DISTILLING UNIFORM DISCRETE DIFFUSION MODELS INTO A ONE-STEP VIDEO GENERATOR

Anonymous authors
Paper under double-blind review

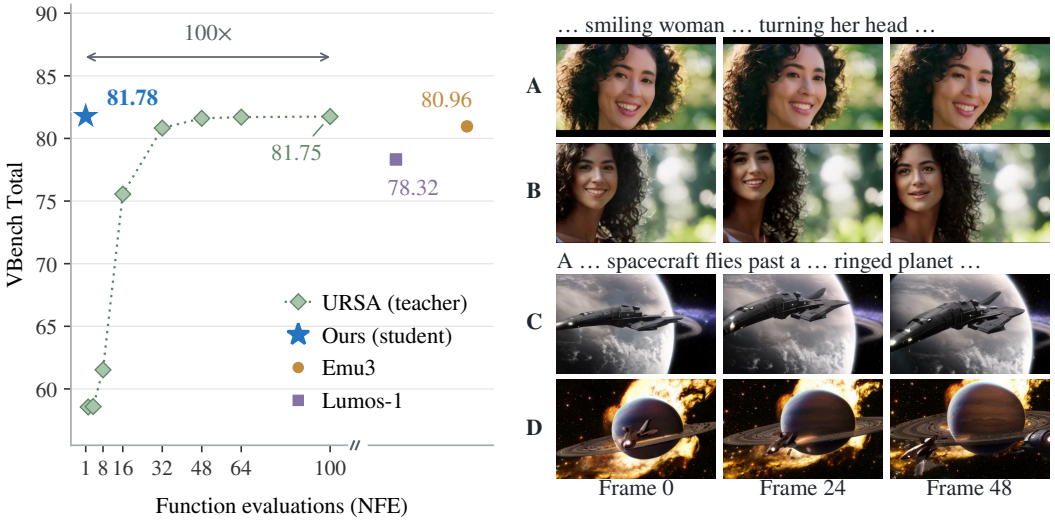


Figure 1: Comparisons with the teacher. ***Left:*** VBench score versus number of function evaluations (NFEs). Our **Discrete Distribution Matching Distillation for Video (DDMD-V)** achieves 81.78 VBench (Huang et al., 2024) Total with 1 NFE, compared with 81.75 for the URSA (Deng et al., 2026) teacher at 100 NFEs. Lumos-1 and Emu3 use reported scores from their respective evaluation protocols (Yuan et al., 2025; Wang et al., 2024); ***Right:*** **Which is which?** Among these videos, some were generated with baseline URSA (100 NFEs each, 50 sampling steps with two classifier-free guidance branches), the others with our DDMD-V (1 NFE each). Can you identify which is which? Each prompt has one teacher video and one student video, shown at frames 0, 24, and 48 with matched prompts and seeds. Row identities and complete prompts appear in App. A. (More comparisons are provided in Fig. 3).

ABSTRACT

Discrete diffusion has emerged as a competitive paradigm for visual generation, yet high-quality generation still relies on repeated refinement steps, limiting its use in low-latency interactive video generation. Compressing this process into a single forward pass is particularly challenging for video, where tens of thousands of categorical variables must jointly establish coherent spatial structure and temporal dynamics without subsequent correction. In this work, we present **DDMD-V**, to our knowledge the first one-step distillation framework for uniform discrete diffusion video generation. DDMD-V extends on-policy distribution matching to the uniform discrete diffusion setting, using student-induced intermediate states along the teacher's metric-dependent corruption path. To complement token-level distribution matching, we further introduce spatiotemporal adversarial regularization in the FSQ embedding space, providing a learned signal over multi-token spatial and temporal configurations. Distilled from URSA-1.7B, our student generates a 49-frame 320×512 video comprising 33,280 discrete tokens in a single network forward pass. It achieves 81.78 VBench score compared with the 81.75

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.

score of the 50-step teacher, while reducing the number of function evaluations (NFEs) from 100 to one when counting the teacher's conditional and unconditional guidance branches. After distillation, the student achieves closely matched visual quality compared to the teacher. In contrast, directly reducing the teacher to one sampling step yields only 58.56 VBench. With BF16 inference, DDMD-V reaches a generation throughput of more than 50 video frames per second on a single H200 GPU. These results demonstrate that a strong discrete video diffusion model can be distilled into a one-step generator with comparable performance, providing a promising path toward low-latency generative video models. We will release the code and model.

1 INTRODUCTION

Generating videos with a discrete diffusion model usually requires repeated sampling steps over the same spatiotemporal representation. Common discrete corruption processes include absorbing-state masking and uniform token replacement (Austin et al., 2021). URSA (Deng et al., 2026) achieves high-quality video generation from a uniform categorical source through iterative global refinement along a metric-dependent probability path. Although each forward pass predicts all tokens in parallel, high-quality generation still relies on multiple refinement steps. As shown in Fig. 1, directly reducing URSA's sampling budget substantially degrades its VBench (Huang et al., 2024) score. Removing this iterative refinement cost while retaining generation quality is therefore an important objective for low-latency discrete video generation and future interactive world model systems.

One-step distillation provides a way to learn a direct generator from an iterative teacher. In continuous diffusion, Distribution Matching Distillation (DMD) (Yin et al., 2024b) trains one-step image generators by matching the student distribution to a pretrained diffusion teacher, while DMD2 (Yin et al., 2024a) improves training through more accurate tracking of the evolving student distribution and adversarial supervision. One-step video generation has also been explored for continuous diffusion models through adversarial post-training (Lin et al., 2025), while recent methods study few-step video distillation through trajectory or transition matching (Nie et al., 2026). In discrete diffusion, Di[M]O (Zhu et al., 2025) and Soft-Di[M]O (Zhu et al., 2026b) distill masked image diffusion models (Bai et al., 2024) into one-step generators, and discrete moment matching distillation (Hoogeboom et al., 2026) studies efficient generation for discrete text and image models with both uniform and masked processes.

However, extending distribution matching to discrete video poses additional challenges. Specifically, a target video generator must model the joint distribution of tens of thousands of categorical video tokens, capturing both spatial structure and temporal dependencies in a single forward pass. Without iterative refinement, these predictions must establish coherent appearance and motion without subsequent opportunities for correction. A token-level distribution-matching signal alone may therefore leave important multi-token spatial and temporal dependencies underconstrained. To our knowledge, one-step distillation of discrete diffusion models for video generation has not yet been demonstrated.

In this work, we address this challenge with **Discrete Distribution Matching Distillation for Video (DDMD-V)**, which distills URSA into a generator that produces an entire video token grid from a uniform categorical source in a single forward pass while keeping similar performance.

First, we introduce on-policy categorical distribution matching for uniform discrete diffusion. The student first generates complete videos from its current distribution, and we construct supervision states by applying URSA's native metric-dependent corruption process to these student samples. At each corrupted state, a frozen guided teacher and a trainable auxiliary denoiser predict clean-token distributions. Their difference provides a distribution-matching update directly to the student's categorical logits, without differentiating through sampled token identities. Because the auxiliary must continually track the evolving student distribution, we update it more frequently than the student and sample corruption times from an intermediate region of the teacher path. Our ablations show that strong one-step performance depends on sufficient coverage of heavily corrupted states, while extending supervision toward the near-clean endpoint provides little additional benefit.

Second, we introduce spatiotemporal adversarial regularization to provide supervision beyond token-wise categorical predictions. A patch discriminator compares local spatiotemporal configurations of student videos against cached multi-step teacher samples in the embedding space of URSA's finite scalar quantization (FSQ) tokenizer. Expected FSQ embeddings provide a differentiable interface from the discriminator back to categorical student logits. The resulting objective complements categorical distribution matching with a learned signal over multi-token spatial and temporal structure, which is particularly important when an entire video must be generated without iterative correction. Our ablations show less VBench degradation with the spatiotemporal adversarial regularization enabled than with it disabled.

Distilled from URSA-1.7B, our student is able to generate 49-frame videos at 320×512 resolution in a single forward pass and achieves a VBench Total score of 81.78, compared with 81.75 for the 50-step teacher (Huang et al., 2024). As shown on the left of Fig. 1, counting conditional and unconditional guidance branches separately, this reduces the number of function evaluations (NFEs) from 100 to one. In contrast, directly reducing the teacher to a single sampling step uses two NFEs and produces a score of only 58.56. The paired examples on the right compare teacher and student clips using matched prompts and corresponding frame indices. Full prompts and row identities are provided in the App. A. These results demonstrate that a strong multi-step uniform discrete video diffusion model can be distilled into a one-step video generator with comparable performance. The main contributions of this work are:

- We introduce DDMD-V, to our knowledge the first one-step distillation framework for uniform discrete diffusion video generation, producing an entire video token grid in a single forward pass with comparable performance on visual quality.
- We develop an on-policy categorical distribution matching objective together with spatiotemporal adversarial regularization in FSQ embedding space, providing complementary token-level and multi-token supervision for one-step video generation.
- Our distillation results on URSA obtain 81.78 VBench versus 81.75 for the 50-step teacher with a $100\times$ reduction in number of function evaluations (NFEs), demonstrating the feasibility of our proposed one-step distillation for uniform discrete video diffusion models.

2 Related Work

Discrete visual generation. Discrete visual models (Austin et al., 2021; Gat et al., 2024) represent images and videos as categorical tokens and generate them through autoregressive prediction or iterative refinement. Masked generative transformers enable parallel token prediction for images (Chang et al., 2022; 2023; Bai et al., 2024; Li et al., 2026b) and videos (Yu et al., 2023), while token-based autoregressive models such as VideoPoet and Emu3 extend next-token prediction to multimodal visual generation (Kondratyuk et al., 2023; Wang et al., 2024). Coarse-to-fine autoregressive modeling further reduces sequential decoding: VAR introduces next-scale generation (Tian et al., 2024), followed by Infinity (Han et al., 2025; Liu et al., 2026b) and GRN (Han et al., 2026). Recently, URSA (Deng et al., 2026) applies a metric-dependent path (Shaul et al., 2025) to uniform discrete video diffusion, and we study whether this iterative refinement can be compressed into a single forward pass.

