DUOMATCHING: JOINT–MARGINAL DISTRIBUTION MATCHING FOR FEW-STEP VIDEO GENERATION

Anonymous authors
Paper under double-blind review

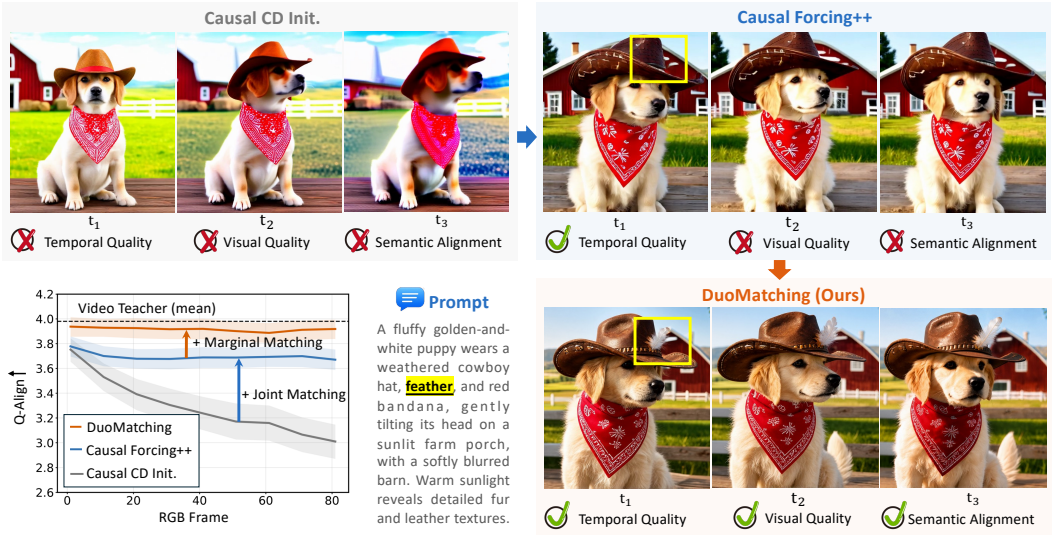

Figure 1: Motivation and results of DuoMatching. Videos generated by autoregressive models often have degraded quality for later rollout frames (see Q-Align (Wu et al., 2023) curves). Joint DMD mitigates this drift, yet the student still retains a visual quality gap relative to the video teacher. The example reveals a missing feather specified in the prompt. To solve this problem, we reformulate distribution matching from a coupled joint-marginal perspective to better approximate the real video distribution, as illustrated by improved visual quality and the recovered feather, while maintaining temporal coherence. We show video results in `index.html` in our supplementary materials.

ABSTRACT

Streaming video generation has benefited from distribution matching distillation (DMD), which matches the joint distribution of video frames to a video teacher's approximation of the real video distribution. Although this joint matching mitigates drift during autoregressive rollouts, limitations remain in visual quality and semantic alignment. To address these limitations, we propose DuoMatching, a distribution matching framework that approximates the real video distribution through a unified joint-marginal formulation. On top of existing joint matching formulations, the additional marginal matching objective provides dedicated frame-level supervision from an image generator, transferring complementary visual and semantic priors from it. To apply this frame-level supervision in video generation, we introduce LatentBridge to resolve the latent representation mismatch between the video student and the image teacher. Latent Variation Sampling further distributes such frame-level supervision across distinct temporal segments, reducing redundancy. Experiments demonstrate that DuoMatching improves visual quality, composition, and semantic alignment while largely preserving motion dynamics. Human evaluations show overall preference rates above 80% against all evaluated baselines.

1 INTRODUCTION

Recent video generation models have achieved remarkable advances in visual quality and temporal coherence (Liu et al., 2024; Yang et al., 2025b). Despite their impressive performance, these models

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.

typically require many sequential denoising steps, resulting in substantial computational cost and inference latency that hinder real-time applications such as interactive world simulation (Parker-Holder et al., 2025) and digital entertainment (Zhao et al., 2026b). Recently, distribution matching distillation (DMD) (Yin et al., 2024b) addresses this computational bottleneck by distilling multi-step diffusion (Ho et al., 2020) or flow models (Lipman et al., 2022) into few-step generators. Although streaming video generation built on this paradigm has approached real-time inference (Sun et al., 2025; Yang et al., 2025a; Liu et al., 2026), preserving generation quality under distribution matching remains a challenge (Feng et al., 2026).

Specifically, we identify remaining limitations in visual quality and semantic alignment of autoregressive generators trained with existing DMD methods. These methods use the video teacher's distribution as a proxy for the real video distribution (Yin et al., 2024b) and match the joint distribution of all frames, a formulation we refer to as joint DMD. As shown in Figure 1, the quantitative results indicate that joint DMD substantially mitigates drift over the rollout, but yields only limited improvement in initial-frame quality. The qualitative comparison highlights remaining limitations in visual quality and semantic alignment, as illustrated by insufficient fine-grained fur detail and the missing feather in this example. These frame-level deficiencies suggest that existing DMD methods do not adequately model the marginal distribution, i.e., the distribution of individual frames. Indeed, existing DMD matches marginals jointly with cross-frame dependencies, without a dedicated objective for individual frames. Adding an additional marginal distribution constraint in training may better approximate the real video distribution.

To model the marginal distribution in video generation, we employ image generation models (Seedream et al., 2025; Cai et al., 2025) which provide a natural frame-level reference distribution. Their strong priors support fine-grained appearance modeling, and their semantic alignment capabilities also help capture prompt-specified attributes and spatial relationships. Marginal distribution matching directly steers the student's frame-level distribution toward the image teacher's distribution, without being constrained by cross-frame dependencies. As illustrated in Figure 2, the marginal distribution matching's signal exhibits more coherent spatial structure, with particularly strong responses on the subject's overly smooth black clothing. Together with the video teacher's joint distribution, the additional marginal distribution from image teachers guides the student to generate more realistic video.

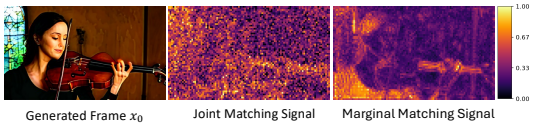

Figure 2: Spatial structure of score differences in joint and marginal distribution matching. Marginal matching exhibits more spatially coherent responses, whereas joint matching produces more scattered responses in this example. Each heatmap is independently normalized to highlight spatial structure rather than absolute magnitude.

We propose DuoMatching, a distribution matching framework that approximates the real video distribution through a unified joint-marginal formulation. For few-step video generation, we instantiate this formulation by training a video generator with a weighted combination of joint DMD and marginal DMD. Specifically, we sample a subset of frames from each generated video and apply marginal DMD independently to each frame using the image teacher, transferring visual and semantic priors to strengthen frame-level supervision. Joint DMD supervises the full sequence using the video teacher, constraining cross-frame dependencies to encourage temporal consistency.

Still, applying marginal distribution matching to a video student is nontrivial. Applying marginal DMD requires frame-level representations compatible with the image teacher. However, temporally compressed video latents do not directly correspond to individual frames, and video and image VAEs may use incompatible latent spaces. Therefore, we propose **LatentBridge**, a lightweight module that resolves the latent representation mismatch, enabling frame-level distribution matching without substantially suppressing encoded dynamics. We further introduce **Latent Variation Sampling** (LVS) to distribute marginal matching across distinct temporal segments, reducing redundant supervision to better balance visual quality and video dynamics.

Experiments demonstrate that DuoMatching improves fine-grained visual quality, visual composition, and semantic alignment not only in causal generation but also in bidirectional generation, while largely preserving temporal quality. Human evaluations further show overall preference rates above 80% against all evaluated baselines.

2 Related Work

Few-step video generation. Recent video generation models (Liu et al., 2024; Seedance et al., 2026; Wiedemer et al., 2025) achieve strong visual quality but rely on expensive iterative denoising, motivating the development of few-step generation methods. Some works adopt adversarial post-training to enable one-step video generation (Lin et al., 2025), while others reduce sampling steps through model distillation (Li et al., 2024; Ding et al., 2025). VideoLCM (Wang et al., 2023a) extends consistency distillation to video generation. CausVid (Yin et al., 2025) uses DMD (Yin et al., 2024a) for autoregressive video generation, enabling KV-cache reuse. Subsequent works address different limitations of causal distillation: Self Forcing (Huang et al., 2026) reduces the mismatch between training and self-generated inference rollouts, while Causal Forcing (Zhu et al., 2026) addresses the architectural mismatch to ensure frame-level injectivity. DuoMatching instead investigates the persistent quality limitations of distribution matching itself to improve the generation quality of few-step generators.

