UNCERTAINTY-AWARE CONSISTENCY DISTILLATION FOR FEW-STEP VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study few-step video generation, *i.e.*, distilling a multi-step video generator, which typically requires tens of sampling steps, incurring substantial latency and compute, into a few-step student. Consistency distillation is a common recipe, in which a multi-step teacher provides the consistency targets for a few-step student. However, these teacher-guided targets are not equally trustworthy, and the content is harder to learn where it varies rapidly over time, *e.g.*, moving foliage shadows or flowing water. We observe that supervision reliability follows the local difficulty of the content rather than semantic complexity: regions that change little yield consistent endpoint predictions, whereas regions with large temporal variation produce larger discrepancies that coincide with the largest perceptual errors. Motivated by this observation, we propose **Uncertainty-Aware Consistency Distillation** (UACD), which reweights consistency supervision at each spatiotemporal region using a local, parameter-free uncertainty estimate. Specifically, we construct two independently perturbed teacher-guided consistency paths, whose student endpoint predictions provide a consensus target; the discrepancy between the student's direct prediction and this target is the uncertainty proxy. We then relax the consistency penalty on high-uncertainty regions through an exponential weight, while keeping the full penalty elsewhere, since the student cannot be expected to match targets that are hard to learn. To preserve perceptual quality under aggressive step reduction, we integrate feature-space adversarial training with semantic alignment. With parameter-efficient LoRA adaptation of the 50-step Wan model, our method achieves state-of-the-art 4-step generation on VBench 2.0 (0.556 mean score) and is preferred over competing methods in a user study[1].

1 INTRODUCTION

Recent advances in video diffusion models, particularly those based on Flow Matching (Song et al., 2020; Kong et al., 2024; Liu et al., 2024; Wan et al., 2025), can now synthesize high-fidelity and temporally coherent videos from text or image prompts. However, these models typically rely on iterative numerical solvers like Euler or Runge-Kutta methods to traverse the probability flow ordinary differential equation (PF-ODE), which requires tens to hundreds of neural network evaluations per video. This cost translates into high inference latency and limits deployment in real-time applications such as interactive media and game development.

Step distillation (Meng et al., 2023; Salimans & Ho, 2022) has emerged as an effective way to cut inference cost while preserving generation quality. Among existing approaches, consistency distillation (Song et al., 2023) is particularly effective: the teacher first takes a small Euler step from a noisy state, after which the student predicts the clean endpoint from the resulting intermediate state to form the distillation target. The student is then trained to produce the same endpoint directly from the original noisy state. While consistency distillation works well for image synthesis, applying it to high-dimensional video generation remains challenging: the reliability of the consistency target varies across spatiotemporal regions, and the content itself is harder to learn where it changes rapidly between frames.

Inspired by recent advances in uncertainty learning (Zhang et al., 2025; Wang et al., 2026), we observe that **not all spatiotemporal regions are equal** in terms of supervision reliability and that the

