$, exponentiation and normalization give

$$\frac{\exp(\ell_k^*)}{\sum_j \exp(\ell_j^*)} = \frac{r(\bar{y}_k)}{\sum_j r(\bar{y}_j)}. \tag{B.25}$$

For bounded test functions ψ, the corresponding self-normalized estimate obeys

$$\frac{\sum_k r(\bar{y}_k)\psi(\bar{y}_k)}{\sum_k r(\bar{y}_k)} \longrightarrow \frac{\mathbb{E}_q[r\psi]}{\mathbb{E}_q[r]} = \mathbb{E}_p[\psi] \quad \text{as } K \to \infty. \tag{B.26}$$

This is an IID population benchmark, not an exact statement about the practical shared-noise groups. Even in the IID case, the finite-K ratio of sample sums is generally biased.

Stabilized mixture weights. The logits in Eq. (5) are computed as

$$\bar{\ell} = \frac{1}{K}\sum_j \ell_j, \qquad v_k = \mathrm{clip}\left(\frac{\ell_k - \bar{\ell}}{\tau}, -\kappa, \kappa\right), \tag{B.27}$$

where $\bar{\ell}$ is the group mean, $\tau > 0$ is the temperature, and $\kappa > 0$ bounds the logits *after* temperature scaling. Centering removes the common anchor offset before clipping, and the subsequent softmax yields the relative estimates $\widehat{r}_k$. The detached, group-shared coefficient $\lambda_{\text{eff}} \in [0, \lambda_{\max}]$, $\lambda_{\max} \le 1$, incorporates the training ramp and reliability gates. Without clipping, temperature scaling replaces r by $r^{1/\tau}$; in the ideal IID limit, the reweighted density is proportional to $p^{1/\tau}q^{1-1/\tau}$ rather than p unless $\tau = 1$. Clipping further changes this weighting and bounds its concentration. By construction,

$$\sum_{k=1}^K \widehat{r}_k = K, \qquad \sum_{k=1}^K h_k = K, \qquad \sum_{k=1}^K \pi_k = 1, \qquad \pi_k = \frac{h_k}{K}. \tag{B.28}$$

For $0 \le \lambda_{\text{eff}} \le 1$, positivity of the normalized ratios also gives

$$1 - \lambda_{\text{eff}} \le h_k \le 1 - \lambda_{\text{eff}} + \lambda_{\text{eff}} K. \tag{B.29}$$

The reference measure can be written explicitly as

$$\begin{aligned} \nu &= (1 - \lambda_{\text{eff}})\mu_{\text{ref}} + \lambda_{\text{eff}}\widehat{\nu}_p, \\ \mu_{\text{ref}} &= \frac{1}{K}\sum_k \delta_{\bar{y}_k}, \qquad \widehat{\nu}_p = \frac{1}{K}\sum_k \widehat{r}_k \delta_{\bar{y}_k}. \end{aligned} \tag{B.30}$$

Here μ_{ref} is the uniform detached reference measure and $\widehat{\nu}_p$ is the estimated teacher-reweighted measure, not a set of independently generated teacher samples.

Shared-noise dependence and conditioning mismatch. The actual particles share one corruption draw ϵ. Conditioned on this draw and the clean prefix, increasing K averages over generator randomness, not over the marginal noisy proposal $q_{\theta,t}$. Even if the exact clean-prefix ratio r_t were available, the conditional large-group limit would be

$$\frac{\mathbb{E}_z[r_t(y_\theta)\psi(y_\theta) \mid \epsilon, u, c]}{\mathbb{E}_z[r_t(y_\theta) \mid \epsilon, u, c]}, \qquad y_\theta = \alpha_t G_\theta(z; u, c) + \sigma_t \epsilon, \tag{B.31}$$

which is not generally $\mathbb{E}_{p_t}[\psi]$. Averaging these normalized ratios over ϵ does not generally remove the difference. In addition, the joint-score queries estimate a noisy-prefix ratio $\tilde{r}_t$, whereas a group is generated from one fixed clean prefix. These facts, together with clipping, tempering, and learned scores, are why we interpret the practical weights as a finite-particle coverage surrogate rather than exact forward-KL importance weights.

B.5 Target-Aligned Local Score Distillation

Separating conditional transport from guidance. At a common noisy input and conditioning state, the unguided and guided teacher predictions satisfy

$$\begin{aligned} D_{\text{cfg}} &= D_{\text{uncond}} + w_{\text{cfg}}(D_p - D_{\text{uncond}}), \\ D_q - D_{\text{cfg}} &= (D_q - D_p) + (D_p - D_{\text{cfg}}). \end{aligned} \tag{B.32}$$

The first component compares student and conditional teacher; the second supplies the additional CFG correction. Because the mass estimates target the conditional teacher, applying h_k to the entire guided correction would also multiply the guidance component by h_k. The difference between that update and the chosen decomposition is

$$h_k(D_{q,k} - D_{\text{cfg},k}) - \left[h_k(D_{q,k} - D_{p,k}) + (D_{p,k} - D_{\text{cfg},k})\right] = (h_k - 1)(D_{p,k} - D_{\text{cfg},k}). \tag{B.33}$$

Equation (6) removes this additional modulation and reduces exactly to the unweighted guided correction when $h_k = 1$. It is a conditional mass correction plus uniform guidance, not the mixed-KL gradient to a single CFG-guided target distribution.

Denoiser units and normalization. Combining the Jacobian factor from Eq. (2) with the Gaussian identity gives the ideal conditional contribution

$$\omega(t)\alpha_t J_\theta^\top h_t^*(s_{q,t} - s_{p,t}) = \omega(t)\frac{\alpha_t^2}{\sigma_t^2} J_\theta^\top h_t^*(D_q^* - D_p^*). \tag{B.34}$$

The implementation instead uses network predictions and the per-rollout normalization

$$n_k = \mathrm{sg}\left[\max\left(\frac{1}{d}\|\bar{x}_k - D_{\mathrm{cfg},k}\|_1, \epsilon_n\right)\right], \qquad \epsilon_n > 0. \tag{B.35}$$

Here ϵ_n is a numerical floor, not a Gaussian noise variable. This sample-dependent normalization is not merely a noise-level weight $\omega(t)$, so the practical normalized update is not asserted to optimize the exact KL mixture in Eq. (B.11).

Injecting a prescribed correction. We replay $x_k = G_\theta(z_k; u, c)$ with the same randomness and unchanged generator parameters as the detached reference $\bar{x}_k$. The forward values agree, while x_k retains the generator graph. The local surrogate in Eq. (7) is

$$\mathcal{L}_{\mathrm{local}} = \frac{1}{2Kd}\sum_k \|x_k - \mathrm{sg}(x_k - g_k)\|_2^2. \tag{B.36}$$

Here $x_k^{\mathrm{tgt}} = \mathrm{sg}(x_k - g_k)$ is a detached gradient-injection target, not a teacher-generated clean video. During differentiation, the scores, h_k, n_k, and this target are fixed. The derivative of Eq. (B.36) is therefore

$$\begin{aligned}\nabla_{x_k}\mathcal{L}_{\mathrm{local}} &= \frac{1}{Kd}(x_k - x_k^{\mathrm{tgt}}) = \frac{g_k}{Kd},\\ \nabla_\theta\mathcal{L}_{\mathrm{local}} &= \frac{1}{Kd}\sum_k J_{\theta,k}^\top g_k,\end{aligned} \tag{B.37}$$

where $J_{\theta,k} = \partial x_k/\partial\theta$ is the Jacobian of the replayed generator graph. The factor $1/(Kd)$ comes from averaging over rollouts and latent coordinates. Stopping the target derivative does not remove the correction: it is the derivative of the outer squared distance that supplies g_k. No generator-update gradient passes through the teacher, fake-score estimator, or mass weights.

B.6 Why Rollout-Set Matching Complements Local Corrections

Mode probabilities need not determine within-mode scores. Consider teacher and student mixtures with common normalized component densities $f_m(y)$ but different positive mode probabilities $\rho_m^{(p)}, \rho_m^{(q)}$:

$$\tilde{p}_t(y) = \sum_m \rho_m^{(p)} f_m(y), \qquad \tilde{q}_t(y) = \sum_m \rho_m^{(q)} f_m(y), \qquad \sum_m \rho_m^{(p)} = \sum_m \rho_m^{(q)} = 1. \tag{B.38}$$

Where a well-separated component m dominates both densities,

$$\begin{aligned}\nabla_y \log \tilde{p}_t(y) &\approx \nabla_y[\log \rho_m^{(p)} + \log f_m(y)] = \nabla_y \log f_m(y),\\ \nabla_y \log \tilde{q}_t(y) &\approx \nabla_y \log f_m(y),\\ \log \tilde{r}_t(y) &\approx \log \rho_m^{(p)} - \log \rho_m^{(q)}.\end{aligned} \tag{B.39}$$

The within-mode corrections can therefore be small while the relative log-ratio offsets differ across modes. For the positive, smoothed densities considered here, these offsets are connected by score-field variation between modes, which the line integral can recover when that field is accurately estimated and integrated. If the exact field were zero everywhere along the connecting path, the offsets could not differ. This distinction concerns weak endpoint corrections, not an absence of information throughout the field, and does not guarantee the accuracy of three-point quadrature.