Distillation of discrete diffusion models. Distribution Matching Distillation (DMD) (Yin et al., 2024b) trains one-step continuous generators by matching their generated distribution to a pretrained diffusion model, while DMD2 (Yin et al., 2024a) improves tracking of the evolving student distribution and incorporates adversarial supervision. Recent work extends fast generation to discrete models. Di4C (Hayakawa et al., 2024) models correlations between output dimensions for few-step discrete diffusion. Di[M]O (Zhu et al., 2025) adapts on-policy distribution matching to masked discrete diffusion using corrupted student samples and an auxiliary denoiser, while Soft-Di[M]O (Zhu et al., 2026b) introduces differentiable expected token embeddings for adversarial and reward-based refinement. Other formulations include discrete moment matching beyond individual token predictions (Hoogeboom et al., 2026). These methods establish one- and few-step discrete generation primarily for images. We instead consider uniform discrete diffusion for video.

Few-step and one-step video generation. Recent work has substantially reduced the sampling cost of continuous video diffusion and flow models. DOLLAR and TMD explore reward-augmented distillation and transition matching for few-step generation (Ding et al., 2025; Nie et al., 2026). For causal video models, Self Forcing and related methods reduce the train-inference mismatch through student-generated rollouts and improved few-step training schemes (Huang et al., 2026; Zhu et al., 2026a; Zheng et al., 2026). At the one-step limit, APT (Lin et al., 2025) and related methods (Yang et al., 2026; Cheng et al., 2026) demonstrate the effectiveness of adversarial supervision for continuous video generation. More recent methods such as AAD-1 (Li et al., 2026a) and One-Forcing (Feng et al., 2026) combine distribution-matching or distillation objectives with video-level adversarial signals to address motion collapse and temporal inconsistency. Our setting differs in representation: we combine on-policy categorical distribution matching with a differentiable adversarial objective in FSQ space, enabling both token-level teacher transfer and learned supervision over multi-token spatiotemporal structure.

3 METHOD

3.1 PRELIMINARIES

Most diffusion-based visual generators (Rombach et al., 2022; Podell et al., 2024; Labs et al., 2025; Esser et al., 2024) operate in continuous latent spaces, where Gaussian noise is progressively transformed into images or videos through iterative denoising. An alternative line of work performs diffusion over discrete visual tokens (Chang et al., 2022; 2023; Bai et al., 2024; Li et al., 2026b; Deng et al., 2026), enabling generation in the same categorical representation space used by token-based generative models. Depending on how clean tokens are corrupted toward the noise endpoint, discrete diffusion models commonly adopt either an *absorbing kernel* or a *uniform kernel* (App. B). In this work, we focus on one-step distillation of uniform discrete diffusion for video generation. Fig. 2 summarizes our method.

3.2 ONE-STEP DISCRETE VIDEO GENERATOR

Let N denote the number of tokens in a complete video representation, K the vocabulary size, and c the text condition. We draw a source grid $\xi \sim \mathrm{Unif}([K])^N$, and let the one-step student G_θ predict logits $z_\theta(\xi, c) \in \mathbb{R}^{N\times K}$. For each position n,

$$\pi_{\theta,n} = \mathrm{softmax}(z_{\theta,n}), \qquad \hat{x}_n \sim \mathrm{Cat}(\pi_{\theta,n}). \tag{1}$$

Although the final token draws are conditionally factorized, their distributions are jointly determined by the shared source grid ξ, the conditioning c, and the full-context computation of G_θ. Marginalizing the source therefore defines the generator G_θ induced distribution:

$$p_\theta(x \mid c) = \mathbb{E}_\xi \left[\prod_{n=1}^{N} \pi_{\theta,n}(x_n \mid \xi, c) \right]. \tag{2}$$

At inference time, this single network evaluation followed by categorical sampling and decoding constitutes the complete generation process.

DDMD-V trains this generator with two complementary signals. First, on-policy categorical distribution matching transfers the teacher's denoising predictions around states induced by the current student. Second, a differentiable spatiotemporal adversarial objective provides field-level supervision over multi-token video structure.

3.3 ON-POLICY CATEGORICAL DISTRIBUTION MATCHING

The one-step student does not visit the intermediate states traversed by the iterative teacher. We therefore construct supervision states from samples generated by the current student itself. This places teacher supervision around the student's evolving output distribution rather than along a fixed teacher trajectory.

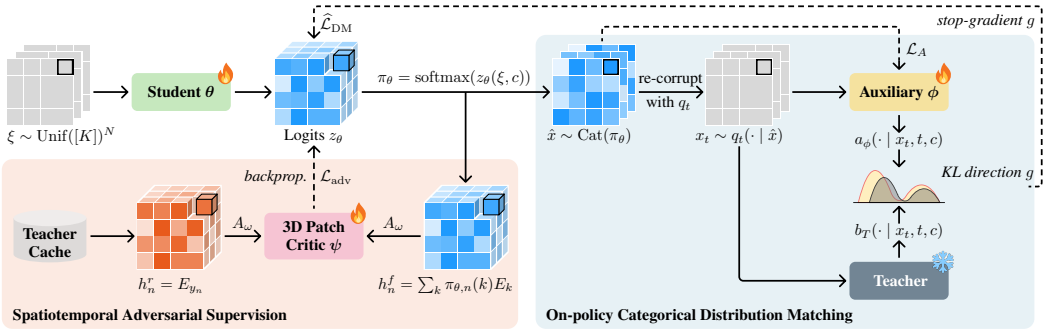


Figure 2: **Overview of DDMD-V.** The one-step student maps a uniform categorical source $\xi \sim \mathrm{Unif}([K])^N$ to logits z_θ. The two training branches share the one-step student but use independent minibatches. **(A)**: For *on-policy categorical distribution matching* (right), we sample student tokens $\hat{x}$ from $\pi_\theta = \mathrm{softmax}(z_\theta)$ and re-corrupt them along the teacher's native probability path to obtain x_t. A frozen teacher and a trainable auxiliary denoiser predict clean-token distributions from the same corrupted state. The auxiliary is trained to reconstruct $\hat{x}$, while the detached teacher-auxiliary KL direction g defines the surrogate student update $\widehat{\mathcal{L}}_{\mathrm{DM}}$ on the original logits. **(B)**: For *spatiotemporal adversarial supervision* (left), student probabilities form differentiable expected FSQ fields $h_n^f = \sum_k \pi_{\theta,n}(k) E_k$, which are compared with cached teacher fields $h_n^r = E_{y_n}$ (E is the FSQ embedding matrix) after shared token replacement A_ω using a 3D patch critic D_ψ. The adversarial loss backpropagates through the expected-FSQ field to the student.

Student-induced intermediate states. We first sample a complete student video grid $\hat{x} \sim p_\theta(\cdot \mid c)$. We then sample a corruption time t from an intermediate interval of the teacher path and re-corrupt the generated video using the teacher's native forward process:

$$x_t \sim q_t(\cdot \mid \hat{x}). \tag{3}$$

Tracking the student distribution. A frozen teacher and a trainable auxiliary denoiser are evaluated at the same corrupted state:

$$b_n = p_{T,n}(\cdot \mid x_t, t, c), \qquad a_n = p^{\phi}_{A,n}(\cdot \mid x_t, t, c). \tag{4}$$

For timestep-free architectures, the explicit dependence on t is simply omitted.

The auxiliary denoiser is trained to reconstruct the sampled student video:

$$\mathcal{L}_A(\phi) = \mathbb{E}\left[-\frac{1}{N}\sum_{n=1}^{N} \log p^{\phi}_{A,n}(\hat{x}_n \mid x_t, t, c)\right]. \tag{5}$$

It therefore tracks the corrupted distribution induced by the current generator, while the frozen teacher provides the target denoising field. Since the student distribution evolves throughout training, we update the auxiliary more frequently than the student.

Categorical distribution-matching direction. We compare teacher and auxiliary predictions with a weighted bidirectional KL:

$$\mathcal{D}_\eta(a, b) = (1-\eta) D_{\mathrm{KL}}(b \| a) + \eta D_{\mathrm{KL}}(a \| b). \tag{6}$$

Let ℓ_A denote the auxiliary logits and define $r_n = \log a_n - \log b_n$. The derivative with respect to the auxiliary logits has the closed form

$$g_n = (1-\eta)(a_n - b_n) + \eta\, a_n \odot (r_n - \langle a_n, r_n \rangle \mathbf{1}). \tag{7}$$

Categorical sampling and corruption are non-differentiable. We therefore detach g_n and transfer this conditional-divergence direction to the student logits at the corresponding original source:

$$\widehat{\mathcal{L}}_{\mathrm{DM}}(\theta) = \frac{1}{BN}\sum_{b=1}^{B}\sum_{n=1}^{N} \langle z_{\theta,n}(\xi_b, c_b), \mathrm{sg}(g_{b,n}) \rangle, \tag{8}$$

where B is the minibatch size and sg denotes stop-gradient.

Eq. 7 is the exact derivative of the conditional categorical divergence with respect to the auxiliary logits; Eq. 8 uses it as a surrogate update for the one-step student. App. C.1 derives the categorical gradient and clarifies the scope of this surrogate construction.

3.4 Spatiotemporal Adversarial Supervision

Distribution matching provides per-position categorical supervision, but a one-step video generator must also form coherent multi-token configurations across space and time without subsequent refinement. We therefore complement categorical distribution matching with a learned adversarial objective.

Differentiable token-embedding fields. Let $E \in \mathbb{R}^{K\times d}$ denote the embedding table associated with the discrete visual representation with d the embedding dimension. We maintain a cache of teacher-generated examples $\mathcal{C} = \{(c_i, \xi_i, y_i)\}$, where y_i is obtained from the iterative teacher conditioned on c_i and source ξ_i. For adversarial training, the student is evaluated using the same (c_i, ξ_i) and these teacher samples are used as adversarial real examples.

A hard categorical sample does not provide a differentiable interface from a field-level objective to the student logits. We therefore use the expected token embedding at each position:

$$h^f_{\theta,n} = \sum_{k=1}^{K} \pi_{\theta,n}(k \mid \xi_i, c_i) E_k, \qquad h^r_n = E_{y_{i,n}}. \tag{9}$$

After reshaping, h^f and h^r form differentiable student and teacher video fields. The adversarial gradient can therefore propagate through h^f_θ to the student's categorical logits; the corresponding Jacobian, including the effect of shared token replacement, is derived in App. C.2.