[1] Anonymous Github: `https://uacd.github.io/UACD/`

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

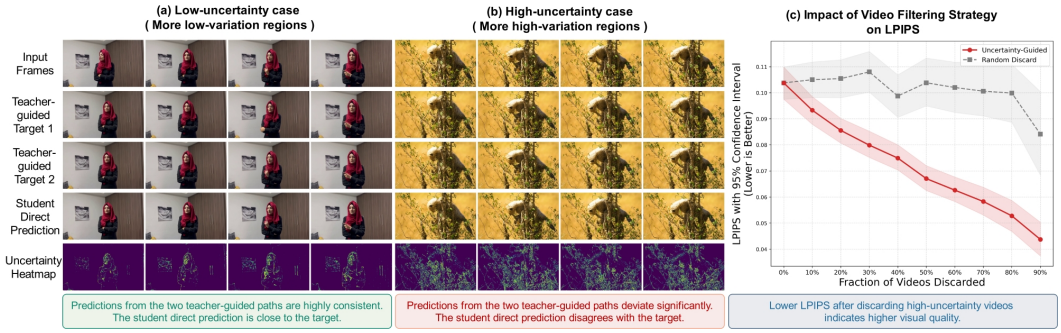


Figure 1: Uncertainty-aware supervision according to local consistency reliability. **(a)** In low-variation regions, two independently constructed teacher-guided consistency paths yield similar student endpoint predictions, producing a small student–consensus discrepancy and thus low uncertainty U (dark heatmap). **(b)** In high-variation regions, where the content is hard to learn (*e.g.*, moving foliage shadows or flowing water), the two paths produce more divergent endpoint predictions, resulting in a larger discrepancy between the student's direct prediction and the consensus target, and thus higher U (bright heatmap). **(c)** Risk-coverage analysis showing that uncertainty-guided filtering consistently reduces LPIPS (lower is better) as high-uncertainty videos are removed, while random discarding has little effect. Our loss relaxes the consistency penalty where uncertainty is high and keeps the full penalty elsewhere.

local reliability is tied to how hard the content is to learn. As illustrated in Figure 1, panels (a) and (b) present the ground-truth videos, two independently perturbed teacher-guided targets, the student's direct prediction, and the resulting uncertainty. Panels (a) and (b) correspond to a low-variation case and a high-variation case, respectively. In the low-variation case, the two teacher-guided predictions and the direct prediction remain consistent, resulting in a stable consensus target and low uncertainty. In contrast, for the high-variation case, these predictions exhibit larger discrepancies, leading to a less reliable consensus target and higher uncertainty. Figure 1(c) presents a risk-coverage analysis on 300 generated videos, where videos with the highest mean uncertainty $\bar{U}$ are progressively removed and LPIPS (Zhang et al., 2018) is measured at each coverage level. Compared with random discarding, uncertainty-guided filtering reduces the average LPIPS of the remaining videos, indicating that $\bar{U}$ tracks perceptual error and can identify, without access to ground-truth frames, the videos on which distillation is least reliable. These are the videos whose content varies most over time, which suggests that temporal variation, rather than semantic complexity, governs the reliability of consistency supervision. However, conventional consistency distillation applies the same penalty to every spatiotemporal region, regardless of how reliable its target is. Regions that are hard to learn therefore contribute as much gradient signal as regions that are easy to learn, forcing the student to match targets that are themselves unreliable and degrading generation fidelity.

To address this issue, we propose **Uncertainty-Aware Consistency Distillation**, a few-step video generation framework that adapts the consistency supervision according to fine-grained spatiotemporal uncertainty. Instead of relying on a single teacher-guided perturbation path, we introduce a dual-path consistency construction that enables parameter-free estimation of local consistency uncertainty. Specifically, we construct two independently perturbed noisy states from the same clean latent using two independent noise samples. The teacher performs a one-step Euler update from each perturbed state, producing two distinct intermediate states. The student then predicts the final endpoint from each intermediate state in a single forward pass, yielding two student endpoint predictions whose average forms a consensus target. In parallel, one of the two perturbed states is randomly selected as the input to the direct student prediction. The discrepancy between the student's direct endpoint prediction and the consensus target provides a simple, parameter-free proxy U for local consistency uncertainty, which we then normalize to $\tilde{U}$ for adaptive supervision weighting. This proxy indicates the reliability of the resulting consistency supervision. We then use this signal to formulate an uncertainty-aware consistency loss. With an exponential weight $w = e^{-\lambda \tilde{U}}$ on the distillation objective, where λ controls the weighting strength, the consistency penalty is relaxed on regions that are hard to learn, while regions with near-zero uncertainty keep the full supervision signal. Requiring the student to match targets that are themselves hard to learn is unnecessary, and enforcing it uniformly only injects unstable gradients into training. To reduce perceptual degradation under aggressive step reduction, we also employ feature-space adversarial refinement with semantic

alignment to preserve high-frequency details and semantic fidelity. The entire framework is trained with parameter-efficient LoRA adapters (Hu et al., 2021), making it directly applicable to billion-scale video diffusion transformers. We apply uncertainty-aware consistency distillation to the Wan video generation model and reduce inference to four steps while maintaining high visual fidelity. On the VBench 2.0 benchmark, our model achieves superior performance compared to existing few-step distillation methods. User studies further indicate a strong preference for our generated videos. Our main contributions are summarized as follows:

- We identify that the reliability of consistency supervision follows the difficulty of the content, *i.e.*, the magnitude of its temporal variation, rather than semantic complexity, motivating fine-grained, region-wise uncertainty modeling for few-step video distillation.
- We introduce a parameter-free local consistency uncertainty proxy based on direct and consensus student predictions, and use this proxy to adapt the supervision strength at each spatiotemporal region according to its local reliability, complemented by adversarial refinement for perceptual quality.
- We achieve state-of-the-art 4-step video generation with the Wan backbone on comprehensive VBench 2.0 evaluations (0.556 Mean Score) and are preferred over prior consistency distillation methods in a user study, while maintaining training efficiency through LoRA adaptation.

2 Related Work

Diffusion Model Distillation. Diffusion distillation (Mansourian et al., 2025; Luhman & Luhman, 2021; Zheng et al., 2023) reduces the inference cost of diffusion models (Blattmann et al., 2023; Ho et al., 2020; 2022) through trajectory-preserving (Luo et al., 2023; Ding et al., 2025; Gu et al., 2023; Wang et al., 2024; Salimans & Ho, 2022; Frans et al., 2025) or distribution-matching (Sauer et al., 2024; Wang et al., 2023; Lu et al., 2025; Chen et al., 2026; Yin et al., 2024) objectives. Video-specific methods include DCM (Lv et al., 2025), DOLLAR (Ding et al., 2025), and adversarial self-distillation (Yang et al., 2026). Our work complements these approaches with uncertainty-aware consistency distillation that adaptively reweights supervision according to local prediction reliability.

Uncertainty Learning. Uncertainty quantification is widely used to improve model reliability and robustness (Kendall & Gal, 2017; Raghu et al., 2019; Nandy et al., 2020; Lee & AlRegib, 2020), with applications in 3D reconstruction (Sabour et al., 2023), image classification (Litrico et al., 2023), and image retrieval (Zhang et al., 2022; Dou et al., 2022; Chen et al., 2024; Wang et al., 2026). Recent works also exploit uncertainty for reward modeling (Zheng & Yang, 2021; Zhang et al., 2021; 2025; 2024) and active view selection (Her et al., 2023; Jiang et al., 2024). To our knowledge, uncertainty has not been directly incorporated into the consistency distillation objective for video generation. We address this gap by deriving a local uncertainty proxy from the discrepancy between the student's direct prediction and a dual-path consensus target.

3 Methodology

3.1 Preliminary

Video Diffusion Model. Flow matching models learn a vector field $v_\psi(x_t, t; c)$ that transports noise $x_1 \sim \mathcal{N}(0, I)$ to data x_0 via the ODE $\mathrm{d}x_t = v_\psi(x_t, t; c)\,\mathrm{d}t$. The model is trained to minimize $\mathbb{E}_{t,x_0,\epsilon}\left[\|v_\psi(x_t, t; c) - (\epsilon - x_0)\|^2\right]$ with $x_t = (1 - \sigma_t)x_0 + \sigma_t\epsilon$. High-quality video synthesis requires solving this ODE with tens to hundreds of iterative solver steps, constituting the primary computational bottleneck for deployment.

Consistency Distillation. Consistency distillation aims to learn a student model $f_\phi(x_t, t; c)$ that maps any point x_t on the probability flow ODE trajectory directly to the clean endpoint x_0 in a single forward pass. The model satisfies the self-consistency property $f_\phi(x_t, t; c) = f_\phi(x_{t'}, t'; c)$ for all t, t' lying on the same trajectory. In practice, this property is enforced along a teacher-guided trajectory: the frozen teacher first performs one Euler step with step size Δt, moving from t to $t' = t - \Delta t$ to obtain an intermediate state $x_{t'}$. The student then predicts the endpoint from $x_{t'}$ to form the distillation target, while also predicting the endpoint directly from x_t. The consistency objective encourages these two student predictions to agree, with the target prediction detached from gradient computation.