Kernel discrepancy between the empirical measures. Let $\mathcal{H}_\beta$ be the reproducing-kernel Hilbert space associated with the Gaussian kernel

$$\mathcal{K}_\beta(y, y') = \exp\left(-\frac{\|y - y'\|_2^2}{2d\beta}\right), \qquad \beta > 0. \tag{B.40}$$

Its feature map Φ_β satisfies $\langle \Phi_\beta(y), \Phi_\beta(y') \rangle_{\mathcal{H}_\beta} = \mathcal{K}_\beta(y, y')$. Using the empirical source and reference in Eq. (8), the squared MMD (Gretton et al., 2012; Li et al., 2015) is

$$\begin{aligned}\mathcal{L}_{\text{set}} &= \left\| \frac{1}{K} \sum_i \Phi_\beta(y_i) - \sum_j \pi_j \Phi_\beta(\bar{y}_j) \right\|_{\mathcal{H}_\beta}^2 \\ &= \frac{1}{K^2} \sum_{i,j} \mathcal{K}_\beta(y_i, y_j) - \frac{2}{K} \sum_{i,j} \pi_j \mathcal{K}_\beta(y_i, \bar{y}_j) + \sum_{i,j} \pi_i \pi_j \mathcal{K}_\beta(\bar{y}_i, \bar{y}_j).\end{aligned} \tag{B.41}$$

Every pair, including the diagonal, is retained. The scale β is the detached median of off-diagonal mean squared distances in the pooled source–reference set, subject to the lower bound $\beta_{\min} > 0$. It is held fixed during differentiation, together with the reference locations and weights. Because the reference reuses the same rollouts, this is a nonnegative empirical matching objective, not an unbiased statistic from two independent sample sets.

Differentiating the source interactions. The derivative with respect to y_i receives contributions from both appearances of y_i in the source–source sum. For the symmetric kernel,

$$\begin{aligned}\nabla_{y_i} \left[\frac{1}{K^2} \sum_{j,k} \mathcal{K}_\beta(y_j, y_k) \right] &= \frac{1}{K^2} \sum_j \left[\nabla_1 \mathcal{K}_\beta(y_i, y_j) + \nabla_2 \mathcal{K}_\beta(y_j, y_i) \right] \\ &= \frac{2}{K^2} \sum_j \nabla_1 \mathcal{K}_\beta(y_i, y_j), \\ \nabla_{y_i} \left[-\frac{2}{K} \sum_{k,j} \pi_j \mathcal{K}_\beta(y_k, \bar{y}_j) \right] &= -\frac{2}{K} \sum_j \pi_j \nabla_1 \mathcal{K}_\beta(y_i, \bar{y}_j).\end{aligned} \tag{B.42}$$

Here i, j, k index rollouts, and ∇_1, ∇_2 differentiate the first and second kernel arguments. The reference–reference term has zero source derivative. Combining the two contributions gives

$$\nabla_{y_i} \mathcal{L}_{\text{set}} = \frac{2}{K} \left[\frac{1}{K} \sum_j \nabla_1 \mathcal{K}_\beta(y_i, y_j) - \sum_j \pi_j \nabla_1 \mathcal{K}_\beta(y_i, \bar{y}_j) \right], \tag{B.43}$$

with

$$\nabla_1 \mathcal{K}_\beta(y, y') = -\frac{y - y'}{d\beta} \mathcal{K}_\beta(y, y'). \tag{B.44}$$

The source–source term is essential; retaining only source–reference attraction would define a different objective.

Identical support does not imply a zero update. At replay initialization, $y_i = \bar{y}_i$ but the source weights are $1/K$ and the reference weights are $\pi_j = h_j/K$. Consequently, the replay-initialization gradient stated in Eq. (11) is

$$\nabla_{y_i} \mathcal{L}_{\text{set}}|_{y=\bar{y}} = \frac{2}{K^2} \sum_j (1 - h_j) \nabla_1 \mathcal{K}_\beta(\bar{y}_i, \bar{y}_j). \tag{B.45}$$

Nonuniform mass weights can therefore induce a geometric correction even when local score differences at the rollouts are weak. When all $h_j = 1$, the measures coincide, the MMD loss is zero, and every source gradient vanishes. Nonuniform weights do not guarantee a useful nonzero update: coincident particles or negligible kernel interactions can still eliminate the signal. The objective compares distributions over the sampled support; it neither supplies absent teacher continuations nor proves recovery of all missing modes.

B.7 Combined Generator Gradient and Fake-Score Training

The collection and replay procedure is specified in Appendix C.3. Setting $\gamma = 0$ retains only local mass-weighted distillation. Also setting $\lambda_{\text{eff}} = 0$ recovers unweighted conditional-chunk DMD, not the holistic Self Forcing data path.

The generator chain rule. The differentiable source satisfies $y_k = \alpha_t G_\theta(z_k; u, c) + \sigma_t \epsilon$, with the same fixed corruption as its reference. Thus $\partial y_k / \partial \theta = \alpha_t J_{\theta,k}$, and

$$\nabla_\theta \mathcal{L}_{\text{set}} = \sum_k \alpha_t J_{\theta,k}^\top \nabla_{y_k} \mathcal{L}_{\text{set}}. \tag{B.46}$$

Combining this expression with Eq. (B.37) yields

$$\nabla_\theta \mathcal{L}_{\text{MF}} = \sum_k J_{\theta,k}^\top \left[\frac{g_k}{Kd} + \gamma \alpha_t \nabla_{y_k} \mathcal{L}_{\text{set}} \right]. \tag{B.47}$$

This is the gradient of the per-update surrogate with reference locations, mass weights, normalization, and kernel scale fixed. The references are collected and replayed with identical randomness before any generator optimizer step, and the set derivative includes both source appearances in every pairwise interaction. With a full differentiable rollout, $J_{\theta,k}$ is the full conditional generator Jacobian. Under Self Forcing's memory-efficient truncation, it is the Jacobian of the retained computation graph: gradients do not traverse the cached prefix or detached intermediate denoising transitions (Huang et al., 2025). This implemented derivative is therefore not the full gradient through all earlier autoregressive states in Eq. (B.2).

Fake-score objective. The fake denoiser tracks the evolving student distribution without applying the generator's mass weights to its samples. For a detached student joint clip $Y_0 = [u, \bar{x}]$, let t and ϵ_Y denote its sampled noise level and standard Gaussian corruption, giving $Y_t = (1 - \sigma_t) Y_0 + \sigma_t \epsilon_Y$. The paired implementation caches these denoising examples for critic replay, as described in Appendix C.3. The superscript joint denotes full-clip predictions, with $D_q^{\text{joint}} = D_p^{\text{joint}} + R_\phi^{\text{joint}}$. The flow-matching objective is

$$\mathcal{L}_q = \mathbb{E}\left[\left\| \frac{Y_t - D_q^{\text{joint}}(Y_t, t, c)}{\sigma_t} - (\epsilon_Y - Y_0) \right\|_{\text{av}}^2 \right], \tag{B.48}$$

where the expectation covers prompts, on-policy joint clips, noise levels, and corruption, and $\| \cdot \|_{\text{av}}^2$ averages over joint-clip latent coordinates. Only ϕ is updated; the teacher predictions and generated clips remain detached.

Flow matching is denoising regression. For the same noisy joint clip,

$$Y_t = (1 - \sigma_t) Y_0 + \sigma_t \epsilon_Y, \qquad \frac{Y_t - D_q^{\text{joint}}}{\sigma_t} - (\epsilon_Y - Y_0) = \frac{Y_0 - D_q^{\text{joint}}}{\sigma_t}. \tag{B.49}$$

The first fraction on the left is the predicted rectified-flow velocity expressed through the clean-latent prediction; the target velocity is $\epsilon_Y - Y_0$. Squaring Eq. (B.49) shows that the critic objective in Eq. (B.48) equals

$$\mathcal{L}_q = \mathbb{E}\left[\frac{1}{\sigma_t^2} \| D_q^{\text{joint}}(Y_t, t, c) - Y_0 \|_{\text{av}}^2 \right]. \tag{B.50}$$

The noise-dependent factor is retained; this is not unweighted clean-latent MSE across timesteps.