Distribution matching. Distribution matching provides a general framework for generative modeling by aligning the generated distribution with the real data distribution (Yin et al., 2024a;b; Wang et al., 2023b). However, the matching process can still suffer from optimization instability. Phased DMD (Fan et al., 2026) eases multi-step optimization by decomposing the denoising process into smaller SNR intervals. Beyond improving the optimization itself, several works introduce additional supervision to enhance the matched distribution. One-Forcing (Feng et al., 2026) adds adversarial supervision to alleviate frame blurring at extremely low NFE, while Reward Forcing (Lu et al., 2026) and DMDR (Jiang et al., 2025) incorporate reward signals to favor preferred outputs. Data-Forcing (Chen et al., 2026) leverages real data to mitigate the mode-seeking behavior of DMD, whereas SA-DMD (Xiao et al., 2026) anchors distribution matching to the source video for better content preservation in video editing. DuoMatching complements joint distribution matching with direct marginal distribution matching to more stably approximate the real video distribution.

Image priors for video generation. Image priors support video generation through image animation (Tian et al., 2021; Guo et al., 2024), and initialization from pretrained image models (Singer et al., 2022; Blattmann et al., 2023; Bai et al., 2026). Beyond initialization, image generation models have been integrated into video sampling to improve visual quality (Zhang et al., 2025; Shao et al., 2025). Some works have also explored incorporating image priors through distribution matching to accelerate video generation, focusing on image-derived video models such as AnimateDiff (Guo et al., 2023). MCM (Zhai et al., 2024) introduces a discriminator trained on image data during consistency distillation, while AVDM2 (Zhu et al., 2024) incorporates score distribution matching using an image teacher to enhance video adversarial training. We revisit distribution matching from a joint-marginal perspective, theoretically analyze the relationship between joint and marginal matching, and formulate both as distribution matching objectives within DuoMatching. We further introduce LatentBridge to generalize this framework to modern video generators with temporally compressed latents.

3 Method

We first formulate video generation as distribution matching, identify the limitations of relying solely on joint matching, and motivate explicit frame-level marginal matching in Section 3.1. In Section 3.2, we reformulate distribution matching through a unified joint-marginal framework to better approximate the real video distribution. To enable marginal distribution matching with image generation models, we introduce LatentBridge to resolve the mismatch between video and image latent spaces in Section 3.3. Finally, Section 3.4 introduces Latent Variation Sampling (LVS) to improve temporal coverage and reduce redundancy in marginal matching.

3.1 Distribution Matching for Video Generation

Video generation aims to learn a generator G_θ whose output distribution Q_θ matches the real data distribution $P^\star$. For simplicity, we omit text conditioning from the notation. The ideal distribution matching objective is

$$\mathcal{J}_{\mathrm{DM}}(\theta) = D_{\mathrm{KL}}(Q_\theta \,\|\, P^\star)\,. \tag{1}$$

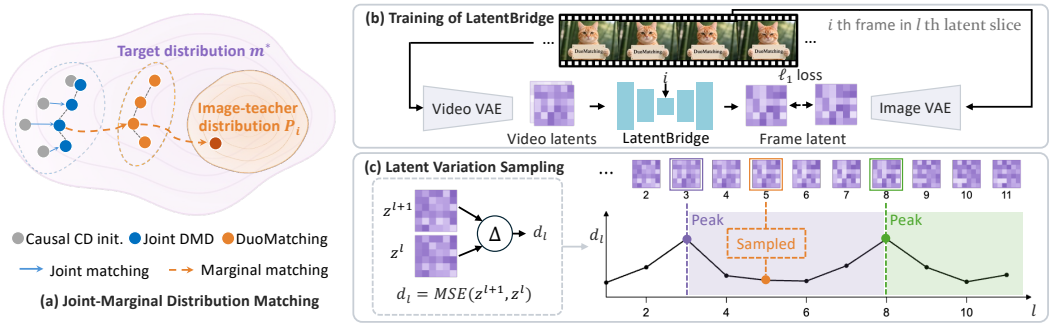

Figure 3: **Overview of DuoMatching.** (a) DuoMatching complements joint distribution matching with an image-teacher reference distribution P_i. Each group schematically represents a video, with nodes depicting frame-level distributions. (b) LatentBridge is trained to map a video latent z^l, conditioned on the preceding slice z^{l-1} and a local frame index i, to the corresponding image latent using an ℓ_1 reconstruction loss. (c) Latent Variation Sampling uses differences between adjacent latent slices to partition the sequence into segments, within which slices are randomly sampled.

Since the score of $P^\star$ is not directly available, we follow prior work and approximate it using a pre-trained video teacher with distribution P_v, which leads to the joint distribution matching objective:

$$\mathcal{J}_{\text{joint}} = D_{\text{KL}}(Q_\theta \,\|\, P_v)\,. \tag{2}$$

Joint distribution matching is performed over all temporal latent slices. In practice of joint DMD, given a generated video latent $\mathbf{z} = G_\theta(\boldsymbol{\xi})$, its noisy version is $\mathbf{z}_\tau = \alpha_\tau \mathbf{z} + \sigma_\tau \boldsymbol{\epsilon}$, where $\boldsymbol{\xi}$ and $\boldsymbol{\epsilon}$ are independent standard Gaussian noise. The video teacher and an auxiliary score model then estimate the teacher and student scores at $\mathbf{z}_\tau$, respectively, to guide the generator update. We write $\mathbf{z}_\tau = (z_\tau^1, \ldots, z_\tau^L)$ and use $\mathbf{z}_\tau^{\setminus l}$ to denote all slices except z_τ^l. For either the teacher or student density p_τ, the joint distribution factorization yields

$$\nabla_{z_\tau^l} \log p_\tau(\mathbf{z}_\tau) = \nabla_{z_\tau^l} \log p_\tau\left(z_\tau^l \mid \mathbf{z}_\tau^{\setminus l}\right). \tag{3}$$

Thus, the score for each latent slice is conditioned on the other slices. Although joint distribution matching also constrains frame-level marginals, a change that improves an individual frame may not be favored if it conflicts with the surrounding frames. For example, adding a missing feather or finer fur details to one frame alone could introduce such inconsistencies with other frames. This potential tension, together with the observed quality limitations in Figure 1, motivates an additional marginal distribution objective.

3.2 DuoMatching: Joint-Marginal Distribution Matching

Let M denote a frame-sampling operator that selects one frame from a video according to a specified sampling policy. The corresponding frame-level marginals of the video teacher and real data are $m_v = MP_v$ and $m^* = MP^*$, respectively. Since m^* is defined over individual frames, it can also be directly approximated by an image generation model with distribution P_i. Under ideal marginal fitting within the same feasible family, the $D_{\text{KL}}(m^* \| P_i)$ is no greater than the $D_{\text{KL}}(m^* \| m_v)$ (see Appendix C for the assumptions and proof). Therefore, we extend the original distribution matching objective from a single proxy distribution to joint and marginal references, forming our DuoMatching framework.

Let $m_\theta = MQ_\theta$ denote the generator's frame-level marginal. We retain the video teacher as the joint distribution reference and use the image teacher for direct marginal supervision, defining the joint-marginal distribution matching objective:

$$\mathcal{J}_{DM}(\theta) = D_{\text{KL}}(Q_\theta \| P_v) + \omega D_{\text{KL}}(m_\theta \| P_i), \qquad \omega > 0, \tag{4}$$

where ω balances the influence of the video teacher's joint distribution estimate and the image teacher's complementary marginal estimate. Under the assumptions detailed in Appendix D, we show that there exists $\omega > 0$ for which the ideal joint-marginal optimum is closer to P^* than P_v

in forward KL. As illustrated in Figure 3(a), the marginal term strengthens frame-level distribution matching beyond joint supervision alone, while transferring complementary visual and semantic priors from the image teacher. The joint term simultaneously constrains cross-frame dependencies, encouraging the student to incorporate these priors coherently throughout the generated sequence. Implementing the joint and marginal terms with their corresponding DMD surrogates gives:

$$\mathcal{L}_{\text{DuoMatching}} = \mathcal{L}_{\text{joint-DMD}} + \omega \mathcal{L}_{\text{marginal-DMD}}. \tag{5}$$