Shared token replacement. Before discrimination, we apply the same random token replacement to the real and student fields. Let $m_n \in \{0, 1\}$ denote a sampled replacement mask and $v_n \sim \mathrm{Unif}([K])$ a replacement token. For $\omega = (m, v)$, define

$$\mathcal{A}_\omega(h)_n = (1 - m_n)h_n + m_n E_{v_n}. \tag{10}$$

We use $\tilde{h}^f = \mathcal{A}_\omega(h^f)$ and $\tilde{h}^r = \mathcal{A}_\omega(h^r)$ with the same sampled ω. Sharing the augmentation prevents the discriminator from exploiting independently sampled corruption patterns as a shortcut.

Spatiotemporal patch discrimination. A 3D patch discriminator D_ψ operates on the embedded video fields and produces a grid of local spatiotemporal logits. Let $\langle\cdot\rangle_{\mathrm{patch}}$ denote averaging over these logits. We optimize the non-saturating logistic objectives

$$\begin{aligned} \mathcal{L}_D(\psi) &= \mathbb{E}\Big[\Big\langle \mathrm{softplus}\Big(D_\psi(\mathrm{sg}(\tilde{h}^f))\Big)\Big\rangle_{\mathrm{patch}} + \Big\langle \mathrm{softplus}\Big(-D_\psi(\tilde{h}^r)\Big)\Big\rangle_{\mathrm{patch}} \Big], \\ \mathcal{L}_{\mathrm{adv}}(\theta) &= \mathbb{E}\Big[\Big\langle \mathrm{softplus}\Big(-D_\psi(\tilde{h}^f)\Big)\Big\rangle_{\mathrm{patch}} \Big]. \end{aligned} \tag{11}$$

The two objectives provide complementary supervision. $\widehat{\mathcal{L}}_{\mathrm{DM}}$ transfers the teacher's categorical denoising field around student-generated states, while $\mathcal{L}_{\mathrm{adv}}$ directly evaluates multi-token spatial and temporal configurations in the generated video field.

3.5 Overall Objective and Inference

The overall student objective is

$$\mathcal{L}_S = \widehat{\mathcal{L}}_{\mathrm{DM}} + \lambda_{\mathrm{adv}}\mathcal{L}_{\mathrm{adv}} + \lambda_{\mathrm{reg}}\mathcal{R}_{\mathrm{logit}}, \tag{12}$$

where $\mathcal{R}_{\mathrm{logit}}$ is a small regularizer on the student logits used for optimization stability. Its exact form is given in App. E. We summarize the core training procedure in App. D. At inference time, only the one-step student is retained. Given a text condition c and an independent uniform categorical source ξ, the student predicts all N output-token distributions in a single forward pass. The resulting categorical grid is sampled once and decoded by the tokenizer decoder.

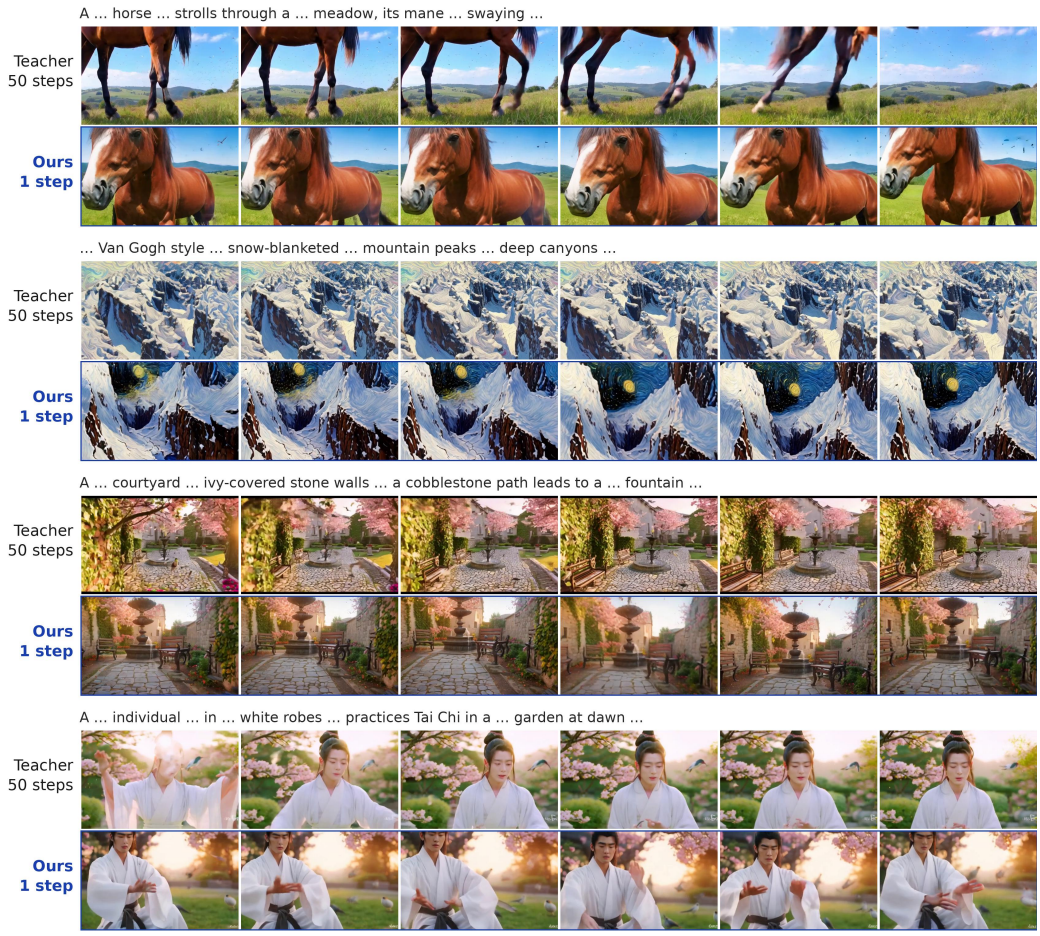


Figure 3: Each pair shows frames $0, 10, 19, 29, 38, 48$ from a 49-frame 320×512 video generated from the same prompt and seed. The 50-step URSA-1.7B teacher with CFG 7.0 (100 NFEs) is shown above, and our one-step student (1 NFE, without a separate CFG branch) below. The examples highlight differences in framing, appearance preservation, scene structure, and articulated motion. For readability, prompts are shortened to content words; complete captions are provided in App. G.

4 Experiments

4.1 Experimental Setup

Teacher and representation. We instantiate DDMD-V with URSA-1.7B (Deng et al., 2026). Both teacher and student generate 49-frame videos at 320×512 resolution. URSA represents each video using a finite scalar quantization (FSQ) tokenizer (Mentzer et al., 2024), producing a $13 \times 40 \times 64$ latent grid ($N = 33{,}280$ categorical variables) with a vocabulary of $K = 64{,}000$ tokens. The teacher uses 50 sampling steps with classifier-free guidance (CFG) scale 7, while the distilled student generates the complete token grid in one forward pass without an additional guidance branch. Counting the teacher's conditional and unconditional branches separately, this corresponds to 100 teacher NFEs versus one student NFE.

Training. Following URSA, training prompts are sampled from Koala-36M (Wang et al., 2025). For distribution matching, we sample corruption times as $t \sim \mathcal{U}(0.05, 0.65)$ and re-corrupt student-generated videos using URSA's native metric-dependent probability path. We use the balanced bidirectional-KL direction with $\eta = 0.5$. For adversarial supervision, we use a cache of 3,200 videos generated by the 50-step teacher and the differentiable expected-FSQ representation described in

Table 1: Text-to-video evaluation on VBench. Lumos-1 (Yuan et al., 2025), Emu3 (Wang et al., 2024), and OmniGen-AR (Wang et al., 2026) are reported by their authors. The others are evaluated under the same 16-dimension VBench evaluation settings. **Bold** indicates the best value within the URSA/DDMD-V comparison block. *Subject*, *Action* and *Object* denote Subject Consistency, Human Action and Object Class.

Model	Params	Steps	NFE	Total	Quality	Semantic	Subject	Action	Object	Color	Aesthetic	Imaging
Discrete / autoregressive video models (as reported)												
Lumos-1	3.6B	–	–	78.3	79.5	73.5	–	–	90.1	82.0	–	58.0
Emu3	8B	–	–	81.0	84.1	68.4	–	–	86.2	88.3	59.6	62.6
OmniGen-AR	0.5B	–	–	74.72	76.60	67.20	–	–	–	–	–	–
OmniGen-AR	1.5B	–	–	80.02	81.51	78.08	–	–	–	–	–	–
URSA sampling sweep under our protocol												
URSA	1.7B	1	2	58.56	67.05	24.56	89.99	3.00	12.42	75.08	29.99	29.50
URSA	1.7B	4	8	61.52	69.15	31.04	86.29	23.60	24.51	64.08	37.12	42.63
URSA	1.7B	8	16	75.55	78.17	65.10	91.40	86.60	80.49	72.94	54.65	52.13
URSA	1.7B	16	32	80.83	82.25	75.17	95.00	94.40	92.23	77.22	61.09	58.80
URSA	1.7B	32	64	81.70	82.89	**76.94**	95.69	95.60	**93.80**	79.85	61.95	61.12
URSA	1.7B	50	100	81.75	82.96	76.88	95.79	93.60	92.66	79.80	**62.10**	**62.23**
Ours	1.7B	**1**	**1**	**81.78**	**83.20**	76.10	**96.23**	**97.00**	92.56	**83.71**	60.53	59.10

Sec. 3.4. The complete training schedule, optimization details, and critic architecture are provided in App. E.

Evaluation protocols. Our primary evaluation uses the 16-dimension VBench (Huang et al., 2024) metrics following URSA (Deng et al., 2026) evaluation settings. For each evaluated model, we generate five videos for each of 946 prompts, yielding 4,730 videos at 49 frames and 320×512 resolution. Teacher and student generations use VBench's corresponding dense rewritten captions. We additionally sample 1,000 prompts from Koala-36M for VideoReward (Liu et al., 2026a) and pairwise preference comparisons between DDMD-V and URSA at 8, 16, and 50 sampling steps. Full evaluation protocols are provided in App. F.2.