What the residual estimator learns. Substituting $D_q^{\text{joint}} = D_p^{\text{joint}} + R_\phi^{\text{joint}}$ gives

$$\mathcal{L}_q = \mathbb{E}\left[\frac{1}{\sigma_t^2} \| R_\phi^{\text{joint}}(Y_t, t, c) - (Y_0 - D_p^{\text{joint}}(Y_t, t, c)) \|_{\text{av}}^2 \right]. \tag{B.51}$$

At fixed (Y_t, t, c), the teacher prediction is a deterministic frozen offset and $1/\sigma_t^2$ is positive and constant with respect to the clean target. The conditional squared-error derivative is proportional to

the residual prediction minus the conditional mean of $Y_0 - D_p^{\text{joint}}$. Setting this derivative to zero under the student joint-clip law gives

$$R^{\text{joint},*}(Y_t, t, c) = \mathbb{E}_q[Y_0 \mid Y_t, t, c] - D_p^{\text{joint}}(Y_t, t, c),$$
$$D_p^{\text{joint}}(Y_t, t, c) + R^{\text{joint},*}(Y_t, t, c) = \mathbb{E}_q[Y_0 \mid Y_t, t, c]. \tag{B.52}$$

This is a function-space optimum, not a guarantee that a finite residual network attains it. It explains why generated student clips suffice to train the residual: the teacher information enters through the known denoising offset, while the regression target distribution remains the student's. The critic loss is not a forward-KL generator objective, and its samples are not reweighted by h_k. Only ϕ is updated in this step.

C Implementation Details

This section specifies the models, sampling procedure, mass-activation rules, and optimizer updates used for Mass Forcing. Table 3 lists all numerical settings, which the text refers to by symbol. All step counts refer to generator optimizer updates unless otherwise stated.

Table 3: Hyperparameters of Mass Forcing. Step counts refer to generator updates.

Symbol	Description	Value
Rollouts and score queries		
B	Chunks per training video	7
F	Latent frames per chunk	3
K	Rollouts per shared prefix	4
–	Video latent shape (time, channel, height, width)	$21 \times 16 \times 60 \times 104$
–	Additional noise on rollout history	0
–	Student denoising timesteps (four-step sampler)	$\{1000, 750, 500, 250\}$
T	Number of diffusion timesteps	1000
s_{shift}	Timestep shift for score queries	5
$[t_{\min}, t_{\max}]$	Score-query timestep range	$[20, 980]$
–	Score-query microbatch size	2
Mass estimation and activation		
τ	Log-ratio temperature	2
κ	Logit clipping bound	2
$\lambda_{\max}$	Maximum mixture coefficient	0.25
N_{ramp}	Ramp length of λ_{eff}	25
$[E_{\text{lo}}, E_{\text{hi}}]$	ESS gate thresholds (fraction of K)	$[0.35, 0.60]$
$[\sigma_{\text{lo}}, \sigma_{\text{hi}}]$	Noise-level range of active cells	$[0.20, 0.65]$
Objectives		
w_{cfg}	Classifier-free guidance scale	3.0
γ	Set-level (MMD) loss weight	0.05
ϵ_n	Floor of the DMD normalizer n_k	10^{-8}
$\beta_{\min}$	Floor of the MMD bandwidth β	10^{-6}
Optimization		
N_{train}	Generator updates	1000
N_{critic}	Critic updates per generator update	5
$N_{\mathcal{B}}$	Critic replay-buffer capacity (entries)	8
N_{ckpt}	Checkpoint interval	100
–	Optimizer (generator and critic)	AdamW
–	AdamW momentum coefficients	$(0,\ 0.999)$
–	Weight decay	0.01
–	Learning rate: generator / critic	2×10^{-6} / 4×10^{-7}
–	Batch size: per rank / global	2 / 16
–	Gradient accumulation steps	1
–	Generator EMA decay	0.99
–	Generator EMA start update	40
–	Generator EMA update interval	10
–	Seed: training / inference	3407 / 0

C.1 Models and On-Policy Rollout Sampling

We use a causal Wan2.1-T2V-1.3B generator initialized from the Self Forcing ODE checkpoint, a frozen bidirectional Wan2.1-T2V-14B teacher, and a bidirectional 1.3B residual estimator (Wan et al., 2025; Huang et al., 2025). The fake denoiser is parameterized as $D_q = D_p + R_\phi$ and initialized from the calibrated residual checkpoint before DMD training. The teacher and text encoder remain frozen.

Prefix and continuation sampling. Each training video consists of B chunks of F latent frames, including the first chunk, with no separate first-frame block. For each batch, we sample b uniformly from $\{1, \dots, B\}$ and synchronize it across workers. We autoregressively generate the first $b-1$ chunks with the current student and detach the resulting prefix and KV cache; no additional context noise is applied during rollout. For $b = 1$, the prefix is empty. From this shared prefix, we sample K alternative realizations of chunk b with independently sampled IID Gaussian noise for initialization and inter-step re-noising, recording the randomness required for deterministic replay. The score networks receive joint clips of $L = Fb$ latent frames, but only the current chunk is optimized in the generator update. The student retains the four-step Self Forcing sampler with denoising-timestep warping enabled.

Score-query timesteps. DMD and critic updates retain the original full-range score-query schedule rather than sampling only inside the ratio band. The inherited sampler draws t_{raw} uniformly from $\{0, \dots, T-1\}$, applies the timestep shift $t = s_{\mathrm{shift}} t_{\mathrm{raw}} / \big(1 + (s_{\mathrm{shift}} - 1) t_{\mathrm{raw}}/T\big)$, and clamps t to $[t_{\min}, t_{\max}]$. The corruption level σ_t is obtained from the scheduler at the resulting timestep. The ratio queries and set-level loss reuse the DMD timestep and endpoint corruption. Critic training uses the full-schedule denoising examples stored in the replay buffer, as detailed in Appendix C.3.

C.2 Mass Estimation and Activation

The ratio estimator uses the pure conditional teacher, shared prefix/current corruption across the K rollouts, a medoid anchor, and three-point Simpson quadrature. We center the log-ratio estimates before temperature scaling and clipping, using the temperature τ and logit bound κ in Eq. (B.27). CFG, with scale w_{cfg}, is used only in the generator correction; it does not enter ratio estimation. Endpoint predictions are reused for both the ratio and local DMD correction. Under the residual parameterization, additional integration midpoints require only residual-network evaluations because the conditional teacher cancels in the score difference.

Effective mixture coefficient. We compute effective sample size from the normalized ratios before mixing them with the uniform weights:

$$E = \frac{(\sum_k \widehat{r}_k)^2}{K \sum_k \widehat{r}_k^2}, \qquad \eta_{\mathrm{ESS}} = \mathrm{clip}\left(\frac{E - E_{\mathrm{lo}}}{E_{\mathrm{hi}} - E_{\mathrm{lo}}}, 0, 1\right). \tag{C.1}$$

Here $E = \mathrm{ESS}/K$, so the ESS attenuation is zero for $E \le E_{\mathrm{lo}}$ and one for $E \ge E_{\mathrm{hi}}$. Let N_G denote the generator-update index, $\chi_{\mathrm{cal}} \in \{0, 1\}$ indicate a valid fake-calibration certificate, and $\chi_{\mathrm{cell}}(\sigma_t, L, b) \in \{0, 1\}$ select prevalidated noise-level and clip-length cells within $\sigma_t \in [\sigma_{\mathrm{lo}}, \sigma_{\mathrm{hi}}]$. The default schedule is

$$\lambda_{\mathrm{eff}} = \lambda_{\max} \min\left(\frac{N_G}{N_{\mathrm{ramp}}}, 1\right) \chi_{\mathrm{cal}}\, \chi_{\mathrm{cell}}(\sigma_t, L, b)\, \eta_{\mathrm{ESS}}. \tag{C.2}$$

Thus the nominal ratio contribution ramps linearly from zero to $\lambda_{\max}$ over the first N_{ramp} generator updates; critic-only initialization steps do not advance this counter. The particle-spread gate is disabled in this configuration. All gate outputs and weights are detached, and normalization is performed separately for each prompt–prefix group.

Outside active cells, $\lambda_{\mathrm{eff}} = 0$ and $h_k = 1$. The local correction then reduces to unweighted conditional DMD, and the mixed-reference MMD loss and gradient vanish under identical replay. These activation rules do not restrict the full noise schedule used for DMD or critic learning.