In practice, joint DMD follows the procedure in Causal Forcing (Zhu et al., 2026). For marginal DMD, given a generated video latent $z = G_\theta(\xi, c)$, LatentBridge maps a temporal latent slice z^l and a local frame index i to an image-compatible representation $u^{l,(i)}$. The mapping is detailed in Section 3.3. We perturb this representation according to the image teacher's noise schedule: $u_\tau^{l,(i)} = \alpha_\tau^{\text{img}} u^{l,(i)} + \sigma_\tau^{\text{img}} \epsilon$, where $\epsilon \sim \mathcal{N}(0, I)$. A frozen image teacher and a trainable fake score estimator provide score estimates $s_{\text{real}}^{\text{img}}$ and $s_{\text{fake}}^{\text{img}}$ at $(u_\tau^{l,(i)}, \tau, c)$, respectively. The fake estimator is trained on the mapped generated latents to track their evolving distribution. The marginal DMD gradient is estimated as

$$\nabla_\theta \mathcal{L}_{\text{marginal}} = \mathbb{E}_{\xi,l,i,\tau,\epsilon}\left[w_{\text{img}}(\tau) \left(s_{\text{fake}}^{\text{img}} - s_{\text{real}}^{\text{img}} \right)^\top \frac{\partial u_\tau^{l,(i)}}{\partial \theta} \right], \tag{6}$$

where $w_{\text{img}}(\tau)$ is a timestep-dependent weighting function. The sampling of l is described in Section 3.4. Gradients propagate through the frozen LatentBridge to the student generator, providing direct frame-level supervision. Although motivated by the quality limitations in autoregressive generation, our joint-marginal framework also generalizes to bidirectional generators.

3.3 LatentBridge: Bridging Video and Image Latents

Because we use image generation models to approximate the marginal distribution, an additional issue arises from the difference between image VAE encoding and video VAE encoding. Most video VAEs (Zhao et al., 2024) use temporal compression, where one video latent maps to multiple actual RGB frames, while each image latent is constrained to one frame. A straightforward solution is to decode video latents into RGB frames and re-encode them with the image VAE, but backpropagating matching gradients through both VAEs incurs substantial memory overhead.

We therefore introduce LatentBridge, a lightweight differentiable module that maps temporally compressed video latents to frame-specific representations in the image teacher's latent space. We pretrain this module using paired video and image latents, as illustrated in Figure 3(b).

Given a temporal latent slice $\mathbf{z}^l$, we uniformly sample a local frame index $i \sim \mathcal{U}\{1, \ldots, r\}$, where r is the number of RGB frames represented by that slice. To account for the causal temporal encoding of the video VAE, we additionally condition on the preceding latent slice $\mathbf{z}^{l-1}$, providing temporal context for the mapping. Conditioned on these slices and i, LatentBridge predicts the latent of the corresponding RGB frame in the image teacher's latent space:

$$u^{l,(i)} = B_\phi(\mathbf{z}^{l-1}, \mathbf{z}^l, i). \tag{7}$$

For training, let $x^{l,i}$ denote the RGB frame corresponding to latent slice l and local frame index i. The target latent is obtained by encoding $x^{l,i}$ with the image teacher's VAE encoder E_{img}. We train B_ϕ using an ℓ_1 reconstruction loss:

$$\mathcal{L}_{LatentBridge} = \mathbb{E}_{\mathbf{z}^{l-1},\mathbf{z}^l,i}\left[\left\| B_\phi(\mathbf{z}^{l-1}, \mathbf{z}^l, i) - E_{\text{img}}(x^{l,i}) \right\|_1 \right]. \tag{8}$$

By conditioning on the sampled frame index, LatentBridge learns to recover distinct frame-level representations from shared temporal latent slices. Optimization through LatentBridge can improve the quality of decoded video frames (see Appendix F for an analysis of frame-level loss transfer). The module can be trained for the latent space of a chosen image teacher, allowing DuoMatching to use different image priors.

3.4 Latent Variation Sampling for Frame-Level Supervision

Marginal distribution matching supervises sampled frames, but concentrating samples in slowly changing regions produces redundant supervision. This incurs unnecessary computational cost and

may suppress motion dynamics by overemphasizing nearly static content. To address the issue, we introduce Latent Variation Sampling, as shown in Figure 3(c). Given the clean video latent sequence $\mathbf{z} = (\mathbf{z}^1, \ldots, \mathbf{z}^L)$, we compute the mean squared difference between adjacent latent slices:

$$d_l = \frac{1}{CHW} \left\| \mathbf{z}^{l+1} - \mathbf{z}^l \right\|_F^2, \qquad l = 1, \ldots, L-1, \tag{9}$$

where C, H, and W are the channel and spatial dimensions of each latent slice. We select the $K-1$ positions with the largest MSE values and split the sequence into K contiguous temporal segments. Denoting the sorted split positions by $b_1 < \cdots < b_{K-1}$, with $b_0 = 0$ and $b_K = L$, we uniformly sample one latent slice from each segment:

$$l_k \sim \mathcal{U}\{b_{k-1} + 1, \ldots, b_k\}, \qquad k = 1, \ldots, K. \tag{10}$$

For each selected slice $\mathbf{z}^{l_k}$, we apply marginal DMD to the image-aligned latent $u^{l_k,(i)}$, with i uniformly sampled. Each segment receives one frame sample per update, distributing the limited supervision budget across temporal regions.

4 Experiments

Implementation details. For causal models, we initialize the generator from the causal consistency distillation checkpoint from Causal Forcing++ (Zhao et al., 2026a) and subsequently optimize it with DMD on VidProM (Wang & Yang, 2024) for 1,000 training steps. For bidirectional models, following CausVid (Yin et al., 2025), we initialize the generator directly from Wan2.1-T2V-1.3B and perform DMD training for 1200 steps on Mixkit (Lin et al., 2024). In both settings, we use the frozen Qwen-Image (Wu et al., 2025) as the default image teacher. The output videos consist of 81 frames and have a resolution of H=480, W=832. LatentBridge is trained separately on a curated subset of OpenVid (Nan et al., 2025) containing 29,400 videos with high aesthetic quality and substantial motion. For DuoMatching, we set the marginal DMD loss weight to $\omega = 0.4$ and sample $K = 4$ temporal latent slices from each video by default. All experiments are conducted on eight NVIDIA A100 GPUs.

Baselines. For causal generation, we compare our method with autoregressive video generation methods, including Self Forcing (Huang et al., 2026), Reward Forcing (Lu et al., 2026), Causal Forcing++ (Zhao et al., 2026a; Zhu et al., 2026) and One-Forcing (Feng et al., 2026), covering both chunk-wise and frame-wise variants when available. For bidirectional generation, we use the bidirectional DMD variant of CausVid (Yin et al., 2025) as our primary baseline.

Metrics. We conduct automatic evaluation on VBench (Huang et al., 2024) using its official prompt suite and evaluation protocol. To more thoroughly assess video generation quality in fine-grained visual quality, visual composition, and semantic alignment, we additionally curate a test set of 400 challenging prompts featuring rich visual details and complex dynamics. We report the Semantic Score, Aesthetic Quality, Imaging Quality, Dynamic Degree, and Motion Smoothness, together with the aggregate VBench score.

4.1 Results

Qualitative comparison with baselines. Figure 4 presents side-by-side comparisons between Causal Forcing++ and DuoMatching. Compared with the baseline, DuoMatching produces finer visual details, such as more distinct surface textures and subtle color variations on the ceramic cat. Benefiting from the semantic alignment capabilities of the image teacher, DuoMatching also better follows prompt specifications, including object counts (a single wooden canoe) and attributes (a teal ceramic cup). In addition, DuoMatching exhibits improved composition, framing the subject as intended rather than incorrectly focusing the camera on the waist. These improvements in fine-grained visual quality, composition, and semantic alignment are achieved while preserving temporal quality, as illustrated by the smooth and consistent paddling motions in the kayaking example. Further comparisons with other baselines are provided in Appendix B.

Quantitative results. The quantitative results in Table 1 show DuoMatching's significant advantages over the compared methods within each matched generation mode and NFE budget. For bidirectional generation, DuoMatching improves upon CausVid in visual quality, semantic alignment,



Figure 4: Qualitative comparison between DuoMatching and Causal Forcing++ across four aspects: fine-grained visual quality, semantic alignment, composition, and dynamics. DuoMatching improves visual details, prompt adherence, and composition while preserving video dynamics.

and video dynamics. For causal generation, DuoMatching achieves higher video generation quality across one-, two-, and four-step settings, yielding considerable improvements in visual quality and semantic alignment while largely preserving motion quality. These results support our joint-marginal formulation, which improves frame-level quality through explicit marginal matching and complementary visual and semantic priors from pretrained image generation models. All gains are achieved without additional inference computation, preserving the real-time generation capability of Causal Forcing++ (Zhao et al., 2026a).