4.2 Quantitative Comparisons

Table 1 compares DDMD-V with its teacher on VBench. Simply truncating the iterative sampler of URSA is ineffective: the teacher scores 58.56 with one sampling step and 61.52 with four steps. Its performance becomes competitive only after substantially more iterative refinement, reaching 81.75 at 50 steps. In contrast, DDMD-V generates videos in one forward pass and obtains a VBench Total of 81.78. Complementary VideoReward results show the same trend and are reported in App. F.1.

Wall-clock acceleration. The NFE reduction translates into substantial measured acceleration. On a single H200 in bfloat16 at batch size one, generating a 49-frame 320×512 clip end to end, including text encoding and VAE decoding, takes 72.26 seconds for the 50-step teacher and 0.93 seconds for the one-step student, a $77.6\times$ reduction in wall-clock latency.

4.3 Qualitative Comparisons

We compare paired generations from the 50-step teacher and DDMD-V using matched prompts and seeds. Fig. 3 shows representative examples spanning subject consistency, appearance, scene structure, and articulated motion. Despite using a single forward pass, the student can preserve global structure throughout the clip: it maintains the full horse within the frame, retains the requested painterly appearance, preserves scene elements such as the fountain and cobblestone path, and produces a clearer sequence of distinct Tai Chi poses.

To distinguish *distillation* from simply truncating the teacher sampler, App. G.2 visualizes the same teacher at progressively smaller sampling budgets.

Human and Model-based Evaluation. We evaluate pairwise preferences using human raters, GPT-6 Astra, and Gemini 3.8 Flash, with each comparison labeled as preferring the teacher, preferring our model, or a tie. As shown in Fig. 4, all three judges strongly prefer our one-step student to

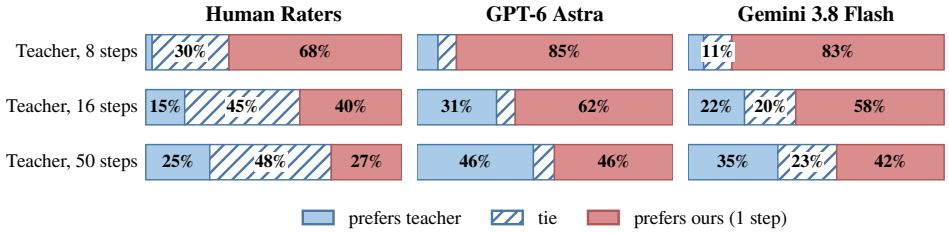


Figure 4: Pairwise preference against the multi-step teacher.

(a) Spatiotemporal critic

Variant	VBench
No critic	78.17
Temporal kernel 1	78.78
Full critic	**81.68**

(b) Corruption range

$t_{\max}$	VBench
0.65	**81.68**
0.80	81.54
0.90	81.33

(c) Corruption coverage

Sampling	VBench
Uniform full band	**81.68**
Entropy + 30% source	81.45
Entropy only	74.63

Table 2: Ablations. Values are VBench after the same training updates. (a) removes adversarial supervision or its temporal convolutional extent. (b) varies the upper bound of uniform corruption-time sampling. (c) changes the sampling distribution: the entropy band is $t \in [0.688, 0.970]$, while "30% source" assigns 30% samples to the more heavily corrupted region $t < 0.688$.

the 8-step teacher. The gap narrows with additional teacher sampling, and the student approaches preference parity with the full 50-step teacher.

4.4 Ablations

We ablate the core design choices of DDMD-V. Within each ablation block, variants differ only in the stated factor. Note these checkpoints are intended for ablations and do not correspond to the final model trained with the complete schedule. Table 2 reports the primary comparisons; broader sensitivity analyses are provided in App. H.

Spatiotemporal adversarial supervision. Removing the critic reduces VBench by 3.51 points, while removing its temporal convolutional extent costs 2.90 points. Per-dimension analysis shows that both ablations substantially degrade subject consistency, background consistency, and temporal flickering. Detailed per-dimension results are reported in App. H.1. Thus, token-level distribution matching alone is insufficient for preserving multi-token video structure in the one-step regime.

Corruption-time coverage. The one-step student benefits primarily from coverage of heavily corrupted, source-near states. Increasing $t_{\max}$ from 0.65 to 0.90 provides no measurable improvement. More importantly, when sampling over nearly the full path, restricting all training states to the entropy-defined band collapses VBench to 74.63. Restoring only 30% of samples from the heavily corrupted region recovers 6.82 points and becomes close to uniform sampling. These results show that sufficient coverage of heavily corrupted states is important, whereas extending supervision toward lightly corrupted states provides little additional benefit.

5 Conclusion

We presented **DDMD-V**, a one-step distillation framework for uniform discrete diffusion video generation. DDMD-V combines on-policy categorical distribution matching along the teacher's native corruption path with spatiotemporal adversarial regularization, enabling a student to generate an entire discrete video token grid in a single forward pass. Distilled from URSA-1.7B, our one-step model achieves comparable performance with the 50-step teacher while reducing the number of function evaluations from 100 to one. This reduction translates to a measured 77.6× wall-clock speedup. As a result, DDMD-V demonstrates that a strong uniform discrete video diffusion model can be effectively distilled into a direct generator, opening a path toward substantially lower-latency discrete video generation.

AI USE STATEMENT

In this work, we used generative AI tools for code development and debugging, feedback on experimental design, and manuscript refinement. All AI-assisted code was executed and verified by the authors, all reported numerical results were checked against experiment outputs, and all technical claims and citations were manually reviewed. The authors take full responsibility for the content of this work.

ETHICS STATEMENT

Except for a human preference study of generated videos, this work does not involve human participants. We do not collect personally identifiable information. Participants in the preference study were informed of the evaluation task and provided informed consent for their judgments to be used for research purposes. Our experiments use existing research models and datasets for text-to-video generation. As with other generative video systems, the resulting models may inherit biases from their training data and may be misused to create misleading synthetic media. We encourage responsible use of released models and code in accordance with applicable licenses and policies.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the model, training objectives, optimization settings, inference procedure, datasets, and evaluation protocol in the main paper and appendix. The appendix additionally documents implementation details and the complete ablation settings used in our experiments. Code, evaluation scripts, and model checkpoints will be released to facilitate reproduction of the reported results.

REFERENCES

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993, 2021.

Jinbin Bai, Tian Ye, Wei Chow, Enxin Song, Qing-Guo Chen, Xiangtai Li, Zhen Dong, Lei Zhu, and Shuicheng Yan. Meissonic: Revitalizing masked generative transformers for efficient high-resolution text-to-image synthesis. In *The Thirteenth International Conference on Learning Representations*, 2024.

Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11305–11315. IEEE, 2022.

Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint arXiv:2301.00704*, 2023.

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

Haoge Deng, Ting Pan, Fan Zhang, Yang Liu, Zhuoyan Luo, Yufeng Cui, Wenxuan Wang, Chunhua Shen, Shiguang Shan, Zhaoxiang Zhang, et al. Uniform discrete diffusion with metric path for video generation. In *International Conference on Learning Representations*, volume 2026, pp. 135312–135327, 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.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first international conference on machine learning*, 2024.

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.

Itai Gat, Tal Remez, Neta Shaul, Felix Kreuk, Ricky TQ Chen, Gabriel Synnaeve, Yossi Adi, and Yaron Lipman. Discrete flow matching. *Advances in Neural Information Processing Systems*, 37: 133345–133385, 2024.

Jian Han, Jinlai Liu, Yi Jiang, Bin Yan, Yuqi Zhang, Zehuan Yuan, Bingyue Peng, and Xiaobing Liu. Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15733–15744. IEEE, 2025.

Jian Han, Jinlai Liu, Jiahuan Wang, Bingyue Peng, and Zehuan Yuan. Generative refinement networks for visual synthesis. In *European Conference on Computer Vision*, pp. 566–585. Springer, 2026.

Satoshi Hayakawa, Yuhta Takida, Masaaki Imaizumi, Hiromi Wakaki, and Yuki Mitsufuji. Distillation of discrete diffusion through dimensional correlations. *arXiv preprint arXiv:2410.08709*, 2024.

Emiel Hoogeboom, David Ruhe, Jonathan Heek, Thomas Mensink, and Tim Salimans. Beyond single tokens: Distilling discrete diffusion models via discrete mmd. *arXiv preprint arXiv:2603.20155*, 2026.

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.

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. *arXiv preprint arXiv:2312.14125*, 2023.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv preprint arXiv:2506.15742*, 2025.

Haobo Li, Yanhong Zeng, Yunhong Lu, Jiapeng Zhu, Hao Ouyang, Qiuyu Wang, Ka Leong Cheng, Yujun Shen, and Zhipeng Zhang. Aad-1: Asymmetric adversarial distillation for one-step autoregressive video generation. *arXiv preprint arXiv:2606.03972*, 2026a.

Shufan Li, Greg Heinrich, Hanrong Ye, Yonggan Fu, Aditya Grover, Jan Kautz, and Pavlo Molchanov. Nemotron-labs-diffusion-image: Advancing masked discrete diffusion for high-resolution image synthesis. *arXiv preprint arXiv:2606.29814*, 2026b.

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.

Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Menghan Xia, Xintao Wang, et al. Improving video generation with human feedback. *Advances in Neural Information Processing Systems*, 38:82155–82192, 2026a.

Jinlai Liu, Jian Han, Bin Yan, Hui Wu, Fengda Zhu, Xing Wang, Yi Jiang, Bingyue Peng, and Zehuan Yuan. Infinitystar: Unified spacetime autoregressive modeling for visual generation. *Advances in Neural Information Processing Systems*, 38:170054–170072, 2026b.

Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen. Finite scalar quantization: Vq-vae made simple. In *International Conference on Learning Representations*, volume 2024, pp. 51772–51783, 2024.

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

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.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Neta Shaul, Itai Gat, Marton Havasi, Daniel Severo, Anuroop Sriram, Peter Holderrieth, Brian Karrer, Yaron Lipman, and Ricky TQ Chen. Flow matching with general discrete paths: A kinetic-optimal perspective. In *International Conference on Learning Representations*, volume 2025, pp. 16397–16429, 2025.

DiffusionGemma Team, Adrien Ali Taïga, James Assiene, Daniele Calandriello, Rahma Chaabouni, João Gante, Tamara von Glehn, Nate Keating, Chris Knutsen, Martin Kukla, et al. Diffusiongemma technical report. *arXiv preprint arXiv:2608.00146*, 2026.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. *Advances in neural information processing systems*, 37:84839–84865, 2024.