C.3 Replay, Optimization, and Evaluation

Generator update. We collect detached reference rollouts, compute their mass weights and score corrections, and then replay the same rollouts with identical randomness before changing the generator parameters. The local correction uses the mean absolute clean-latent error normalizer in Eq. (B.35) with numerical floor ϵ_n. The MMD source is uniformly weighted, while its fixed reference uses $\pi_k = \mathrm{sg}(h_k/K)$. We use the complete empirical RBF-MMD, including source–source interactions and diagonal terms, with mean squared latent distances and the detached median bandwidth in Appendix B.6, floored at $\beta_{\min}$. Gradient accumulation preserves the $1/(Kd)$ local-loss reduction and the full MMD source derivative, including both occurrences of each sample in the source–source term.

Critic update and optimizer settings. Each generator optimizer step is followed by N_{critic} critic updates using detached joint student clips from a replay buffer $\mathcal{B}$ holding up to $N_{\mathcal{B}}$ entries. Each replay entry retains the clean and noisy joint clips, corruption noise, timestep, and conditional teacher prediction, so critic replay does not require a new generator rollout or teacher forward pass. The critic uses the flow-matching objective in Eq. (B.48) without ratio reweighting. Both networks are optimized with AdamW; learning rates, momentum coefficients, weight decay, batch sizes, and the training seed are listed in Table 3. We retain the baseline's mixed-precision training and gradient checkpointing, and update only the retained generator replay graph rather than backpropagating through the detached prefix.

Training duration and inference. Training runs for N_{train} generator updates, and hence $N_{\text{critic}}N_{\text{train}}$ critic updates, saving a checkpoint every N_{ckpt} generator updates. We maintain an exponential moving average (EMA) of the generator weights and use it for evaluation. MovieGen-100 validation generates full BF-frame latent videos with a fixed inference seed. At inference, only the four-step causal generator is deployed, with IID Gaussian sampling and no teacher, fake denoiser, ratio estimation, or MMD computation.

C.4 Training Algorithm

Algorithm 1 gives the default training loop. Assignments indexed by k apply to all K rollouts; score queries share (u_t, t, c) as in the main text. PrefixRollout generates the preceding $b-1$ chunks and their KV cache (Appendix C.1); Bandwidth uses the pooled-median rule in Appendix B.6 with floor $\beta_{\min}$; all hyperparameters are listed in Table 3.

Algorithm 1 Mass Forcing Training

Require: Initialized G_θ, calibrated residual R_ϕ; frozen D_p, D_{uncond}, with $D_q = D_p + R_\phi$.
Require: $B, F, K, \tau, \kappa, \lambda_{\max}, \gamma, w_{\text{cfg}}, E_{\text{lo}}, E_{\text{hi}}, N_{\text{ramp}}, N_{\text{critic}}, N_{\mathcal{B}}, N_{\text{train}}$ (Table 3).

Initialize generator EMA; $\mathcal{B} \leftarrow \emptyset$ ▷ Critic replay buffer, capacity $N_{\mathcal{B}}$
for $N_G = 0, \dots, N_{\text{train}} - 1$ **do**
  // Phase 1: On-policy references and mass estimation (no grad)
  Sample prompt c, $b \sim \text{Unif}\{1, \dots, B\}$; $L \leftarrow Fb$
  $u \leftarrow \text{sg}(\text{PrefixRollout}(G_\theta, c, b))$ ▷ Save prefix KV state
  Sample independent z_k; $\bar{x}_k \leftarrow \text{sg}(G_\theta(z_k; u, c))$ ▷ Record replay randomness
  Sample t from the full schedule; $\epsilon_u, \epsilon \sim \mathcal{N}(0, I)$ ▷ Shared across k
  $u_t \leftarrow \alpha_t u + \sigma_t \epsilon_u, \quad \bar{y}_k \leftarrow \alpha_t \bar{x}_k + \sigma_t \epsilon$
  $D_{p,k}, D_{q,k}, D_{\text{uncond},k} \leftarrow D_p(\bar{y}_k), D_q(\bar{y}_k), D_{\text{uncond}}(\bar{y}_k)$
  Retain $\bar{D}^{\text{joint}}_{p,k}$ from the same teacher query ▷ Detached full-clip prediction
  $D_{\text{cfg},k} \leftarrow D_{\text{uncond},k} + w_{\text{cfg}}(D_{p,k} - D_{\text{uncond},k})$
  $h_k \leftarrow 1, \quad \lambda_{\text{eff}} \leftarrow 0$
  if $\chi_{\text{cal}}\chi_{\text{cell}}(\sigma_t, L, b) = 1$ **then** ▷ Validated ratio cells
    $k^* \leftarrow \arg\min_j \sum_k \|\bar{y}_j - \bar{y}_k\|_2, \quad a \leftarrow \bar{y}_{k^*}$
    $\Delta_t(y) \leftarrow -\alpha_t R_\phi(y)/\sigma_t^2$ ▷ Reuse endpoints; residual-only midpoints
    $\ell_k \leftarrow \frac{1}{6}\left\langle \Delta_t(a) + 4\Delta_t\left(\frac{a+\bar{y}_k}{2}\right) + \Delta_t(\bar{y}_k), \bar{y}_k - a \right\rangle$ ▷ Simpson
    $\bar{\ell} \leftarrow K^{-1}\sum_k \ell_k, \quad v_k \leftarrow \text{clip}((\ell_k - \bar{\ell})/\tau, -\kappa, \kappa)$
    $\hat{r}_k \leftarrow K\exp(v_k)/\sum_j \exp(v_j)$
    $E \leftarrow \frac{(\sum_k \hat{r}_k)^2}{K\sum_k \hat{r}_k^2}, \quad \eta_{\text{ESS}} \leftarrow \text{clip}\left(\frac{E - E_{\text{lo}}}{E_{\text{hi}} - E_{\text{lo}}}, 0, 1\right)$
    $\lambda_{\text{eff}} \leftarrow \lambda_{\max}\min(N_G/N_{\text{ramp}}, 1)\eta_{\text{ESS}}, \quad h_k \leftarrow 1 + \lambda_{\text{eff}}(\hat{r}_k - 1)$
  end if
  $n_k \leftarrow \text{sg}\left[\max\left(d^{-1}\|\bar{x}_k - D_{\text{cfg},k}\|_1, \epsilon_n\right)\right]$
  $g_k \leftarrow \text{sg}\left(\frac{h_k(D_{q,k} - D_{p,k}) + D_{p,k} - D_{\text{cfg},k}}{n_k}\right), \quad \pi_k \leftarrow \text{sg}(h_k/K)$
  Cache $\{([u, \bar{x}_k], [u_t, \bar{y}_k], t, [\epsilon_u, \epsilon], c, \bar{D}^{\text{joint}}_{p,k})\}_{k=1}^K$ in $\mathcal{B}$

  // Phase 2: Generator update (references and weights fixed)
  for $k = 1, \dots, K$ **do**
    Restore prefix KV state; $x_k \leftarrow G_\theta(z_k; u, c)$ ▷ Same z_k; $x_k = \bar{x}_k$
    $y_k \leftarrow \alpha_t x_k + \sigma_t \epsilon$
  end for
  $\mu_\theta \leftarrow K^{-1}\sum_k \delta_{y_k}, \quad \nu \leftarrow \sum_k \pi_k \delta_{\bar{y}_k}$
  $\beta \leftarrow \text{sg}(\text{Bandwidth}(\{y_k, \bar{y}_k\}_{k=1}^K))$
  $\mathcal{L}_{\text{local}} \leftarrow \frac{1}{2Kd}\sum_k \|x_k - \text{sg}(x_k - g_k)\|_2^2$
  $\mathcal{L}_{\text{set}} \leftarrow \text{MMD}^2_{\mathcal{K}_\beta}(\mu_\theta, \nu)$
  Update θ using $\nabla_\theta(\mathcal{L}_{\text{local}} + \gamma\mathcal{L}_{\text{set}})$; update generator EMA

  // Phase 3: Fake-denoiser updates (unweighted replay)
  for $n = 1, \dots, N_{\text{critic}}$ **do**
    $(Y_0, Y_t, t, \epsilon_Y, c, \bar{D}^{\text{joint}}_p) \sim \mathcal{B}$ ▷ All entries detached
    $D^{\text{joint}}_q \leftarrow \bar{D}^{\text{joint}}_p + R^{\text{joint}}_\phi(Y_t, t, c)$ ▷ Reuse cached teacher output
    $\mathcal{L}_q \leftarrow \left\|(Y_t - D^{\text{joint}}_q)/\sigma_t - (\epsilon_Y - Y_0)\right\|^2_{\text{av}}$
    Update ϕ using $\nabla_\phi \mathcal{L}_q$
  end for
end for
return Trained generator G_θ (EMA weights)

Here $\bar{D}^{\text{joint}}_{p,k}$ is the detached full-clip conditional-teacher prediction retained from the endpoint query. The MMD gradient includes both source appearances; critic updates reuse cached inputs and teacher outputs without mass weighting or new teacher queries.