Human evaluation. We further conduct a human evaluation following a two-alternative forced-choice (2AFC) protocol. We recruit 23 participants, each of whom evaluates 40 prompt-matched pairs of videos generated by different methods and indicates their preference in terms of visual quality, semantic alignment, temporal quality, and overall quality. As shown in Table 2, DuoMatching achieves overall preference rates above 80% against every baseline. It receives higher preference rates for visual quality and semantic alignment, supporting the effectiveness of explicit marginal distribution matching in improving frame-level quality. Preference rates for Temporal & Motion are near or above parity, indicating that temporal quality is largely preserved.

Table 1: Quantitative comparison on VBench with Self Forcing (Huang et al., 2026), One-Forcing (Feng et al., 2026), Reward Forcing (Lu et al., 2026) and Causal Forcing++ (Zhao et al., 2026a). "Full" denotes bidirectional full-video generation, while "Frame" and "Chunk" denote the causal autoregressive generation unit. NFE denotes the number of function evaluations per generation unit. Results should be compared within matched generation modes and NFE budgets. Bold indicates the best result within each matched group; ours are shaded in light gray.

Method	Mode	NFE	Semantic ↑	Aesthetic ↑	Imaging ↑	Dynamic ↑	Smoothness ↑	Total ↑
Bidirectional DMD methods								
CausVid	Full	4	69.23	64.15	73.00	36.67	**98.84**	80.63
Ours	Full	4	**70.89**	**64.55**	**73.77**	**39.17**	98.82	**81.45**
Causal autoregressive methods								
Frame-wise generation (1 step)								
Causal Forcing++	Frame	1	57.17	67.12	68.83	**70.56**	98.49	79.82
One-Forcing	Frame	1	69.25	65.08	68.28	66.11	99.03	82.09
Ours	Frame	1	**69.95**	**68.62**	**69.42**	69.44	**99.07**	**82.65**
Frame-wise generation (2 steps)								
Causal Forcing++	Frame	2	68.42	62.28	69.82	**94.44**	98.20	80.67
Ours	Frame	2	**72.84**	**66.97**	**73.48**	93.61	**98.39**	**83.53**
Chunk-wise generation (4 steps)								
CausVid	Chunk	4	68.69	62.16	66.97	55.83	98.04	80.19
Self Forcing	Chunk	4	71.41	66.83	69.66	68.61	95.52	81.52
Reward Forcing	Chunk	4	56.23	63.62	71.72	73.61	98.44	80.10
Causal Forcing++	Chunk	4	70.84	64.47	70.13	**80.56**	98.14	82.79
Ours	Chunk	4	**71.61**	**66.90**	**72.41**	76.67	**99.06**	**83.51**

Table 2: Human preference study under the same Wan2.1 video teacher. Each entry reports the percentage of pairwise judgments favoring our method over the corresponding baseline.

Comparison	Visual Quality ↑	Semantic Alignment ↑	Temporal & Motion ↑	Overall ↑
Ours vs. CausVid	92.17%	89.13%	51.30%	91.30%
Ours vs. Causal Forcing++	81.30%	79.57%	49.57%	80.87%
Ours vs. One-Forcing	93.48%	91.74%	90.87%	94.35%
Ours vs. Reward Forcing	83.91%	82.61%	60.87%	83.91%

4.2 Ablation Study

Gains from marginal matching and stronger image priors. Our ablation suggests that the improvements above stem from both explicit marginal matching and stronger image priors. As shown in Table 3, using Wan2.1-14B as the image teacher, without introducing additional priors from an external image generation model, already improves performance across all reported metrics. We further compare SDXL (Podell et al., 2024), FLUX.2-4B (Labs, 2025), and Qwen-Image (Wu et al., 2025). Across these three image teachers, higher teacher HPSv3 scores are accompanied by consistent improvements in Semantic, Aesthetic, and Imaging scores. Thanks to the spatiotemporal decoupling enabled by LatentBridge, Dynamic Degree and Motion Smoothness remain stable across different image teachers, indicating that teacher capability primarily affects visual quality and semantic alignment with relatively trivial changes in motion dynamics and smoothness.

Enabling frame-level marginal matching with LatentBridge. We investigate the role of LatentBridge in enabling marginal matching. Qwen-Image and Wan share the same VAE encoder, allowing us to directly apply marginal DMD to temporally compressed video latent slices as a baseline. As shown in Table 4, such direct application improves semantic alignment and visual quality, but suppresses the dynamic information encoded in the latent representation, resulting in videos with limited motion. By mapping temporally compressed video latents to frame-specific image latents, LatentBridge separates spatial information in video latents and aligns it with the target image latent space. LatentBridge yields further gains in visual quality and semantic alignment while largely retaining video dynamics. While Decode-Encode runs out of memory under our training configuration, LatentBridge incurs only a small additional memory overhead over the Direct baseline.

Table 3: Ablation of image teachers, comparing Wan2.1-14B (Wan et al., 2025), SDXL (Podell et al., 2024), FLUX.2-4B (Labs, 2025), and Qwen-Image (Wu et al., 2025) against a baseline without an image teacher. Teacher HPSv3 (Ma et al., 2025) measures human preference alignment of images generated by each teacher, while the remaining metrics evaluate the resulting video student.

	Teacher	Video Student					
Image Teacher	HPSv3 ↑	Semantic ↑	Aesthetic ↑	Imaging ↑	Dynamic ↑	Smoothness ↑	Total ↑
None	–	68.42	62.28	69.82	94.44	98.20	80.67
Wan2.1-14B	5.87	68.77	64.16	69.90	94.72	99.05	81.43
SDXL	6.57	68.50	64.78	70.01	93.89	98.12	81.47
FLUX.2-4B	8.86	69.06	65.29	71.58	94.17	99.23	82.64
Qwen-Image	9.05	72.84	66.97	73.48	93.61	98.39	83.53

Table 4: Ablation of marginal DMD and latent mapping strategies. Direct applies marginal DMD directly to temporally compressed video latent slices. Decode-Encode decodes video latents into RGB frames and re-encodes sampled frames with the image VAE, retaining gradients through both operations. Peak memory denotes the maximum GPU memory allocated to tensors during a complete training iteration. OOM indicates an out-of-memory failure.

Configuration		Evaluation Metrics				Peak Memory
Marginal DMD	Latent Mapping	Semantic ↑	Imaging ↑	Dynamic ↑	Total ↑	(GB) ↓
×	—	68.42	69.82	94.44	80.67	56.49
✓	Direct	71.74	71.66	76.94	81.78	59.40
✓	Decode-Encode	—	—	—	—	OOM
✓	LatentBridge	72.84	73.48	93.61	83.53	59.41

Effective allocation of frame-level supervision with Latent Variation Sampling (LVS). We first examine the effect of the number of sampled latent slices K, then compare LVS with uniform random and temporally stratified sampling. As shown in Table 5, increasing K from 2 to 4 improves the Total score for both uniform sampling and LVS, whereas further increasing K to 8 reduces both Dynamic and Total scores. This indicates that increasing K raises computational cost without necessarily improving performance, highlighting the importance of an effective supervision allocation strategy. LVS outperforms uniform sampling across all tested K and temporally stratified sampling at $K = 4$. These results demonstrate that LVS enables more effective frame-level supervision while preserving temporal quality and limiting computational cost.

Table 5: Ablation of the number of sampled latent slices K and sampling strategies. Temporally stratified sampling randomly selects one latent slice from each of K segments of equal length.

Sampling strategy	$K = 2$	$K = 4$	$K = 8$
	Dynamic ↑ / Total ↑	Dynamic ↑ / Total ↑	Dynamic ↑ / Total ↑
Uniform random	93.06 / 80.99	91.39 / 82.61	89.44 / 81.04
Temporally stratified	– / –	91.67 / 82.68	– / –
LVS (Ours)	94.17 / 81.15	93.61 / 83.53	92.22 / 82.55

5 Conclusion

We presented DuoMatching, a unified joint-marginal distribution matching framework for few-step video generation. By complementing joint DMD with explicit marginal DMD, our framework strengthens frame-level supervision and transfers visual and semantic priors from image models. LatentBridge enables marginal matching for temporally compressed video latents, while Latent Variation Sampling reduces redundant supervision. Experiments demonstrate improvements in visual quality, composition, and semantic alignment while largely preserving motion dynamics.

AI USE STATEMENT

Generative AI tools were employed to refine the manuscript's language, including grammar, phrasing, and clarity. Their use did not extend to developing research ideas, designing methods or experiments, interpreting findings, or formulating mathematical claims. The authors reviewed and revised all AI-assisted text to verify its technical correctness and preserve the intended meaning. We assume full responsibility for the final manuscript, including content prepared with generative AI assistance.