Junke Wang, Xun Wang, Qiushan Guo, Peize Sun, Weilin Huang, Zuxuan Wu, and Yu-Gang Jiang. Omnigen-ar: Autoregressive any-to-image generation. *Advances in Neural Information Processing Systems*, 38:53050–53071, 2026.

Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, et al. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8428–8437. IEEE, 2025.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024.

Yongqi Yang, Huayang Huang, Xu Peng, Xiaobin Hu, Donghao Luo, Jiangning Zhang, Chengjie Wang, and Yu Wu. Towards one-step causal video generation via adversarial self-distillation. In *International Conference on Learning Representations*, volume 2026, pp. 34858–34876, 2026.

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.

Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10459–10469. IEEE, 2023.

Hangjie Yuan, Weihua Chen, Jun Cen, Hu Yu, Jingyun Liang, Shuning Chang, Zhihui Lin, Tao Feng, Pengwei Liu, Jiazheng Xing, et al. Lumos-1: On autoregressive video generation from a unified model perspective. *arXiv e-prints*, pp. arXiv–2507, 2025.

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. *arXiv preprint arXiv:2606.25473*, 2026.

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

Yuanzhi Zhu, Xi Wang, Stéphane Lathuilière, and Vicky Kalogeiton. Di [m] o: Distilling masked diffusion models into one-step generator. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18606–18618. IEEE, 2025.

Yuanzhi Zhu, Xi Wang, Stéphane Lathuilière, and Vicky Kalogeiton. Soft-di [m] o: Improving one-step discrete image generation with soft embeddings. In *International Conference on Learning Representations*, volume 2026, pp. 2884–2931, 2026b.

A TEASER IDENTITIES AND COMPLETE PROMPTS

In both teaser alternatives, rows A and D are generated by URSA (teacher, 100 NFE), and rows B and C by our student (1 NFE).

The text above the videos retains verbatim spans of the prompts, with ellipses indicating omissions. The complete strings below include the `motion=9.0,` prefix used during generation. Both models generate 49 frames at 320×512 and 12 fps, using the same prompt and seed for each pair. The teacher uses 50 sampling steps with guidance scale 7; the student uses one step with guidance scale 1. NFE counts conditional and unconditional branches separately.

Rows A and B: "motion=9.0, A cinematic close-up portrait of a smiling woman with curly dark hair, her entire face visible, turning her head gently toward the camera in a sunlit garden, soft background bokeh."

Rows C and D: "motion=9.0, A sleek spacecraft flies past a vast ringed planet, illuminated clouds and distant stars in the background, cinematic tracking shot."

B DISCRETE DIFFUSION BACKGROUND

We give the categorical transition kernels underlying the distinction between absorbing kernel and uniform kernel discrete diffusion.

Absorbing kernel. Let the vocabulary contain K data tokens and an additional absorbing mask token M. For a clean token $i \in [K]$,

$$q_t^{\text{abs}}(j \mid i) = \alpha_t \mathbf{1}[j = i] + (1 - \alpha_t)\mathbf{1}[j = \text{M}], \qquad \alpha_0 = 0,\ \alpha_1 = 1. \tag{13}$$

Equivalently, its transition matrix is

$$Q_t^{\text{abs}} = \begin{bmatrix} \alpha_t I_K & (1 - \alpha_t)\mathbf{1}_K \\ \mathbf{0}_K^\top & 1 \end{bmatrix}. \tag{14}$$

At $t = 0$, every data token maps deterministically to the mask state. Masked visual generators such as Meissonic (Bai et al., 2024) and one-step distillation methods such as Di[M]O and Soft-Di[M]O (Zhu et al., 2025; 2026b) operate in this source setting.

Uniform kernel. Uniform diffusion (Team et al., 2026) instead corrupts clean tokens toward the uniform categorical distribution over the vocabulary:

$$q_t^{\text{uni}}(j \mid i) = \alpha_t \mathbf{1}[j = i] + \frac{1 - \alpha_t}{K}, \tag{15}$$

or, in matrix form,

$$Q_t^{\text{uni}} = \alpha_t I_K + (1 - \alpha_t)\frac{1}{K}\mathbf{1}\mathbf{1}^\top. \tag{16}$$

Its source distribution at $t = 0$ is therefore

$$x_0^n \sim \text{Unif}([K]) \tag{17}$$

independently at each token position. URSA (Deng et al., 2026) preserves this uniform categorical source while using a metric-dependent transition path over the visual codebook rather than the isotropic interpolation above.

C DERIVATIONS AND GRADIENT INTERFACES

We provide the derivations underlying the categorical distribution-matching direction and the differentiable interfaces between categorical student predictions and the spatiotemporal critic. All categorical probability vectors below are produced by softmax and therefore have strictly positive entries.

C.1 Categorical Gradient and Surrogate Student Update

Let

$$a = \mathrm{softmax}(\ell_A), \qquad b \in \Delta^{K-1},$$

where b is treated as fixed. The softmax Jacobian is

$$J(a) = \frac{\partial a}{\partial \ell_A} = \mathrm{diag}(a) - aa^\top. \tag{18}$$

For the forward conditional KL,

$$D_{\mathrm{KL}}(b\|a) = \sum_k b_k \log \frac{b_k}{a_k}, \tag{19}$$

we have

$$\nabla_{\ell_A} D_{\mathrm{KL}}(b\|a) = J(a)\left(-\frac{b}{a}\right) \tag{20}$$

$$= a - b, \tag{21}$$

where the division is element-wise and we used $\sum_k b_k = 1$.

For the reverse conditional KL, define

$$r = \log a - \log b.$$

Since

$$\nabla_a D_{\mathrm{KL}}(a\|b) = r + \mathbf{1},$$

we obtain

$$\nabla_{\ell_A} D_{\mathrm{KL}}(a\|b) = J(a)(r + \mathbf{1}) \tag{22}$$

$$= J(a)r \tag{23}$$

$$= a \odot (r - \langle a, r\rangle \mathbf{1}), \tag{24}$$

where the second equality follows from $J(a)\mathbf{1} = 0$.

Therefore, for the weighted bidirectional divergence

$$\mathcal{D}_\eta(a, b) = (1-\eta) D_{\mathrm{KL}}(b\|a) + \eta D_{\mathrm{KL}}(a\|b), \tag{25}$$

its exact derivative with respect to the auxiliary logits is

$$\nabla_{\ell_A} \mathcal{D}_\eta(a, b) = (1-\eta)(a - b) + \eta\, a \odot (r - \langle a, r\rangle \mathbf{1}), \tag{26}$$

which gives Equation 7.

Scope of the student surrogate. Categorical sampling $\hat{x} \sim \mathrm{Cat}(\pi_\theta)$ and subsequent corruption $x_t \sim q_t(\cdot \mid \hat{x})$ are treated as non-differentiable. After evaluating the teacher and auxiliary at x_t, we detach the resulting direction g and define

$$\widehat{\mathcal{L}}_{\mathrm{DM}} = \frac{1}{BN} \sum_{b=1}^{B} \sum_{n=1}^{N} \langle z_{\theta,n}(\xi_b, c_b), \mathrm{sg}(g_{b,n}) \rangle. \tag{27}$$

Because g is stop-gradient in this transfer,

$$\nabla_\theta \widehat{\mathcal{L}}_{\mathrm{DM}} = \frac{1}{BN} \sum_{b,n} \left(\frac{\partial z_{\theta,n}(\xi_b, c_b)}{\partial \theta} \right)^\top \mathrm{sg}(g_{b,n}). \tag{28}$$

Thus, Equation 7 is an exact conditional categorical-divergence derivative with respect to the auxiliary logits, whereas Equation 8 defines a surrogate student update by transferring this detached direction to the student's logits at the corresponding source.

C.2 Expected-Embedding Gradient

Consider one token position with student logits z, probabilities

$$\pi = \mathrm{softmax}(z),$$

and fixed token embeddings $\{E_k\}_{k=1}^{K}$. The differentiable student field used in Equation 9 is

$$h = \sum_{k=1}^{K} \pi_k E_k = \mathbb{E}_{x\sim \mathrm{Cat}(\pi)}[E_x]. \tag{29}$$

Using

$$\frac{\partial \pi_k}{\partial z_j} = \pi_k \left(\mathbf{1}[k=j] - \pi_j\right), \tag{30}$$

the derivative of the embedding field with respect to logit z_j is

$$\frac{\partial h}{\partial z_j} = \sum_k E_k \frac{\partial \pi_k}{\partial z_j} \tag{31}$$

$$= \pi_j E_j - \pi_j \sum_k \pi_k E_k \tag{32}$$

$$= \pi_j (E_j - h). \tag{33}$$

Hence a differentiable loss on the expected embedding field provides a well-defined gradient to the categorical student logits without differentiating through a sampled token identity.

In the actual adversarial branch, shared token replacement applies

$$\widetilde{h}_n = (1 - m_n) h_n + m_n E_{v_n}.$$

For fixed sampled (m_n, v_n),

$$\frac{\partial \widetilde{h}_n}{\partial z_{n,j}} = (1 - m_n)\, \pi_{n,j} (E_j - h_n). \tag{34}$$

Thus replaced positions contribute no adversarial gradient to the student, while unreplaced positions retain the expected-embedding Jacobian.

C.3 Straight-Through Token Interface

For the support-matched critic interface, sample

$$s \sim \mathrm{Cat}(\pi)$$

and treat the sampled index s as detached from the logits. Let h denote the expected embedding from Equation 29. We define the straight-through field as

$$h^{\mathrm{ST}} = E_s + h - \mathrm{sg}(h). \tag{35}$$

Numerically, the forward value is

$$h^{\mathrm{ST}} = E_s, \tag{36}$$

so the critic observes a valid sampled codebook embedding.

Under the straight-through backward estimator, s is treated as non-differentiable and $\mathrm{sg}(h)$ has zero derivative. Consequently,

$$\frac{\partial h^{\mathrm{ST}}}{\partial z} = \frac{\partial h}{\partial z}. \tag{37}$$

The straight-through and expected-embedding interfaces therefore use the same local surrogate Jacobian with respect to the student logits, while presenting different forward values to the critic.

Algorithm 1 DDMD-V training.