D Detailed Experimental Results

Tables 4–7 report dimension-level results using the VBench evaluation taxonomy (Huang et al., 2024). For the main comparisons, the methods, ordering, and model groups follow Table 1. Dimension-level scores are raw evaluator outputs multiplied by 100, whereas aggregate scores use the normalized scale; higher values indicate better performance and dashes mark unavailable results.

D.1 MovieGen

MovieGen provides text prompts but not the metadata required by VBench's semantic evaluators. We construct this metadata using Qwen3-VL-32B-Instruct. The annotation prompt and validation procedure are provided in Appendix D.1.1. No prompts in the evaluated MovieGen set were identified as satisfying VBench's criteria for Spatial Relationship, so we exclude this dimension rather than assigning it a zero score. For the remaining eight dimensions, we retain VBench's standard per-dimension normalization and relative weights, then renormalize these weights to sum to one. Since VBench assigns equal weight to each semantic dimension, the resulting Semantic score is the mean of the eight normalized scores. All compared methods and ablation variants use the same metadata, dimension set, and aggregation procedure, ensuring a consistent within-benchmark comparison. Quality aggregation is unchanged, and the Total score remains 0.8 Quality plus 0.2 Semantic.

Table 4: Detailed quality evaluation on MovieGen prompts. Scores are shown as percentages (↑).

Model	**Subject Consist.**	**Background Consist.**	**Temporal Flickering**	**Motion Smoothness**	**Dynamic Degree**	**Aesthetic Quality**	**Imaging Quality**	**Quality Score**
Bidirectional Video Diffusion Models								
LTX-Video	97.30	96.67	98.47	99.35	31.90	62.75	70.36	82.44
Wan2.1	94.29	96.12	97.16	98.31	59.02	65.15	67.61	82.73
AVD	94.20	94.13	95.83	98.32	61.11	62.91	67.41	81.54
Autoregressive Video Diffusion Models								
NOVA	92.24	93.99	98.26	99.01	50.45	56.05	52.57	78.36
Pyramid Flow	94.20	95.33	98.56	99.30	34.40	64.09	64.24	81.07
SkyReels-V2	96.23	96.94	97.71	98.77	50.25	64.86	64.43	82.51
MAGI-1	97.80	97.66	99.01	99.45	22.13	65.19	64.93	81.81
Distilled Autoregressive Video Models								
CausVid	94.74	94.29	96.88	98.06	57.33	62.79	67.81	81.72
Self Forcing	94.17	94.85	96.75	98.32	63.41	63.88	69.84	82.77
Causal Forcing	94.37	94.92	95.52	97.64	77.57	64.42	70.87	83.28
Causal-rCM	95.93	95.88	97.55	98.64	48.26	65.35	69.50	82.80
Mass Forcing (Ours)	95.10	95.02	95.94	98.41	83.35	65.92	69.87	84.53

Note. Consist. denotes consistency.

Table 5: Detailed semantic evaluation on MovieGen prompts. Scores are shown as percentages (↑).

Model	Object Class	Multiple Objects	Human Action	Color	Scene	Appear. Style	Temp. Style	Overall Consist.	Semantic Score
Bidirectional Video Diffusion Models									
LTX-Video	83.75	55.77	76.92	84.68	60.66	21.58	23.16	23.15	72.21
Wan2.1	84.20	51.92	100.0	93.43	58.48	22.13	22.44	26.02	76.41
AVD	84.43	56.25	100.0	90.10	47.37	22.07	23.14	24.86	74.69
Autoregressive Video Diffusion Models									
NOVA	82.39	59.13	84.62	83.34	50.24	23.06	23.15	25.57	73.14
Pyramid Flow	89.03	54.81	100.0	87.54	60.24	22.64	20.67	25.20	76.24
SkyReels-V2	82.61	54.33	92.31	91.67	52.96	21.88	20.96	24.47	73.33
MAGI-1	87.50	65.87	100.0	87.79	58.09	23.03	22.34	25.41	77.95
Distilled Autoregressive Video Models									
CausVid	85.57	75.96	92.31	87.69	51.27	22.00	20.77	24.40	75.62
Self Forcing	86.36	65.87	92.31	92.50	60.54	21.87	23.28	25.66	77.70
Causal Forcing	88.69	71.63	100.0	95.63	59.78	21.97	21.79	25.45	79.42
Causal-rCM	89.55	61.54	92.31	93.25	62.71	21.67	21.61	25.31	77.21
Mass Forcing (Ours)	85.63	67.79	92.31	96.39	63.53	21.98	22.97	25.79	78.78

Note. Appear., Temp., and Consist. denote appearance, temporal, and consistency.

D.1.1 SEMANTIC-METADATA ANNOTATION PROMPT

For each MovieGen record, the annotation model receives the expanded English prompt used for video generation as its sole semantic ground truth, without rewriting it at annotation time. We reproduce the system instruction and per-record user message below. The official vocabularies of 79 VBench object classes, 86 VBench scene labels, and 400 Kinetics-400 action labels are appended to the system instruction at runtime. Responses are parsed as JSON and checked against these vocabularies and the prompt text; unsupported or ambiguous fields are discarded before conversion to VBench evaluator queries.

System instruction.

```
You are a strict VBench semantic-metadata annotator. You receive one
already-expanded English prompt that will be used verbatim to generate
a video. Treat that expanded prompt as the only ground truth. Extract
only semantic targets explicitly supported by its text.

Return exactly one JSON object and no markdown. Required keys:
{
  "primary_semantic_dimension": one of ["object_class",
  "multiple_objects", "human_action", "color", "spatial_relationship",
  "scene", "appearance_style", "temporal_style",
  "overall_consistency"],
  "primary_feature": a concise English feature string directly
  supported by the prompt,
  "confidence": a number in [0, 1],
  "semantic_metadata": {
    "object_class": {"object": "single common detector class noun"} or
    null,
    "multiple_objects": {"object_a": "common detector class noun",
    "object_b": "different common detector class noun"} or null,
    "human_action": {"kinetics_400_label": "exact label from the
    provided allowlist"} or null,
    "color": {"object": "common detector class noun", "color": "one
    allowed color"} or null,
    "spatial_relationship": {"object_a": "common detector class noun",
    "object_b": "different common detector class noun", "relationship":
    "one allowed relationship"} or null,
    "scene": {"scene": "short literal setting noun phrase"} or null,
    "appearance_style": {"appearance_style": "explicit visual style
    phrase"} or null,
    "temporal_style": {"temporal_style": "explicit temporal style
    phrase"} or null
  },
  "rationale": one short sentence.
}

Rules:
1. Always choose one primary_semantic_dimension. Use
overall_consistency when no narrower semantic dimension is reliable.
overall_consistency is implicit and must not appear inside
semantic_metadata.
2. Do not infer or embellish anything beyond the supplied expanded
prompt. Use null whenever a field is absent, ambiguous, merely implied,
or cannot satisfy the exact schema.
3. object_class must be one exact label from the supplied official
VBench object-class allowlist and must denote a visibly present main
subject. Map men/women/children to "person", planes/jets to "airplane",
and other subclasses only when the mapping is unambiguous. Do not use a
label absent from the allowlist.
4. multiple_objects requires two different labels from that same
object-class allowlist, both explicitly visible. Do not use two
instances of one class, scenery, an action, or an abstract concept.
```

```
5. color is valid only when the prompt explicitly binds exactly one of
[black, blue, gray, green, grey, orange, pink, purple, red, silver,
white, yellow] to an object-class allowlist label. Return that label
separately.
6. spatial_relationship requires two different object-class allowlist
labels and an explicit relation exactly equal to one of [on the left
of, on the right of, on the top of, on the bottom of]. Do not translate
inside/behind/beside/through/above/below/reflecting/over/under into one
of these relations.
7. scene must be one exact label from the supplied official VBench
scene allowlist and must be explicitly supported by the prompt. Do not
return an inferred setting or a longer description.
8. appearance_style is valid only for an explicit rendering or visual
style, such as photorealistic, watercolor, anime, oil painting,
stop-motion, or cinematic. Mood, clothing, image quality, camera angle,
lighting, and colorfulness alone are not appearance styles.
9. temporal_style is valid only for an explicit temporal presentation
such as time-lapse, slow motion, stop motion, reverse motion, looping,
or step-printing. Ordinary actions and camera motion are not temporal
styles.
10. human_action is valid only when every content word of one complete
Kinetics-400 label appears in order in the supplied prompt (articles
may differ) and describes a visible person. Do not choose a merely
similar label: ordinary walking is not "walking the dog", and ordinary
drinking is not "tasting beer". Copy the exact lowercase allowlist
label or return null.
11. Keep all values in English. Never include a metadata key not listed
in the schema.
```

Per-record user message. The placeholder {EXTENDED_PROMPT} is replaced verbatim with the corresponding English generation prompt.

```
Expanded MovieGen generation prompt (the sole metadata ground truth):
{EXTENDED_PROMPT}

Extract VBench semantic metadata now and return exactly the required
JSON object.
```

D.2 VBENCH

Table 6: Detailed quality evaluation on VBench prompts. Scores are shown as percentages (↑).