REFERENCES

Yunpeng Bai, Yossi Gandelsman, Michaël Gharbi, and Qixing Huang. Instruction-based video editing by repurposing an image editing model, 2026. URL https://arxiv.org/abs/2608.14790.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Huanqia Cai, Sihan Cao, Ruoyi Du, Peng Gao, Aiming Hao, Steven Hoi, Zhaohui Hou, Shijie Huang, Dengyang Jiang, Yuming Jiang, et al. Z-image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Siyi Chen, Shaowei Liu, Yixuan Jia, Zian Wang, Huan Ling, Qing Qu, and Jun Gao. Data-forcing distillation: Restoring diversity and fidelity in few-step video generation. *arXiv preprint arXiv:2606.18478*, 2026.

Zihan Ding, Chi Jin, Difan Liu, Haitian Zheng, Krishna Kumar Singh, Qiang Zhang, Yan Kang, Zhe Lin, and Yuchen Liu. Dollar: Few-step video generation via distillation and latent reward optimization. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 17961–17971. IEEE, 2025.

Xiangyu Fan, Zesong Qiu, Zhuguanyu Wu, Fanzhou Wang, Zhiqian Lin, Tianxiang Ren, Dahua Lin, Ruihao Gong, and Lei Yang. Phased dmd: Few-step distribution matching distillation via score matching within subintervals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41667–41676, 2026.

Jiaqi Feng, Justin Cui, Yuanhao Ban, and Cho-Jui Hsieh. One-forcing: Towards stable one-step autoregressive video generation. *arXiv preprint arXiv:2605.23458*, 2026.

Xiefan Guo, Jinlin Liu, Miaomiao Cui, Liefeng Bo, and Di Huang. I4vgen: Image as free stepping stone for text-to-video generation. *arXiv preprint arXiv:2406.02230*, 2024.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *Advances in Neural Information Processing Systems*, 38:167283–167308, 2026.

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

Dengyang Jiang, Dongyang Liu, Zanyi Wang, Qilong Wu, Liuzhuozheng Li, Hengzhuang Li, Xin Jin, David Liu, Changsheng Lu, Zhen Li, et al. Distribution matching distillation meets reinforcement learning. *arXiv preprint arXiv:2511.13649*, 2025.

Black Forest Labs. FLUX.2: Frontier Visual Intelligence. https://bfl.ai/blog/flux-2, 2025.

Jiachen Li, Weixi Feng, Tsu-Jui Fu, Xinyi Wang, Sugato Basu, Wenhu Chen, and William Y Wang. T2v-turbo: Breaking the quality bottleneck of video consistency model with mixed reward feedback. *Advances in neural information processing systems*, 37:75692–75726, 2024.

Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024.

Shanchuan Lin, Xin Xia, Yuxi Ren, Ceyuan Yang, Xuefeng Xiao, and Lu Jiang. Diffusion adversarial post-training for one-step video generation. *arXiv preprint arXiv:2501.08316*, 2025.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Wei Liu, Ziyu Chen, Zizhang Li, Yue Wang, Hong-Xing Yu, and Jiajun Wu. Realwonder: Real-time physical action-conditioned video generation. *arXiv preprint arXiv:2603.05449*, 2026.

Yixin Liu, Kai Zhang, Yuan Li, Zhiling Yan, Chujie Gao, Ruoxi Chen, Zhengqing Yuan, Yue Huang, Hanchi Sun, Jianfeng Gao, et al. Sora: A review on background, technology, limitations, and opportunities of large vision models. *arXiv preprint arXiv:2402.17177*, 2024.

Yunhong Lu, Yanhong Zeng, Haobo Li, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Jiapeng Zhu, Hengyuan Cao, Zhipeng Zhang, Xing Zhu, et al. Reward forcing: Efficient streaming video generation with rewarded distribution matching distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34385–34397, 2026.

Yuhang Ma, Xiaoshi Wu, Keqiang Sun, and Hongsheng Li. Hpsv3: Towards wide-spectrum human preference score. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15086–15095. IEEE, 2025.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. In *International Conference on Learning Representations*, volume 2025, pp. 1045–1064, 2025.

Jack Parker-Holder, Shlomi Fruchter, et al. Genie 3: A new frontier for world models. *Google DeepMind Blog*, 2025.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations*, volume 2024, pp. 1862–1874, 2024.

Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

Team Seedream, Yunpeng Chen, Yu Gao, Lixue Gong, Meng Guo, Qiushan Guo, Zhiyao Guo, Xiaoxia Hou, Weilin Huang, Yixuan Huang, et al. Seedream 4.0: Toward next-generation multimodal image generation. *arXiv preprint arXiv:2509.20427*, 2025.

Shitong Shao, Lichen Bai, Haoyi Xiong, Zeke Xie, et al. Iv-mixed sampler: Leveraging image diffusion models for enhanced video synthesis. In *International Conference on Learning Representations*, volume 2025, pp. 12396–12424, 2025.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. Worldplay: Towards long-term geometric consistency for real-time interactive world modeling. *arXiv preprint arXiv:2512.14614*, 2025.

Yu Tian, Jian Ren, Menglei Chai, Kyle Olszewski, Xi Peng, Dimitris N Metaxas, and Sergey Tulyakov. A good image generator is what you need for high-resolution video synthesis. *arXiv preprint arXiv:2104.15069*, 2021.

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

Wenhao Wang and Yi Yang. Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. *Advances in Neural Information Processing Systems*, 37:65618–65642, 2024.

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

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in neural information processing systems*, 36:8406–8441, 2023b.

Thaddäus Wiedemer, Yuxuan Li, Paul Vicol, Shixiang Shane Gu, Nick Matarese, Kevin Swersky, Been Kim, Priyank Jaini, and Robert Geirhos. Video models are zero-shot learners and reasoners. *arXiv preprint arXiv:2509.20328*, 2025.

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

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023.

Yicheng Xiao, Wenxun Dai, Xinran Qin, Lin Song, Maoquan Zhang, Hang Xu, Yukang Chen, Yitong Li, Guohui Zhang, Yuan Zhang, et al. Joyai-video-edit: Real-time open-ended video editing with autoregressive diffusion. *arXiv preprint arXiv:2608.03974*, 2026.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, et al. Longlive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*, 2025a.

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 *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025b.

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

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Fredo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6613–6623. IEEE, 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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22963–22974. IEEE, 2025.

Yuanhao Zhai, Kevin Lin, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Chung-Ching Lin, David Doermann, Junsong Yuan, and Lijuan Wang. Motion consistency model: Accelerating video diffusion with disentangled motion-appearance distillation. *arXiv preprint arXiv:2406.06890*, 2024.

Yabo Zhang, Yuxiang Wei, Xianhui Lin, Zheng Hui, Peiran Ren, Xuansong Xie, and Wangmeng Zuo. Videoelevator: Elevating video generation quality with versatile text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 10266–10274, 2025.

Min Zhao, Hongzhou Zhu, Kaiwen Zheng, Zihan Zhou, Bokai Yan, Xinyuan Li, Xiao Yang, Chongxuan Li, and Jun Zhu. Causal forcing++: Scalable few-step autoregressive diffusion distillation for real-time interactive video generation. *arXiv preprint arXiv:2605.15141*, 2026a.

Sijie Zhao, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Muyao Niu, Xiaoyu Li, Wenbo Hu, and Ying Shan. Cv-vae: A compatible video vae for latent generative video models. *Advances in Neural Information Processing Systems*, 37:12847–12871, 2024.

Yuyang Zhao, Yicheng Pan, Qiyuan He, Jincheng Yu, Junsong Chen, Tian Ye, Haozhe Liu, Enze Xie, and Song Han. Sana-streaming: Real-time streaming video editing with hybrid diffusion transformer. *arXiv preprint arXiv:2605.30409*, 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. *arXiv preprint arXiv:2602.02214*, 2026.

Yuanzhi Zhu, Hanshu Yan, Huan Yang, Kai Zhang, and Junnan Li. Accelerating video diffusion models via distribution matching. *arXiv preprint arXiv:2412.05899*, 2024.

A HYPERPARAMETER ABLATIONS FOR DUOMATCHING.

We examine the effect of the marginal DMD loss weight ω. As shown in Table 6, increasing ω from 0.1 to 0.4 improves semantic alignment and imaging quality while maintaining a similar Dynamic score. Further increasing ω to 0.8 reduces both metrics and substantially lowers the Dynamic score, suggesting that excessive marginal supervision compromises the balance between visual quality and video dynamics. Among the tested settings, $\omega = 0.4$ achieves the highest scores across all reported metrics, and we adopt it as the default.