Require: frozen teacher T; student G_θ; auxiliary denoiser A_ϕ; critic D_ψ; prompt distribution $\mathcal{D}$; teacher cache $\mathcal{C} = \{(c, \xi, y)\}$; embedding table E; teacher corruption process q_t; update ratio R

repeat
  Initialize accumulated student gradient to zero
  for $r = 1, \dots, R$ **do**
  // On-policy categorical distribution matching
    Sample $c \sim \mathcal{D}$ and $\xi \sim \mathrm{Unif}([K])^N$
    Compute $\pi_\theta(\xi, c) = \mathrm{softmax}(G_\theta(\xi, c))$ and sample $\hat{x} \sim \mathrm{Cat}(\pi_\theta)$
    Sample corruption level t and $x_t \sim q_t(\cdot \mid \hat{x})$
    Compute frozen teacher predictions $b = T(x_t, c)$
    Update ϕ using $\mathcal{L}_A(\hat{x}, x_t, c)$
    Recompute $a = A_\phi(x_t, c)$ and obtain $g = \nabla_{\ell_A} \mathcal{D}_\eta(a, b)$
  // Spatiotemporal adversarial supervision
    Independently sample $(c', \xi', y) \sim \mathcal{C}$
    Compute student probabilities $\pi'_\theta = \mathrm{softmax}(G_\theta(\xi', c'))$
    Construct $h^f_n = \sum_k \pi'_{\theta,n}(k) E_k$ and $h^r_n = E_{y_n}$
    Apply the same token replacement $\tilde{h}^f = \mathcal{A}_\omega(h^f)$, $\tilde{h}^r = \mathcal{A}_\omega(h^r)$
    Update ψ using $\mathcal{L}_D(\tilde{h}^f, \tilde{h}^r)$
    Compute $\mathcal{L}_{\mathrm{adv}}$ with the updated critic
  // Accumulate student update
    Recompute $z_\theta(\xi, c)$ and form $\widehat{\mathcal{L}}_{\mathrm{DM}} = \frac{1}{N} \sum_n \langle z_{\theta,n}, \mathrm{sg}(g_n) \rangle$
    Accumulate the gradient of $\widehat{\mathcal{L}}_{\mathrm{DM}} + \lambda_{\mathrm{adv}} \mathcal{L}_{\mathrm{adv}} + \lambda_{\mathrm{reg}} \mathcal{R}_{\mathrm{logit}}$
  end for
  Update θ once using the accumulated gradient
until convergence
return one-step generator G_θ

D Training Algorithm

Algorithm 1 summarizes the core DDMD-V training stage. Distribution matching uses fresh on-policy student samples, whereas adversarial supervision uses an independent minibatch of cached teacher pairs.

E Additional Experimental Details

E.1 Training Schedule and Initialization

The reported model is obtained through a staged training schedule. All stages use the full 49-frame latent geometry, independent uniform categorical source grids, and the balanced bidirectional-KL direction ($\eta = 0.5$). At step 0, the student and the auxiliary denoiser are both initialized from the released URSA-1.7B checkpoint. The spatiotemporal critic is initialized randomly.

Table 3: Training schedule of the reported model.

Stage	Steps	Critic field	λ_{adv}
DM-only training	0–1500	–	0
Adversarial warmup	1500–2025	Raw Expected Embedding	0.03
LR refinement	2025–2125	Raw Expected Embedding	0.03

E.2 Corruption Sampling and Regularization

For the reported training trajectory, corruption times for distribution matching are sampled as

$$t \sim \mathcal{U}(0.05, 0.65), \tag{38}$$

where $t = 0$ denotes the uniform categorical source and larger t corresponds to less corrupted states. The corrupted grids are sampled using URSA's native metric-dependent probability path.

In addition to the distribution-matching and adversarial objectives, we apply a small maximum-logit penalty

$$\mathcal{R}_{\text{logit}} = \frac{1}{BN} \sum_{b=1}^{B} \sum_{n=1}^{N} \left(\max_{k} z_{\theta,bnk} \right)^2, \tag{39}$$

with $\lambda_{\text{reg}} = 10^{-3}$. This term acts only as an optimization regularizer on the unnormalized logits.

E.3 Optimization and Update Schedule

The student and auxiliary denoiser use AdamW with $\beta = (0.9, 0.95)$, weight decay 0.01, and cosine learning-rate decay. The nominal peak learning rate is 10^{-5} with a floor of 10^{-6} and a 50-student-update warmup. The discriminator uses AdamW with learning rate 2×10^{-4}, $\beta = (0, 0.99)$, and zero weight decay.

Student and auxiliary computations use BF16, whereas the discriminator is optimized in FP32. Training uses eight H200 GPUs with ZeRO-2. Global parameter-gradient norms are clipped at 1.

Each GPU processes two fresh distribution-matching videos per microbatch and one independently sampled cached teacher pair for adversarial training. Student gradients are accumulated over five microbatches before one student optimizer step. The auxiliary denoiser and discriminator are updated once per microbatch, resulting in auxiliary-to-student and discriminator-to-student update ratios of 5:1. Thus, one student update uses 80 fresh videos and 40 cached teacher pairs across eight GPUs.

E.4 Teacher Cache and Training Data

Fresh training prompts are sampled from Koala-36M (Wang et al., 2025), with a maximum text length of 320 tokens.

The adversarial cache contains 3,200 videos generated by URSA-1.7B using 50 sampling steps and CFG scale 7.0. Each cached item stores the text condition, its initial uniform categorical source grid, and the corresponding teacher output. The student is evaluated from the same cached source grid and text condition, and the teacher output is used only as an adversarial real example.

Before discrimination, we draw a replacement rate $\rho \sim \mathcal{U}(0, 0.5)$ and independently replace token positions with probability ρ. Real and generated fields use the same replacement mask and replacement token IDs.

E.5 Spatiotemporal Critic

The critic operates directly on raw FSQ fields of shape $6 \times 13 \times 40 \times 64$. All convolutional layers use spectral normalization. The feature layers use LeakyReLU with slope 0.2, and the three intermediate blocks use GroupNorm with 32 groups.

Table 4: Spatiotemporal critic architecture. Output shapes report latent temporal and spatial dimensions.

Layer	Channels	Kernel	Stride	Output (T, H, W)
Initial	$6 \to 96$	$(3, 4, 4)$	$(1, 2, 2)$	$(13, 20, 32)$
Block 1	$96 \to 192$	$(3, 4, 4)$	$(1, 2, 2)$	$(13, 10, 16)$
Block 2	$192 \to 384$	$(3, 4, 4)$	$(2, 2, 2)$	$(7, 5, 8)$
Block 3	$384 \to 384$	$(3, 4, 4)$	$(2, 2, 2)$	$(4, 2, 4)$
Output	$384 \to 1$	$(3, 3, 3)$	$(1, 1, 1)$	$(4, 2, 4)$

The critic contains approximately 11.5M parameters and produces 32 patch logits per video. It receives neither an explicit timestep nor a text condition.

F Additional Quantitative Evaluation

F.1 VideoReward

We additionally evaluate 1,000 Koala-36M prompts used in the preference study with VideoReward (Liu et al., 2026a). As shown in Table 5, the one-step model substantially outperforms the truncated 8- and 16-step teachers and achieves an Overall score of 0.103, closely matching the full 50-step teacher at 0.090.

Table 5: VideoReward evaluation. Results are measured on the 1,000 Koala-36M prompts used in the preference study. Higher is better.

Model	VQ	MQ	TA	Overall
URSA, 8 steps	-1.315	-0.717	0.340	-1.691
URSA, 16 steps	-0.913	-0.397	1.117	-0.193
URSA, 50 steps	-0.804	**-0.315**	1.208	0.090
Ours, 1 step	**-0.764**	-0.347	**1.214**	**0.103**

F.2 Pairwise Preference Evaluation

Evaluation set. We randomly sample 1,000 prompts from Koala-36M (Wang et al., 2025) and use the same prompt set for all pairwise preference comparisons. For each prompt, we compare our one-step DDMD-V student against the URSA teacher at 8, 16, and 50 sampling steps. Each comparison has three possible outcomes: *prefer teacher*, *tie*, or *prefer DDMD-V*. The same generated video pairs are used across the human and model-based evaluations.

Human evaluation. We recruit 20 human raters, and each rater evaluates 150 video pairs, yielding 3,000 individual human judgments. Videos are displayed without model identity and raters are asked the following question:

> "Which video looks better? Choose the video with higher visual quality and greater realism: sharpness, absence of artifacts or distortion, stable objects, plausible motion. If you genuinely cannot separate them, choose Tie."

Model-based evaluation. We additionally evaluate the same teacher-student comparisons using GPT-6 Astra and Gemini 3.8 Flash as multimodal judges. Each judge receives visual observations from the paired videos and returns one of the same three outcomes: *prefer teacher*, *tie*, or *prefer DDMD-V*. The judge is instructed to compare overall perceptual video quality.

Each video is represented by uniformly sampled frames. For a video of 49 frames, we select frame indices

$$\mathrm{round}\left(\frac{i \cdot 48}{k-1}\right), \qquad i = 0, \ldots, k-1, \tag{40}$$

at the original resolution of 320×512. We use $k = 8$ frames per video and provide the dense rewritten generation caption used at inference time.

For GPT-6 Astra, the sampled frames of video A are provided in temporal order, followed by the sampled frames of video B. For Gemini 3.8 Flash, the sampled frames of each video are tiled into a single JPEG image in temporal order (a 2×4 grid for $k = 8$), in order to satisfy the service limits of at most 10 images and 5 MiB per request.

G Qualitative Examples and Complete Prompts

For all qualitative comparisons, videos are generated from the complete conditioning captions listed below. The text displayed above each video strip in the figures is shortened to content words for readability. VBench associates generated clips with the corresponding raw benchmark prompts; our generation pipeline uses the dense rewritten captions reported here. Unless otherwise stated, all videos contain 49 frames at 320×512 resolution.

G.1 Main Teacher-Student Comparison

The examples in Fig. 3 are selected to illustrate representative differences in subject consistency, appearance preservation, scene structure, and articulated motion.