Model	Subject Consist.	Background Consist.	Temporal Flickering	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality	Quality Score
Bidirectional Video Diffusion Models								
LTX-Video	96.46	97.17	99.45	99.33	43.06	65.58	68.34	83.78
Wan2.1	96.34	98.03	99.29	98.43	62.50	68.79	67.02	85.17
AVD	96.21	94.77	96.52	98.40	73.61	65.03	66.54	83.51
Autoregressive Video Diffusion Models								
NOVA	94.25	95.44	99.18	99.23	44.44	59.43	55.01	79.96
Pyramid Flow	93.78	96.66	99.30	99.24	59.72	65.30	63.08	83.51
SkyReels-V2	97.35	98.24	99.48	98.94	63.89	66.07	63.29	84.86
MAGI-1	97.88	97.91	99.79	99.50	16.67	64.62	62.75	81.38
Distilled Autoregressive Video Models								
CausVid	95.87	95.67	99.26	98.02	65.28	64.84	66.64	83.92
Self Forcing	96.15	95.89	98.98	98.62	61.11	65.90	69.53	84.51
Causal Forcing	95.01	95.46	97.95	97.59	86.11	66.49	69.28	85.22
Causal-rCM	95.95	96.55	99.24	98.36	69.44	66.24	68.38	85.10
Mass Forcing (Ours)	97.56	96.17	97.18	98.97	88.89	67.08	66.93	86.17

Note. Consist. denotes consistency.

Table 7: Detailed semantic evaluation on VBench prompts. Scores are shown as percentages (↑).

Model	Object Class	Multiple Objects	Human Action	Color	Spatial Rel.	Scene	Appear. Style	Temp. Style	Overall Consist.	Semantic Score
Bidirectional Video Diffusion Models										
LTX-Video	93.28	66.01	94.00	75.06	62.46	48.62	20.73	24.87	25.47	73.42
Wan2.1	92.88	79.80	95.00	83.49	71.82	49.78	22.40	26.85	27.53	79.04
AVD	90.27	70.88	96.00	81.53	66.49	50.29	21.97	26.23	26.06	76.32
Autoregressive Video Diffusion Models										
NOVA	88.05	70.88	95.00	77.00	65.94	48.76	23.44	26.47	26.94	76.11
Pyramid Flow	88.77	65.40	94.00	82.48	67.56	51.96	22.29	26.29	27.05	76.21
SkyReels-V2	87.58	59.22	95.00	87.71	54.76	54.65	21.31	25.54	26.30	74.19
MAGI-1	90.66	61.20	96.00	87.01	55.50	59.74	23.43	25.54	26.99	76.59
Distilled Autoregressive Video Models										
CausVid	88.61	83.23	97.00	86.86	60.33	49.71	21.14	25.62	26.64	77.11
Self Forcing	93.04	80.03	97.00	84.21	68.68	55.23	21.72	26.54	27.51	79.40
Causal Forcing	92.64	85.82	95.00	86.81	74.61	50.15	21.55	26.57	27.01	79.83
Causal-rCM	93.91	83.92	94.00	82.85	74.83	47.24	21.67	26.40	27.03	78.84
Mass Forcing (Ours)	93.20	83.08	99.00	88.68	71.42	53.27	21.62	25.94	27.30	80.23

Note. Appear., Temp., and Consist. denote appearance, temporal, and consistency; Spatial Rel. denotes spatial relationship.

D.3 Ablation Studies

Tables 8–11 provide the dimension-level results for Table 2, together with the Quality, Semantic, and Total scores. The Ours rows repeat the reference results from the main comparison tables above. The w/o CFG decoupling variant omits both CFG decoupling and MMD, whereas w/o MMD retains CFG decoupling but removes set-level supervision. These variants form a progression from coupled local weighting, through decoupled local weighting, to the full local-and-set objective; the gap between w/o CFG decoupling and Ours therefore includes both changes. The MovieGen and VBench evaluations contain 1,003 and 946 videos per ablation variant, respectively. MovieGen semantic results omit Spatial Relationship and aggregate the eight available dimensions; VBench uses all nine semantic dimensions.

D.3.1 MovieGen

Table 8: Detailed quality ablation results on MovieGen prompts (0–100 scale; ↑).

Variant	Subject Consist.	Background Consist.	Temporal Flickering	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality	Quality Score
Ours	95.10	95.02	95.94	98.41	83.35	65.92	69.87	84.53
w/o CFG decoupling	95.90	95.53	96.82	98.53	61.81	65.91	69.05	83.42
w/o MMD	95.69	95.12	96.00	98.56	65.00	65.91	69.86	83.35
Shuffled-ratio control	95.72	95.18	96.35	98.72	59.62	66.16	69.88	83.22
$K = 2$	95.09	94.99	95.57	98.52	74.78	65.38	68.21	83.43
Endpoint-only	95.19	94.94	95.54	98.34	74.68	66.19	69.14	83.59
$\lambda_{\max} = 0.5$	95.24	94.94	96.34	98.45	64.31	65.73	69.78	83.22
$\lambda_{\max} = 1.0$	95.00	94.87	96.22	98.48	72.98	65.44	69.31	83.67

Table 9: Detailed semantic ablation results on MovieGen prompts (0–100 scale; ↑).

Variant	Object Class	Multiple Objects	Human Action	Color	Scene	Appear. Style	Temp. Style	Overall Consist.	Semantic Score	Total Score
Ours	85.63	67.79	92.31	96.39	63.53	21.98	22.97	25.79	78.78	83.38
w/o CFG decoupling	86.25	67.79	92.31	80.11	56.52	21.84	22.02	25.41	75.24	81.79
w/o MMD	86.76	72.60	92.31	96.59	58.09	21.90	22.81	25.67	78.59	82.40
Shuffled-ratio control	86.08	75.48	92.31	94.88	52.92	22.02	21.58	25.63	77.48	82.07
$K = 2$	86.82	75.96	100.00	86.53	53.53	22.16	22.32	25.45	77.90	82.32
Endpoint-only	86.65	62.50	92.31	93.18	61.59	22.01	23.19	25.76	77.63	82.40
$\lambda_{\max} = 0.5$	86.14	68.27	100.00	80.63	58.42	21.97	22.69	25.66	76.97	81.97
$\lambda_{\max} = 1.0$	84.60	66.83	100.00	91.13	61.26	21.94	22.88	25.60	78.38	82.61

D.3.2 VBENCH

Table 10: Detailed quality ablation results on VBench prompts (0–100 scale; ↑).

Variant	Subject Consist.	Background Consist.	Temporal Flickering	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality	Quality Score
Ours	97.56	96.17	97.18	98.97	88.89	67.08	66.93	86.17
w/o CFG decoupling	97.11	95.89	98.30	98.69	65.28	66.90	68.60	84.77
w/o MMD	97.13	95.21	96.96	98.66	77.78	67.44	69.21	85.19
Shuffled-ratio control	96.91	95.38	97.08	98.77	69.44	67.79	69.03	84.68
$K = 2$	96.61	95.16	96.71	98.71	87.50	66.13	67.49	85.29
Endpoint-only	96.48	95.16	96.56	98.45	86.11	67.56	68.53	85.34
$\lambda_{\max} = 0.5$	96.57	95.61	97.58	98.62	62.50	67.54	69.25	84.25
$\lambda_{\max} = 1.0$	96.47	95.12	97.45	98.67	79.17	66.77	68.81	85.21

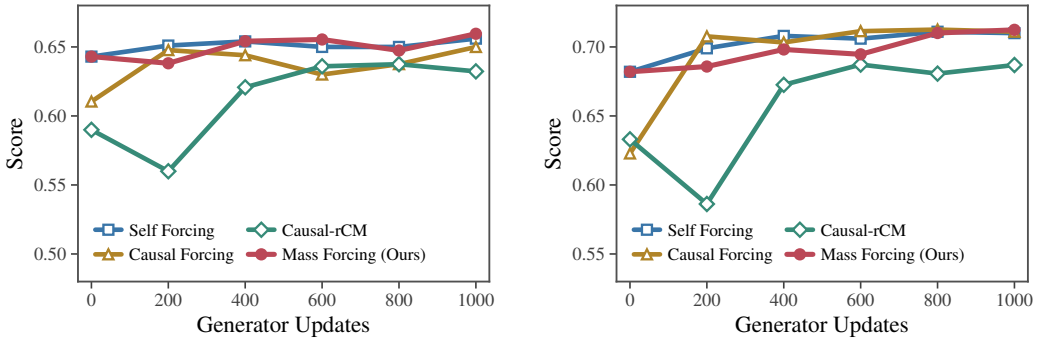
Table 11: Detailed semantic ablation results on VBench prompts (0–100 scale; ↑).