B ADDITIONAL QUALITATIVE COMPARISONS

Figure 5 presents additional qualitative comparisons with Reward Forcing, One-Forcing, and CausVid. Compared with Reward Forcing, DuoMatching renders the ice dragon with more clearly defined structures and finer details, while better preserving the boxer's appearance during rapid motion. In the pianist example, it also produces a more plausible spatial relationship between the pianist, keyboard, and piano, better reflecting the intended interaction. Compared with One-Forcing and CausVid, DuoMatching alleviates overexposure and produces more realistic results, with more natural lighting and richer details in animal fur, feathers, clothing, and surrounding environments. In particular, under the $NFE = 1$ setting, DuoMatching substantially improves motion stability over One-Forcing, reducing abrupt changes in subject appearance and position across frames. Together, these examples illustrate improvements in generation quality across different video distillation baselines.

C PROOF OF THE MARGINAL FITTING COMPARISON

Using the notation of the main text, let P^* denote the real video distribution, P_v the video reference, and P_i the image reference. Using the same frame-sampling operator M, define $m^* = MP^*$ and $m_v = MP_v$. Let $\mathcal{P}$ be a feasible family of video distributions and let $\mathcal{F} = \{MP : P \in \mathcal{P}\}$. For this idealized comparison, assume matched marginal capacity: $\mathcal{F}$ is also the feasible family of image distributions. We assess reference fitting using forward KL.

For any $P \in \mathcal{P}$, construct the joint distributions of a video x and a sampled frame y using the same sampling kernel $M(y \mid x)$. The KL chain rule gives

$$D_{\mathrm{KL}}(P^* \| P) = D_{\mathrm{KL}}(m^* \| MP) + C(P), \tag{11}$$

where

$$C(P) = \mathbb{E}_{y \sim m^*} \left[D_{\mathrm{KL}}\big(P^*(x \mid y) \| P(x \mid y)\big) \right]. \tag{12}$$

The conditional distributions are defined by this video-and-frame sampling procedure. Thus, joint fitting accounts for both marginal and conditional distribution discrepancies.

In this idealized comparison, assume that P_i and P_v attain the following finite minima:

$$P_i \in \arg\min_{m \in \mathcal{F}} D_{\mathrm{KL}}(m^* \| m), \qquad P_v \in \arg\min_{P \in \mathcal{P}} D_{\mathrm{KL}}(P^* \| P). \tag{13}$$

Since $P_i \in \mathcal{F}$, there exists a video distribution $P^\dagger \in \mathcal{P}$ whose frame-level marginal satisfies $MP^\dagger = P_i$. We assume that such a distribution can be chosen with $D_{\mathrm{KL}}(P^* \| P^\dagger) < \infty$. This auxiliary distribution allows us to compare joint fitting under P_v with fitting the image reference P_i as the marginal.

The marginal optimality of P_i gives

$$D_{\mathrm{KL}}(m^* \| P_i) \le D_{\mathrm{KL}}(m^* \| m_v). \tag{14}$$

The joint optimality of P_v, together with the KL decomposition, gives

$$D_{\mathrm{KL}}(m^* \| m_v) + C(P_v) \le D_{\mathrm{KL}}(m^* \| P_i) + C(P^\dagger). \tag{15}$$

Combining these inequalities yields

$$0 \le D_{\mathrm{KL}}(m^* \| m_v) - D_{\mathrm{KL}}(m^* \| P_i) \le C(P^\dagger) - C(P_v). \tag{16}$$

Thus, under matched marginal capacity, marginal-only fitting achieves no greater marginal error than joint fitting, and this advantage is bounded by the additional conditional fitting error of $P^\dagger$ relative to P_v.

D Proof of Joint-Marginal Improvement

Using the same frame-sampling operator M, define $m^* = MP^*$ and $m_v = MP_v$. Assume that the image reference has no greater marginal approximation error than the video-teacher marginal, measured by forward KL, and that the two distributions differ:

$$D_{\mathrm{KL}}(m^* \| P_i) \le D_{\mathrm{KL}}(m^* \| m_v), \qquad P_i \neq m_v. \tag{17}$$

Consider the ideal distribution-space optimum

$$Q_\omega \in \arg\min_Q \left\{ D_{\mathrm{KL}}(Q \| P_v) + \omega D_{\mathrm{KL}}(MQ \| P_i) \right\}, \qquad m_\omega = MQ_\omega, \tag{18}$$

where the minimization is over all video distributions. We assume that the reference and optimal distributions have positive densities on common supports and that all displayed KL divergences are finite. We further assume that the optimal density path Q_ω exists for all sufficiently small $\omega \ge 0$, is right-differentiable at $\omega = 0$, and permits interchanging the differentiations below with integration and the sampling operator. We use the same symbols for distributions and their densities. At $\omega = 0$, the unique minimizer is $Q_0 = P_v$. To establish local improvement, we show that $D_{\mathrm{KL}}(P^* \| Q_\omega)$ has a strictly negative right derivative at $\omega = 0$.

Since M is linear, the first variations of the two KL terms with respect to $Q(x)$ are $\log(Q(x)/P_v(x)) + 1$ and $\mathbb{E}_{Y \sim M(\cdot|x)}[\log((MQ)(Y)/P_i(Y)) + 1]$, respectively.

$$(MQ)(y) = \int M(y \mid x) Q(x)\, dx, \qquad \frac{\delta (MQ)(y)}{\delta Q(x)} = M(y \mid x),$$

$$\frac{\delta}{\delta Q(x)} D_{\mathrm{KL}}(MQ \| P_i) = \int M(y \mid x) \left[\log \frac{(MQ)(y)}{P_i(y)} + 1 \right] dy.$$

$$\mathcal{A}_\omega(Q, \lambda) = D_{\mathrm{KL}}(Q \| P_v) + \omega D_{\mathrm{KL}}(MQ \| P_i) + \lambda \left(\int Q(x)\, dx - 1 \right).$$

Imposing the normalization constraint and absorbing terms independent of x into a constant gives

$$0 = \log \frac{Q_\omega(x)}{P_v(x)} + 1 + \omega\, \mathbb{E}_{Y \sim M(\cdot|x)} \left[\log \frac{m_\omega(Y)}{P_i(Y)} + 1 \right] + \lambda_\omega,$$

$$c_\omega = -1 - \omega - \lambda_\omega.$$

$$\log \frac{Q_\omega(x)}{P_v(x)} + \omega\, \mathbb{E}_{Y \sim M(\cdot|x)} \left[\log \frac{m_\omega(Y)}{P_i(Y)} \right] = c_\omega. \tag{19}$$

Define

$$g(x) = \mathbb{E}_{Y \sim M(\cdot|x)} \left[\log \frac{m_v(Y)}{P_i(Y)} \right]. \tag{20}$$

Taking the right derivative of the optimality condition at $\omega = 0$, with dots denoting these derivatives, yields

$$\begin{aligned} \dot{c}_0 &= \frac{\dot{Q}_0(x)}{Q_0(x)} + \mathbb{E}_{Y \sim M(\cdot|x)} \left[\log \frac{m_0(Y)}{P_i(Y)} \right] \\ &= \frac{\dot{Q}_0(x)}{P_v(x)} + g(x). \end{aligned} \tag{21}$$

Differentiating $\int Q_\omega(x)\, dx = 1$ gives $\int \dot{Q}_0(x)\, dx = 0$. Therefore, integrating the preceding equality against P_v gives

$$\underbrace{\int \dot{Q}_0(x)\, dx}_{=0} + \int P_v(x) g(x)\, dx = \dot{c}_0 \underbrace{\int P_v(x)\, dx}_{=1}.$$

$$\begin{aligned} \dot{c}_0 &= \mathbb{E}_{X \sim P_v}[g(X)] \\ &= \mathbb{E}_{X \sim P_v} \mathbb{E}_{Y \sim M(\cdot|X)} \left[\log \frac{m_v(Y)}{P_i(Y)} \right] \\ &= \mathbb{E}_{Y \sim m_v} \left[\log \frac{m_v(Y)}{P_i(Y)} \right] \\ &= D_{\mathrm{KL}}(m_v \| P_i). \end{aligned} \tag{22}$$

Let $F(\omega) = D_{\mathrm{KL}}(P^* \| Q_\omega)$. Substituting the expression for $\dot{Q}_0/P_v$ gives

$$\begin{aligned} F'_+(0) &= -\int P^*(x) \frac{\dot{Q}_0(x)}{P_v(x)}\, dx \\ &= -\mathbb{E}_{X\sim P^*}\left[\dot{c}_0 - g(X)\right] \\ &= \mathbb{E}_{X\sim P^*}[g(X)] - D_{\mathrm{KL}}(m_v \| P_i) \\ &= \mathbb{E}_{Y\sim m^*}\left[\log \frac{m_v(Y)}{P_i(Y)}\right] - D_{\mathrm{KL}}(m_v \| P_i) \\ &= \mathbb{E}_{Y\sim m^*}\left[\log \frac{m^*(Y)}{P_i(Y)} - \log \frac{m^*(Y)}{m_v(Y)}\right] - D_{\mathrm{KL}}(m_v \| P_i) \\ &= D_{\mathrm{KL}}(m^* \| P_i) - D_{\mathrm{KL}}(m^* \| m_v) - D_{\mathrm{KL}}(m_v \| P_i) \\ &< 0. \end{aligned} \tag{23}$$

The last equality follows by expanding the two forward KL divergences. Their difference is non-positive by assumption, while $D_{\mathrm{KL}}(m_v \| P_i) > 0$ because $P_i \neq m_v$. Hence, $F'_+(0) < 0$.