- **A horse taking a peaceful walk.** A majestic chestnut horse with a glossy coat leisurely strolls through a sun-dappled meadow, its mane gently swaying in the breeze. The scene transitions to a close-up of the horse's serene eyes, reflecting the tranquility of its surroundings. As it walks, the horse's hooves softly tread on the lush, green grass, creating a rhythmic, calming sound. The backdrop features rolling hills and a clear blue sky, with occasional birds soaring overhead. The horse pauses to graze, its movements slow and deliberate, embodying peace and contentment in the idyllic landscape.
- **Snow rocky mountains peaks canyon, Van Gogh style.** In a mesmerizing Van Gogh style, snow-blanketed rocky mountain peaks tower majestically, their rugged surfaces adorned with swirling, vibrant strokes of white and blue. Deep canyons, shadowed and mysterious, twist and bend through the high-elevated terrain, creating a labyrinth of natural beauty. The canyons' winding paths are accentuated by the dynamic, textured brushstrokes, capturing the essence of movement and depth. The entire scene is bathed in a surreal, dreamlike quality, with the snow and rock formations blending seamlessly into a tapestry of swirling colors and intricate patterns, evoking the timeless artistry of Van Gogh.
- **Courtyard.** A serene courtyard bathed in the golden glow of late afternoon sunlight, surrounded by ivy-covered stone walls and vibrant flower beds. In the center, a cobblestone path leads to an ornate, wrought-iron fountain, its gentle trickle adding to the tranquil ambiance. Wooden benches with intricate carvings are strategically placed under the shade of blossoming cherry trees, inviting quiet reflection. Birds chirp melodiously from the branches, while a gentle breeze rustles the leaves, creating a symphony of nature. The scene captures a perfect blend of rustic charm and peaceful solitude, offering a moment of escape from the hustle and bustle of daily life.
- **A person is Tai Chi.** A serene individual, dressed in flowing white robes, practices Tai Chi in a tranquil garden at dawn. The scene opens with a close-up of their calm face, eyes closed, breathing deeply. As the camera pans out, they gracefully move through a series of slow, deliberate motions, their hands and feet in perfect harmony. The garden, lush with greenery and blooming flowers, is bathed in the soft, golden light of the rising sun. Birds chirp in the background, and a gentle breeze rustles the leaves, enhancing the peaceful atmosphere. The person's movements are fluid and meditative, embodying balance and inner peace.

G.2 Teacher Sampling-Budget Comparison

Fig. 5 complements the quantitative sampling sweep in Table 1 by visualizing how direct truncation of the teacher differs from one-step distillation.

Reducing the teacher from 50 to 16 and then 8 steps produces increasingly visible degradation in geometry, appearance, and temporal consistency. In contrast, DDMD-V remains qualitatively close to the full 50-step reference while requiring only one network evaluation.

The teacher uses CFG 7.0 at every sampling budget, corresponding to two network evaluations per sampling step; DDMD-V uses one network evaluation without a separate CFG branch. The complete conditioning captions for these examples are listed below.

- **A storm trooper vacuuming the beach.** A lone stormtrooper, clad in iconic white armor, stands on a sunlit beach, holding a futuristic vacuum cleaner. The scene opens with the stormtrooper methodically vacuuming the golden sand, the ocean waves gently lapping in the background. Seagulls fly overhead, casting fleeting shadows on the pristine shore. The stormtrooper's movements are precise and deliberate, contrasting humorously with the serene beach setting. As the camera zooms in, the details of the armor gleam under the bright sunlight, and the vacuum hums softly, creating an amusing juxtaposition of sci-fi and everyday life. The scene concludes with the stormtrooper pausing to look out at the horizon, the vast ocean stretching endlessly, blending the surreal with the mundane.
- **A cat playing in park.** A playful tabby cat with striking green eyes frolics in a sunlit park, its fur glistening in the warm afternoon light. The cat pounces on a fluttering butterfly, its movements

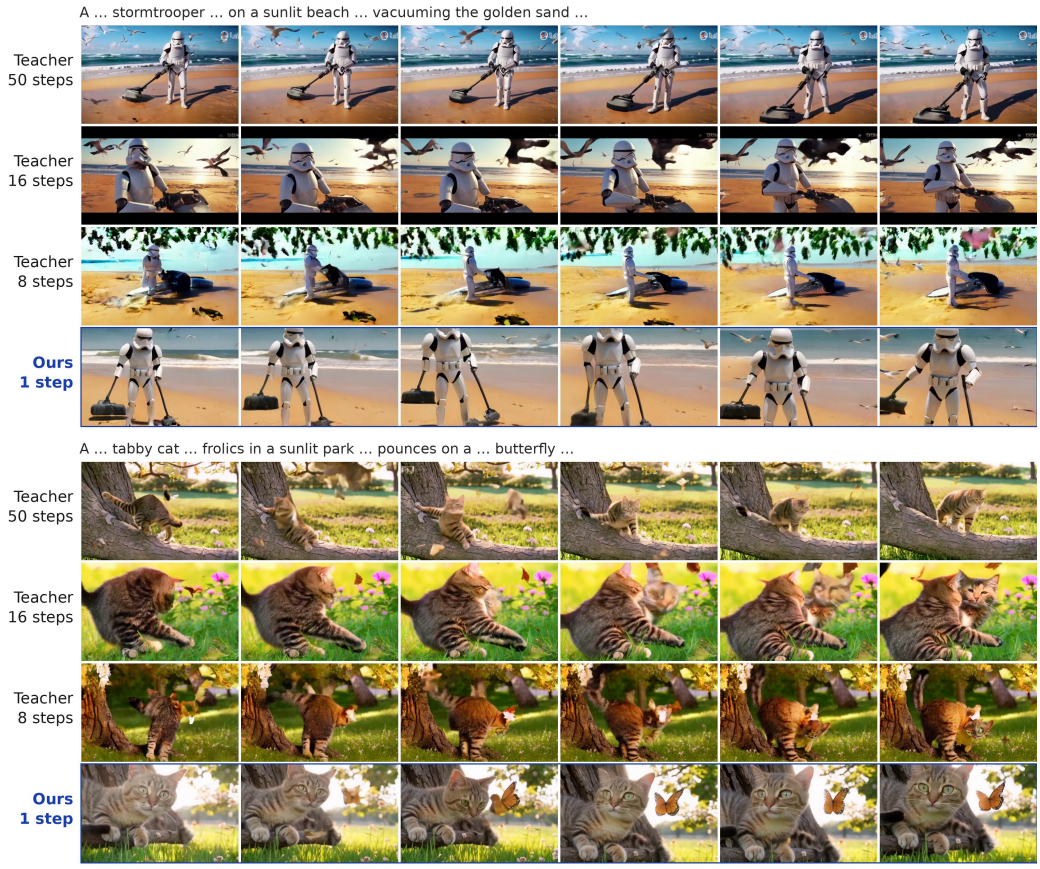


Figure 5: Rows compare the URSA teacher at 50, 16, and 8 sampling steps (100, 32, and 16 NFEs, respectively) with our one-step student at 1 NFE. Frame indices and prompt formatting follow Fig. 3. As the teacher sampling budget is reduced, visual structure and temporal consistency deteriorate, whereas the distilled one-step model remains close to the full 50-step reference. The complete teacher sampling sweep, including the 4-step setting, is reported quantitatively in Table 1.

agile and graceful, surrounded by lush green grass and blooming flowers. It then chases a falling leaf, leaping and twisting mid-air, showcasing its playful nature. The scene shifts to the cat climbing a sturdy oak tree, its claws gripping the bark as it ascends with ease. Finally, the cat rests on a low branch, its tail swaying gently, as it surveys the vibrant park, filled with the sounds of chirping birds and rustling leaves.

G.3 Additional Qualitative Examples

Figure 6 adds seven teacher-student comparisons spanning appearance style, imaginative composition, artistic style, fluid dynamics and object identity. Each pair uses the same prompt and the same sampling seed for the teacher (50 steps, 100 NFE) and for our one-step student (1 NFE). The complete conditioning captions for these examples are listed below.

- **A corgi's head depicted as an explosion of a nebula.** A corgi's head, with its adorable features, transforms into a mesmerizing cosmic explosion. The fur seamlessly blends into swirling nebulae, with vibrant hues of deep blues, purples, and pinks. Stars and galaxies twinkle within the corgi's eyes, creating an ethereal glow. Wisps of cosmic dust and gas radiate outward, forming intricate patterns that mimic the corgi's fur texture. The background is a vast expanse of space, dotted with distant stars, enhancing the surreal and otherworldly atmosphere. The entire scene captures the whimsical fusion of a beloved pet and the grandeur of the universe.
- **Origami dancers in white paper, 3D render, on white background, studio shot, dancing modern dance.** In a pristine studio with a white backdrop, intricately folded origami dancers crafted from crisp white paper come to life in a mesmerizing 3D render. These delicate figures, with sharp, precise folds, perform an elegant modern dance, their movements fluid and synchronized. The camera captures close-ups of their intricate details, highlighting the artistry of each fold. As they twirl and leap, their shadows create a subtle play of light and depth on the white background, enhancing the ethereal quality of the scene. The entire performance exudes a sense of grace and innovation, blending traditional art with contemporary dance.
- **A fat rabbit wearing a purple robe walking through a fantasy landscape.** A plump rabbit, adorned in a flowing purple robe with golden embroidery, ambles through an enchanting fantasy landscape. The rabbit's large, expressive eyes take in the vibrant surroundings, where towering mushrooms with glowing caps and bioluminescent flowers light up the path. The sky above is a swirl of pastel colors, with floating islands and waterfalls defying gravity. As the rabbit walks, its robe sways gently, revealing intricate patterns that shimmer in the magical light. The air is filled with the soft hum of mystical creatures, and the ground beneath is a mosaic of sparkling stones and lush, emerald grass.
- **A robot DJ playing the turntable on a futuristic Tokyo rooftop.** A sleek, metallic robot DJ with glowing blue eyes stands on a neon-lit rooftop in futuristic Tokyo, surrounded by towering skyscrapers adorned with holographic advertisements. The night sky is illuminated by vibrant, pulsating lights, reflecting off the rain-soaked surfaces. The robot, with intricate circuitry and mechanical arms, expertly manipulates the turntables, creating an electrifying mix. Heavy rain pours down, adding a dramatic effect as the droplets sizzle on the robot's exterior. The scene is a blend of sci-fi and fantasy, with the cityscape's cyberpunk aesthetic enhancing the surreal atmosphere. The robot's movements are precise and rhythmic, embodying the fusion of technology and artistry in this captivating, rain-drenched night.
- **A beautiful coastal beach in spring, in the style of Hokusai.** A breathtaking coastal beach scene unfolds in spring, depicted in the iconic Ukiyo-e style of Hokusai. The waves, meticulously detailed, gently lap against the golden sand, creating a rhythmic dance of water and shore. Cherry blossoms in full bloom frame the scene, their delicate petals contrasting with the deep blue of the ocean. Traditional Japanese fishing boats, with their sails billowing, dot the horizon, adding a sense of timelessness. The sky, painted in soft pastels, transitions from a serene dawn to a vibrant midday, capturing the essence of a perfect spring day by the sea.
- **A balloon full of water exploding in extreme slow motion.** A vibrant, water-filled balloon hangs suspended in mid-air against a dark backdrop, its surface glistening under the spotlight. Suddenly, a pin pierces the balloon, and in extreme slow motion, the rubber bursts apart, creating a mesmerizing cascade of water droplets. The liquid forms intricate, fleeting shapes, each droplet catching the light and sparkling like tiny diamonds. The balloon's remnants peel away, revealing

the water's graceful dance as it disperses into the air. The entire scene unfolds with breathtaking clarity, capturing the beauty and chaos of the explosion in exquisite detail.