Variant	Object Class	Multiple Objects	Human Action	Color	Spatial Rel.	Scene	Appear. Style	Temp. Style	Overall Consist.	Semantic Score	Total Score
Ours	93.20	83.08	99.00	88.68	71.42	53.27	21.62	25.94	27.30	80.23	84.98
w/o CFG decoupling	97.63	83.08	96.00	83.98	69.86	56.03	21.51	26.48	27.20	80.16	83.85
w/o MMD	95.57	78.96	96.00	83.50	68.78	47.38	21.50	26.65	27.51	78.27	83.81
Shuffled-ratio control	94.46	82.24	97.00	83.48	71.65	53.20	21.89	26.56	27.64	79.89	83.72
$K = 2$	91.30	79.73	96.00	84.21	64.76	48.98	21.82	25.97	26.95	77.48	83.73
Endpoint-only	93.75	83.16	97.00	84.42	74.09	51.74	21.93	26.68	27.76	80.18	84.31
$\lambda_{\max} = 0.5$	93.51	82.91	96.00	84.42	73.50	53.20	21.87	26.56	27.48	80.00	83.40
$\lambda_{\max} = 1.0$	93.12	79.65	98.00	84.59	67.05	53.63	21.84	26.61	27.43	79.17	84.00

D.4 Visual Quality During DMD Training

The loss of temporal dynamics in the baseline trajectories need not coincide with declining visual quality. Figure 5 compares Aesthetic Quality and Imaging Quality for Self Forcing, Causal Forcing, Causal-rCM, and Mass Forcing at generator updates $0, 200, 400, 600, 800$, and $1{,}000$. For Self Forcing, Aesthetic Quality increases from 0.643 to 0.656, while Imaging Quality increases from 0.682 to 0.710. Causal Forcing and Causal-rCM also finish above their initial scores in both metrics, although Causal-rCM exhibits an early decline before recovering. These improvements are not monotonic and coexist with the late-training decline in baseline Dynamic Degree shown in Figure 1.

At update $1{,}000$, Mass Forcing reaches 0.660 Aesthetic Quality and 0.712 Imaging Quality, the highest final-checkpoint scores among the four plotted methods, while retaining 83% Dynamic Degree (Figure 1). The baseline trajectories expose a checkpoint-selection trade-off between visual quality and dynamics, whereas Mass Forcing sustains high scores on both measures at the end of training.



(a) Aesthetic Quality

(b) Imaging Quality

Figure 5: **Visual quality during DMD training.** Aesthetic and Imaging Quality for Self Forcing, Causal Forcing, Causal-rCM, and Mass Forcing. Markers show checkpoint scores on the original 0–1 scale; lines connect adjacent observations. Self Forcing and Mass Forcing share the same ODE initialization at update 0.

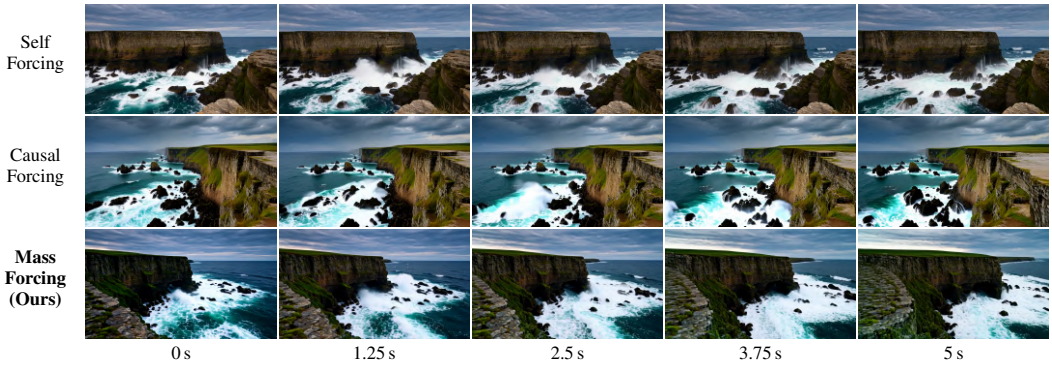
E Extended Qualitative Comparisons

Figures 6 and 7 compare prompt alignment and temporal variation, with full prompts and key requirements highlighted in bold. Rows show Self Forcing, Causal Forcing, and Mass Forcing; columns correspond to 0, 1.25, 2.5, 3.75, and 5 seconds.

In Figure 6, Mass Forcing shows more pronounced changes in the waves and foam, making the requested crashing surf more evident across time. For the dewdrop, it renders a larger, more detailed miniature world of reflected foliage within the droplet and develops the close-up across frames, whereas the baselines retain smaller reflections and show more limited compositional change.

In Figure 7, the dragon exhibits clearer changes in head pose and framing while retaining detailed silver-edged scales. The listening man shows more expressive changes in his smile, gaze, and upper-body posture, better conveying the requested amused chuckle than the comparatively restrained baseline sequences. The illustrated boy provides a stylized example with visible changes in expression and gaze, while retaining the clasped hands and lifted-foot pose specified in the prompt.

Prompt. *A dramatic landscape photograph capturing a steep cliffside with a stone pedestal jutting out, leading down to reveal the* ***powerful waves crashing against jagged rocks*** *far below. The cliffs are rugged and weathered, with green moss covering parts of the rock faces. The water is turbulent, with white foam rising from the impact of the waves against the rocks. The sky above is a mix of deep blues and grays, creating a moody and atmospheric scene. The photo has a natural and realistic texture, emphasizing the* ***dynamic movement of the waves*** *and the dramatic drop-off. A wide-angle shot from a low angle, looking up towards the cliff.*



Prompt. *A close-up zoom-in on a morning dewdrop perched on a leaf, capturing the* ***intricate reflection of its surroundings within its translucent surface****. The dewdrop glistens under the morning sunlight, revealing* ***a miniature world of shimmering leaves, distant flowers, and faint shadows****. The leaf itself is crisp and green, with visible veins and edges. The background is a blurred landscape, hinting at a dewy meadow bathed in early morning light. The dewdrop sparkles with a soft, ethereal glow, emphasizing the beauty of nature's smallest wonders. A macro shot from a slightly tilted angle.*

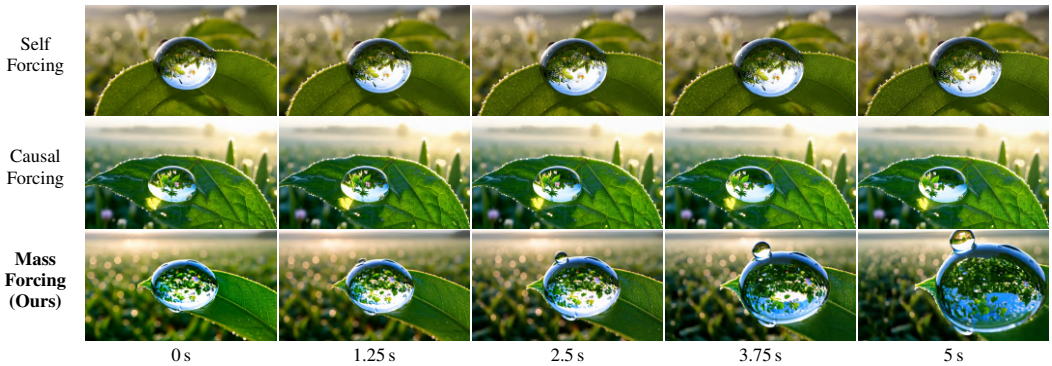
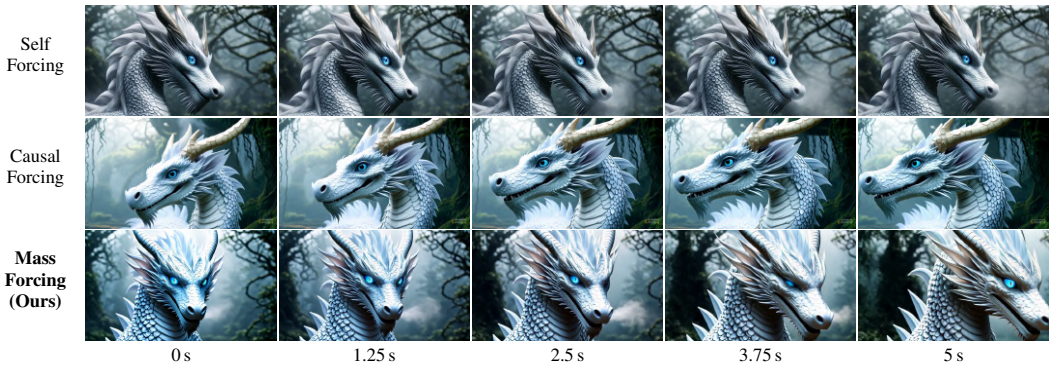


Figure 6: **Scene dynamics and fine-grained prompt details.** Top: waves breaking against a cliff. Bottom: a dewdrop containing a miniature reflected world.