By right differentiability, $F(\omega) = F(0) + \omega F'_+(0) + o(\omega)$ as $\omega \downarrow 0$. Since $F'_+(0) < 0$ and $Q_0 = P_v$, there exists $\omega_0 > 0$ such that

$$D_{\mathrm{KL}}(P^* \| Q_\omega) < D_{\mathrm{KL}}(P^* \| P_v), \qquad 0 < \omega < \omega_0. \tag{24}$$

Thus, there exists a positive marginal weight ω for which the ideal joint-marginal optimum is closer to P^* than P_v in forward KL.

E Efficiency and Reconstruction Quality of LatentBridge

LatentBridge provides a lightweight mapping between video and image latent spaces. LatentBridge consists of eight FiLM residual blocks with 256 hidden channels, totaling only 10.75M trainable parameters. It is trained on 29,400 videos at 480p resolution with an effective batch size of 32. Training for 3,000 iterations takes approximately 15 minutes on 8 NVIDIA A100 GPUs. Despite its compact size and short training time, LatentBridge produces reconstructions that closely match the targets in scene structure and visual details, as shown in Figure 6. The examples across Qwen-Image, FLUX, and SDXL further illustrate its applicability to different image latent spaces.

F Transferring Loss Improvement through LatentBridge

We establish a sufficient condition under which a loss reduction measured through LatentBridge also holds for the corresponding decoded video frames. Let $z_\theta = G_\theta(\xi, c)$, where ξ is the generator noise and c is the text condition. For a sampled latent slice l and local frame index i, let $t(l, i)$ denote the corresponding RGB frame index. Define the predicted and image-latent representations as

$$u_\theta = B_\phi(z_\theta^{l-1}, z_\theta^l, i), \qquad v_\theta = E_{\mathrm{img}}\big([D_{\mathrm{video}}(z_\theta)]_{t(l,i)}\big), \tag{25}$$

where D_{video} is the video VAE decoder, E_{img} is the image VAE encoder, and B_ϕ is frozen. Both representations use the same generated video and sampled frame position.

Let $\ell(u, c)$ be a fixed scalar frame-level loss, with lower values indicating better performance under this loss. Assume that, on a domain containing the representations considered below, it satisfies

$$|\ell(u, c) - \ell(v, c)| \le L_\ell \|u - v\|_1 \tag{26}$$

for a finite constant $L_\ell > 0$. Define the two objectives

$$\mathcal{J}_B(\theta) = \mathbb{E}[\ell(u_\theta, c)], \qquad \mathcal{J}_R(\theta) = \mathbb{E}[\ell(v_\theta, c)], \tag{27}$$

and assume that these expectations are finite. Consider generator parameters θ_0 and θ_1 before and after an update or a sequence of updates. Suppose the expected mapping errors at these parameters satisfy

$$\mathbb{E}\|u_{\theta_j} - v_{\theta_j}\|_1 \le \varepsilon_j, \qquad j \in \{0, 1\}. \tag{28}$$

Table 6: Ablation of the marginal DMD loss weight ω.

Metric	$\omega = 0.1$	$\omega = 0.4$	$\omega = 0.8$
Semantic ↑	69.60	72.84	71.49
Imaging ↑	70.79	73.48	71.74
Dynamic ↑	93.61	93.61	87.78
Total ↑	81.03	83.53	81.57

The Lipschitz condition then gives, at each endpoint,

$$\begin{aligned} |\mathcal{J}_B(\theta_j) - \mathcal{J}_R(\theta_j)| &\le \mathbb{E}\big[|\ell(u_{\theta_j}, c) - \ell(v_{\theta_j}, c)|\big] \\ &\le L_\ell \mathbb{E}\|u_{\theta_j} - v_{\theta_j}\|_1 \le L_\ell \varepsilon_j. \end{aligned} \tag{29}$$

Suppose the objective evaluated through LatentBridge decreases by $\delta = \mathcal{J}_B(\theta_0) - \mathcal{J}_B(\theta_1) > 0$. Adding and subtracting the two bridge objectives yields

$$\begin{aligned} \mathcal{J}_R(\theta_1) - \mathcal{J}_R(\theta_0) &= [\mathcal{J}_R(\theta_1) - \mathcal{J}_B(\theta_1)] + [\mathcal{J}_B(\theta_1) - \mathcal{J}_B(\theta_0)] \\ &\quad + [\mathcal{J}_B(\theta_0) - \mathcal{J}_R(\theta_0)] \\ &\le -\delta + L_\ell(\varepsilon_0 + \varepsilon_1). \end{aligned} \tag{30}$$

Consequently, if $\delta > L_\ell(\varepsilon_0 + \varepsilon_1)$, the same loss evaluated on image latents encoded from the actual decoded video frames strictly decreases. In particular, when $\varepsilon_0, \varepsilon_1 \le \varepsilon$, a decrease exceeding $2L_\ell\varepsilon$ is sufficient.

As shown in Figure 7, the mapping errors remain consistent across training steps. Thus, when these errors remain bounded, a decrease in the fixed frame-level loss through LatentBridge exceeding $L_\ell(\varepsilon_0 + \varepsilon_1)$ guarantees a decrease in the same loss evaluated through Decode–Encode.

G Training Procedure

Algorithm 1 summarizes the training procedure of DuoMatching. The generator is optimized with joint DMD and marginal DMD, while both fake estimators are updated on freshly generated samples. The teachers and LatentBridge remain frozen, with gradients propagated through LatentBridge during generator updates.