- **An orange bicycle.** A vibrant orange bicycle stands alone on a cobblestone street, its frame gleaming under the soft morning sunlight. The bike, with its classic design, features a wicker basket on the front, filled with fresh flowers. The scene transitions to a close-up of the bike's intricate details: the shiny spokes, the leather saddle, and the vintage bell. Next, the bicycle is seen leaning against a rustic brick wall, ivy creeping up behind it, adding a touch of nature to the urban setting. Finally, the bike is captured in motion, its wheels spinning gracefully as it glides down a tree-lined path, the leaves rustling gently in the breeze.

H Additional Ablations

Within each ablation block, variants differ only in the stated factor.

H.1 Per-Dimension Analysis of Spatiotemporal Critic Ablations

To diagnose the source of the performance degradation in the critic ablation, we report the corresponding VBench dimensions related to appearance consistency and temporal behavior.

Table 6: Per-dimension analysis of the spatiotemporal critic ablation. The full spatiotemporal critic primarily improves content and temporal consistency.

Variant	Total	Subject	Background	Flicker
Spatiotemporal critic	81.68	95.77	93.63	96.19
Temporal kernel 1	78.78	89.50	91.04	94.39
No critic	78.17	88.53	90.71	94.52

Removing adversarial supervision produces the largest degradation in Subject Consistency, which falls from 95.77 to 88.53. Restricting the critic to temporal kernel size 1 yields a similar pattern, with Subject Consistency decreasing to 89.50. Background Consistency also decreases from 93.63 to 90.71 and 91.04, respectively, while Temporal Flickering drops from 96.19 to 94.52 and 94.39. Thus, the diagnostic dimensions are consistent with the spatiotemporal critic helping maintain subject, background, and temporal consistency.

H.2 Adversarial Weight

We vary the adversarial weight while keeping the others fixed.

Table 7: Adversarial-weight sensitivity.

λ_{adv}	0	0.01	**0.03**	0.10
VBench	78.17	79.45	**81.68**	81.28

A non-trivial adversarial signal is important: $\lambda_{\mathrm{adv}} = 0.03$ improves over 0.01 by 2.23 points. However, increasing the weight to 0.10 does not improve over 0.03, indicating that performance is not sensitive to the precise weight once the adversarial signal is sufficiently strong.

H.3 Distribution-Matching Direction

Our conditional distribution-matching direction uses

$$\mathcal{D}_\eta(a, b) = (1 - \eta) D_{\mathrm{KL}}(b \| a) + \eta D_{\mathrm{KL}}(a \| b).$$

We compare forward KL ($\eta = 0$), the balanced Jeffreys direction ($\eta = 0.5$), and reverse KL ($\eta = 1$).

Forward KL is worse than the balanced direction by 1.12 while reverse KL is slightly lower. We therefore use the balanced direction.

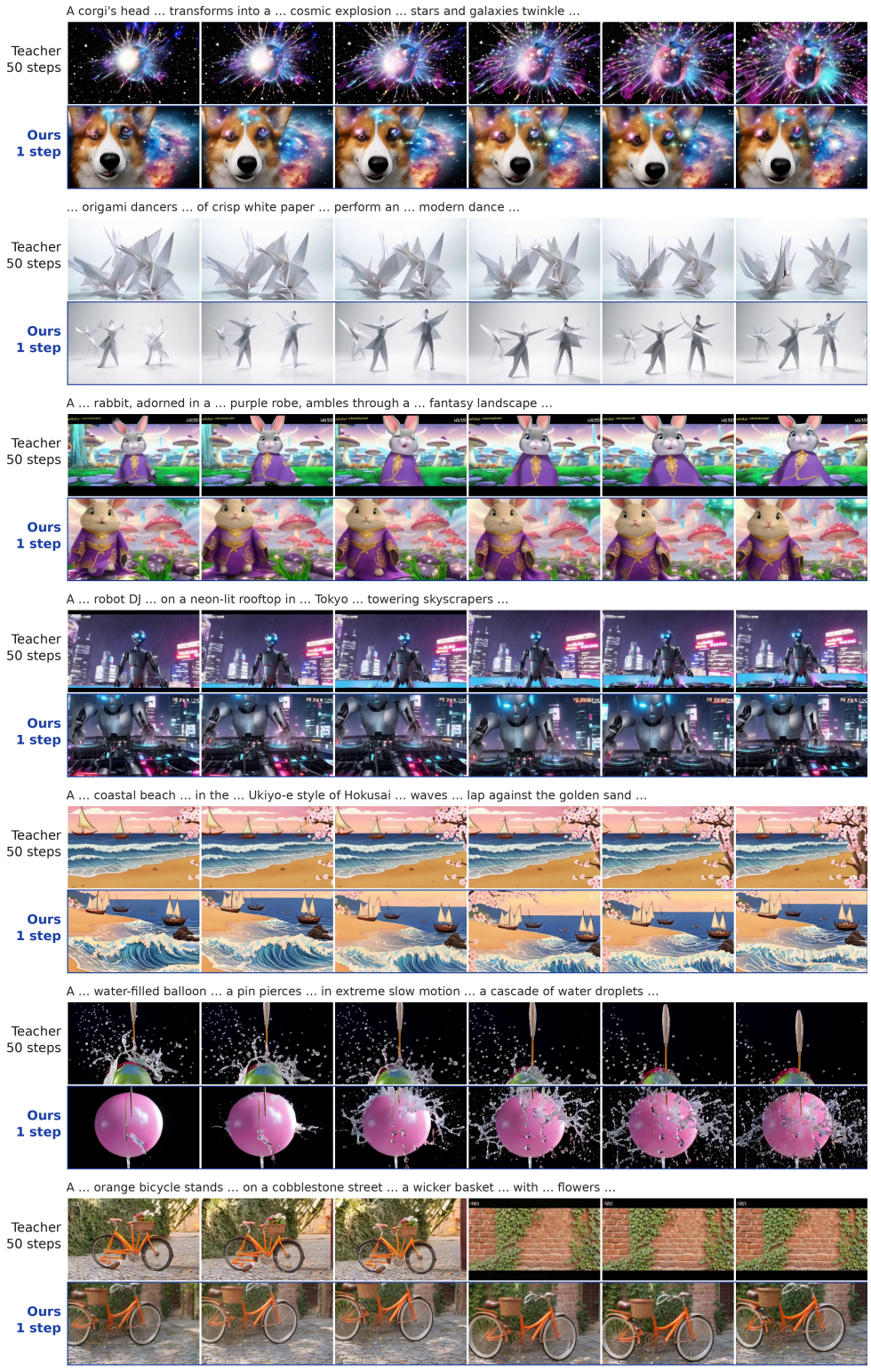


Figure 6: Additional teacher-student comparisons. Per prompt, top: teacher, 50 steps (100 NFE); bottom: ours, 1 step (1 NFE). Six frames uniformly spaced over the 49-frame clip; both rows share the prompt and the seed.

Table 8: Distribution-matching direction.

Direction	VBench
Forward KL	80.56
Jeffrey (ours)	**81.68**
Reverse KL	81.30

H.4 Metric-Aware Matching Bridge

URSA uses a metric-dependent corruption path over the FSQ codebook. We therefore test whether the same geometry should be inserted explicitly into the distribution-matching bridge. Our default formulation obtains the student-side categorical field from the trainable auxiliary denoiser p_A^{ϕ}. As an alternative, we replace this learned categorical bridge with an analytic metric kernel constructed from FSQ codebook distances, whose concentration is controlled by γ.

Table 9: Categorical versus analytic metric-aware matching. The analytic variants use FSQ metric kernels with the indicated concentration parameter γ.

Bridge	VBench
Categorical (ours)	**81.68**
Analytic metric kernel, $\gamma = 1$	81.11
Analytic metric kernel, $\gamma = 3$	80.52

Explicitly encoding the codebook geometry does not improve the learned categorical formulation in the tested settings, so we retain the auxiliary-based categorical bridge.

H.5 Critic Input Representation

Table 10: Critic input representation.

Critic input	VBench
Straight-through	81.55
Norm-matched soft	81.59
Raw expected embedding (ours)	**81.68**

The real critic input is a discrete FSQ codebook embedding, whereas the student predicts a categorical distribution. We investigate three differentiable representations for connecting student logits to the critic. For probabilities $p = \mathrm{softmax}(z)$ and FSQ embeddings $\{e_k\}$, define

$$\mu = \sum_k p_k e_k, \qquad h \sim \mathrm{Cat}(p).$$

The *raw-soft* representation directly uses μ. The straight-through (ST) representation

$$f_{\mathrm{ST}} = e_h + \mu - \mathrm{sg}(\mu)$$

presents a sampled codebook vector in the forward pass while propagating the gradient of μ. Finally, the norm-matched soft representation uses

$$f_{\mathrm{norm}} = a\mu, \qquad a = \mathrm{sg}\left(\frac{\|e_h\|}{\max(\|\mu\|, 10^{-6})}\right),$$

which preserves the soft direction while matching the magnitude of the sampled hard embedding.

The three representations perform similarly. Raw-soft exceeds ST, but its advantage is only 0.13 points. Norm matching changes performance by only about 0.09 points relative to raw-soft. We use the raw expected embedding in the final model for its simplicity and slightly stronger performance.