Prompt. *A close-up shot of a majestic white dragon with **pearlescent, silver-edged scales**, icy blue eyes, and elegant ivory horns. The dragon's face is detailed with subtle wrinkles and sharp, defined features, capturing a regal and serene expression. Its **breath forms a gentle mist**, adding to the ethereal quality. The scales are meticulously textured, reflecting light in a way that highlights their depth and shine. Set against a softly blurred background, the scene is bathed in a soft, ambient glow, emphasizing the dragon's majesty and otherworldly presence. The background hints at a misty forest, with blurred outlines of ancient trees and vines, creating a mystical atmosphere.*



Prompt. *A middle-aged man with a **kind smile and amused chuckle** listens intently to a funny story being told. He has short graying hair and wears a casual blue shirt with a pair of khaki pants. His hands are loosely clasped together, and he **leans slightly forward** in his comfortable armchair. The background shows a cozy living room with warm lighting, books on a nearby shelf, and a small potted plant on the windowsill. The scene captures the moment when he is fully engaged, with a soft focus on his face and a slight tilt of his head. A close-up shot from a slightly elevated angle.*



Prompt. *A hand-drawn simple line art illustration of a young boy with a **look of wonder and amazement** on his face, gazing up at the sky. He has curly brown hair and bright blue eyes that sparkle with curiosity. His small hands are clasped together in front of him, and he stands on a grassy hill, **one foot slightly lifted**. The background features a clear blue sky with fluffy clouds and distant mountains, creating a serene and peaceful atmosphere. A close-up shot from a slightly lower angle, capturing the child's innocent and awe-filled expression.*

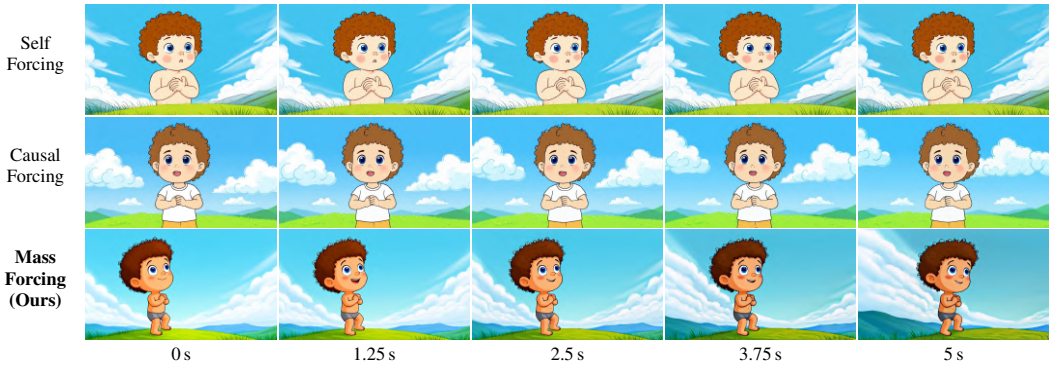


Figure 7: **Character appearance and motion.** Top: a white dragon with detailed scales. Middle: a man reacting to a humorous story. Bottom: an illustrated boy gazing at the sky.

F Qualitative Ablation

Figure 8 compares the eight variants in Table 2 on the same prompt, with five frames per video at matched timestamps. The labels follow the quantitative ablations: w/o CFG decoupling also omits MMD, w/o MMD retains decoupling, and Ours includes both components.

Mass Forcing makes the requested running action visually explicit: the elephant exhibits extended, alternating leg poses, forward body motion, and a changing dust trail while the herd remains visible in the background. The side-on view keeps this locomotion legible across the sequence, matching the prompt's emphasis on an elephant running in the context of its herd. Without MMD, the elephant increasingly turns toward and approaches the camera, making its relation to the distant herd less prominent; without both components, the sequence depicts an elephant already among the herd with more restrained changes in pose. The other variants also produce motion, but differ in framing, stride visibility, and depiction of the group. This example complements the aggregate results by showing how the full method combines visible subject motion with the scene content specified by the prompt.

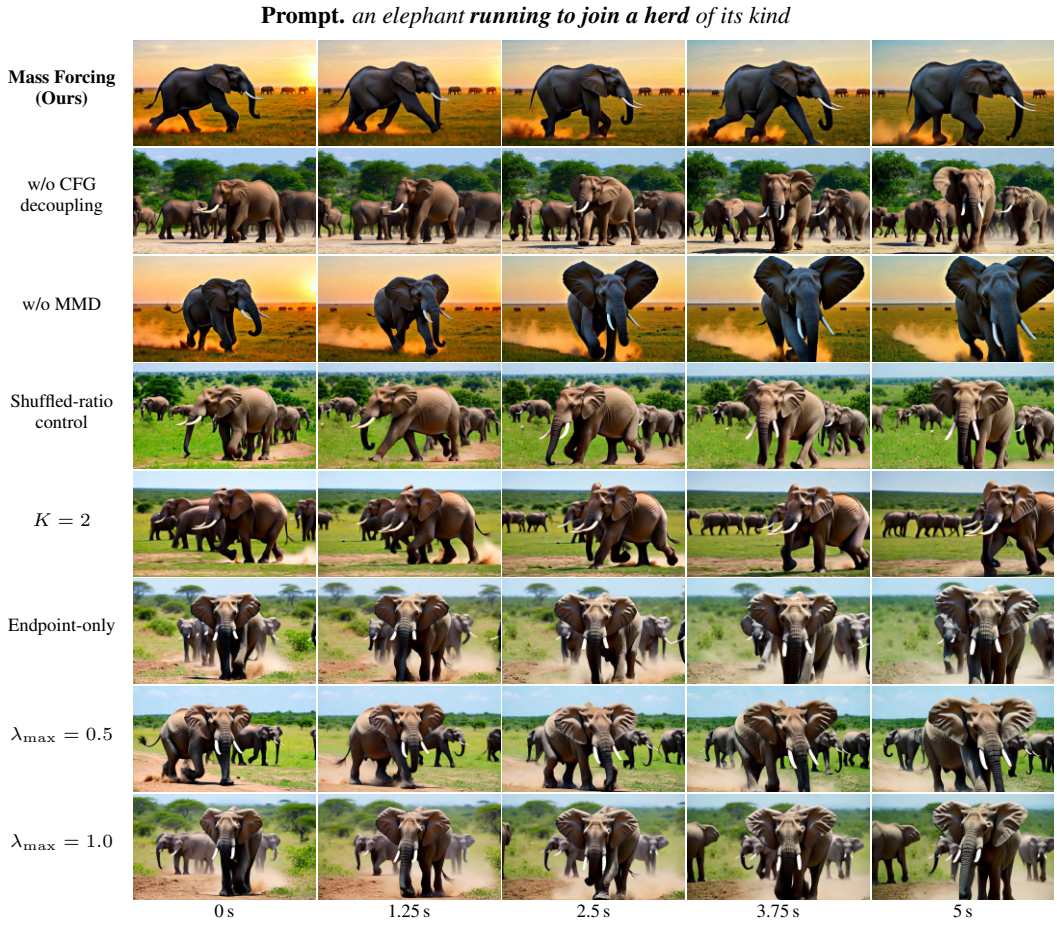


Figure 8: **Qualitative comparison of the ablation variants.** Rows follow Table 2; columns show matched timestamps. Ours renders a pronounced running stride and evolving dust trail with the herd visible in the scene. The w/o CFG decoupling variant also omits MMD.