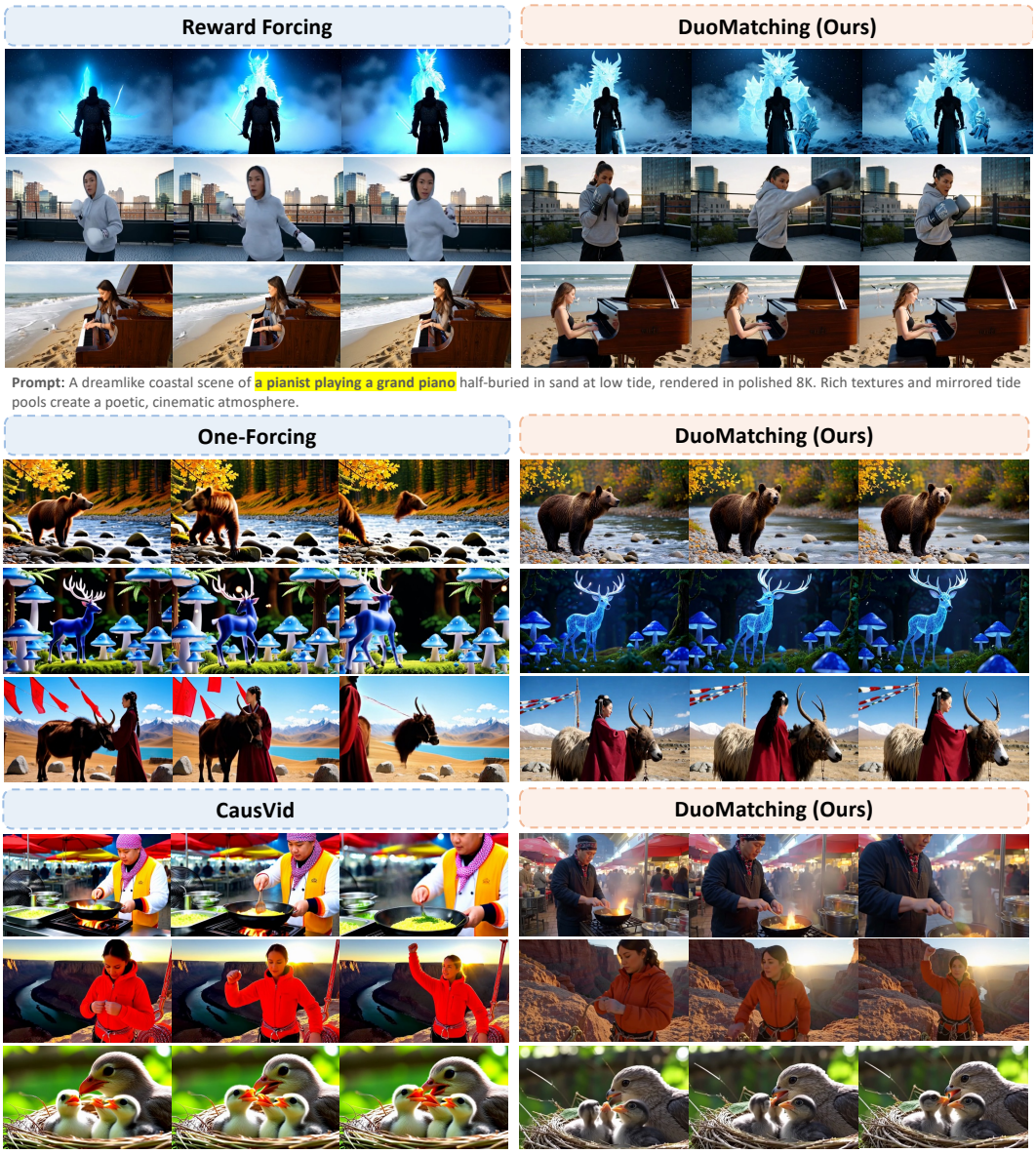

Figure 5: Additional qualitative comparisons of DuoMatching with Reward Forcing, One-Forcing, and CausVid. DuoMatching improves visual details, semantic alignment, and composition while preserving video dynamics.

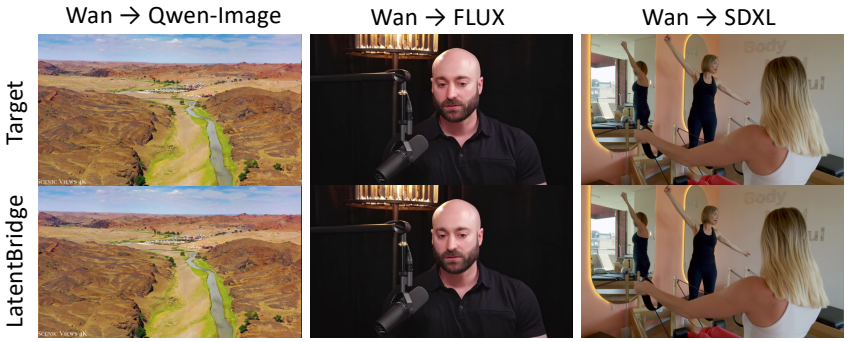

Figure 6: Qualitative evaluation of LatentBridge across image latent spaces. Top: target latents obtained by encoding the corresponding RGB frames with each image VAE. Bottom: reconstructions from latents predicted by LatentBridge from video latents of Wan2.1. Both rows use the corresponding image VAE decoder.

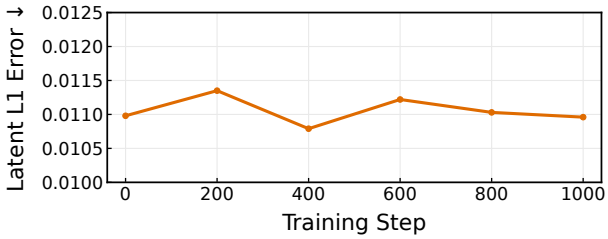

Figure 7: Latent reconstruction error of LatentBridge at different stages of generator training. Given a generated video latent slice z^l, with z^{l-1} and the local frame index i as conditioning, LatentBridge predicts the corresponding image latent. We measure its ℓ_1 error against the target obtained through Decode-Encode over 100 samples.

Algorithm 1 Training of DuoMatching

Require: Generator G_θ; frozen LatentBridge B_ϕ
Require: Frozen teachers μ^v_{real}, $\mu^{\text{img}}_{\text{real}}$
Require: Trainable fake estimators μ^v_{fake}, $\mu^{\text{img}}_{\text{fake}}$
Require: Marginal weight ω; sampling budget K; update interval R; warmup length $T_{\text{warm}} > 0$; training iterations N

for $n = 0, \dots, N-1$ **do**
  if $n \bmod R = 0$ **then**
    Generator update
    Sample text condition c and noise $\xi \sim \mathcal{N}(0, I)$
    $z \leftarrow G_\theta(\xi, c)$
    $\{l_k\}_{k=1}^K \leftarrow \text{LVS}(z, K)$
    for $k = 1, \dots, K$ **do**
      Sample local frame index i_k
      $u_k \leftarrow B_\phi(z^{l_k-1}, z^{l_k}, i_k)$ ▷ $u_k = u^{l_k,(i_k)}$
    end for
    $U \leftarrow \text{Stack}(u_1, \dots, u_K)$ ▷ Retain gradients through B_ϕ
    Sample image timestep τ and $\epsilon \sim \mathcal{N}(0, I)$
    $U_\tau \leftarrow \alpha^{\text{img}}_\tau \,\text{sg}(U) + \sigma^{\text{img}}_\tau \epsilon$
    $s^{\text{img}}_{\text{real}} \leftarrow \text{Score}(\mu^{\text{img}}_{\text{real}}; U_\tau, \tau, c)$ ▷ Image teacher; no autograd
    $s^{\text{img}}_{\text{fake}} \leftarrow \text{Score}(\mu^{\text{img}}_{\text{fake}}; U_\tau, \tau, c)$ ▷ Fake estimator; no autograd
    $\Delta s \leftarrow s^{\text{img}}_{\text{fake}} - s^{\text{img}}_{\text{real}}$ ▷ Score difference
    $g \leftarrow \alpha^{\text{img}}_\tau w_{\text{img}}(\tau) \Delta s$ ▷ DMD gradient with respect to U
    $\widetilde{\mathcal{L}}_{\text{marginal}} \leftarrow \frac{1}{2}\,\text{MSE}\big(U, \text{sg}(U - g)\big)$ ▷ Marginal DMD surrogate loss
    $\mathcal{L}_{\text{joint}} \leftarrow \text{JointDMD}(z, c, \mu^v_{\text{real}}, \mu^v_{\text{fake}})$ ▷ Joint DMD loss
    $\omega_n \leftarrow \omega \min(1, n/T_{\text{warm}})$
    $\widetilde{\mathcal{L}}_{\text{DuoMatching}} \leftarrow \mathcal{L}_{\text{joint}} + \omega_n \widetilde{\mathcal{L}}_{\text{marginal}}$ ▷ Joint DMD + marginal DMD
    Update θ using $\nabla_\theta \widetilde{\mathcal{L}}_{\text{DuoMatching}}$
  end if
  Fake estimator updates
  Sample fresh text condition c' and noise $\xi' \sim \mathcal{N}(0, I)$
  $\bar{z} \leftarrow \text{sg}(G_\theta(\xi', c'))$
  $\{\bar{l}_k\}_{k=1}^K \leftarrow \text{LVS}(\bar{z}, K)$
  for $k = 1, \dots, K$ **do**
    Sample local frame index $\bar{i}_k$
    $\bar{u}_k \leftarrow \text{sg}\big(B_\phi(\bar{z}^{\bar{l}_k-1}, \bar{z}^{\bar{l}_k}, \bar{i}_k)\big)$
  end for
  $\bar{U} \leftarrow \text{Stack}(\bar{u}_1, \dots, \bar{u}_K)$
  Sample fresh image timestep τ and $\epsilon \sim \mathcal{N}(0, I)$
  $\bar{U}_\tau \leftarrow \alpha^{\text{img}}_\tau \bar{U} + \sigma^{\text{img}}_\tau \epsilon$
  Update $\mu^{\text{img}}_{\text{fake}}$ with flow-matching target $\epsilon - \bar{U}$ ▷ Fake estimator for marginal DMD
  Update μ^v_{fake} using the video flow-matching loss on $(\bar{z}, c')$ ▷ Fake estimator for joint DMD
end for