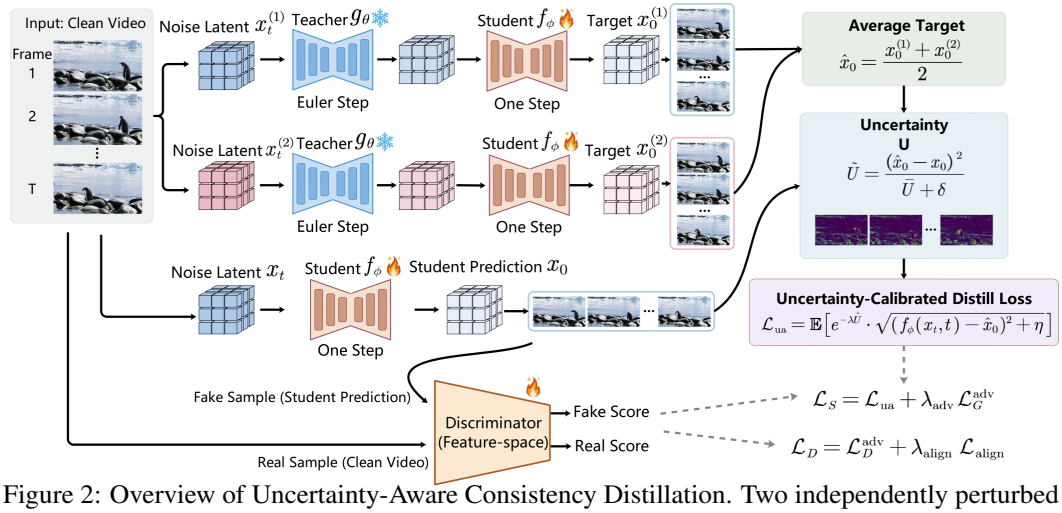


Figure 2: Overview of Uncertainty-Aware Consistency Distillation. Two independently perturbed paths are each advanced by one Euler step using the frozen teacher, producing two teacher-guided intermediate states. The student then predicts the endpoint $x_0^{(1)}$ and $x_0^{(2)}$ from each intermediate state, and their average forms a detached consensus consistency target $\hat{x}_0$. One of the two perturbed states is randomly selected for the direct student prediction input x_t, and its discrepancy with the consensus target defines the normalized local uncertainty map $\tilde{U}$, which reweights the consistency distillation loss to attenuate supervision at regions with higher uncertainty. A feature-space discriminator provides adversarial and feature-matching objectives, together with a semantic alignment loss.

3.2 Uncertainty-Aware Consistency Distillation

Dual-Path Target Construction. To estimate local consistency uncertainty without introducing an additional uncertainty model, we construct two independently perturbed teacher-guided consistency paths. Given a clean latent x_0, we randomly sample a timestep t and two independent noise perturbations $\epsilon_1, \epsilon_2 \sim \mathcal{N}(0, I)$ to construct two noisy states:

$$x_t^{(1)} = \sigma_t \epsilon_1 + (1 - \sigma_t) x_0, \quad x_t^{(2)} = \sigma_t \epsilon_2 + (1 - \sigma_t) x_0, \tag{1}$$

where σ_t denotes the noise level associated with timestep t (Lipman et al., 2023). For each perturbed input, the teacher performs a one-step Euler update $g_\theta(\cdot)$ from timestep t to $t' = t - \Delta t$ to obtain an intermediate point, which is then fed to the student $f_\phi(\cdot)$ to predict the endpoint. For notational simplicity, we omit the conditioning variable c throughout this section. Averaging these two predictions yields a consensus consistency target:

$$\hat{x}_0 = \frac{1}{2}\left(f_\phi(g_\theta(x_t^{(1)}, t), t') + f_\phi(g_\theta(x_t^{(2)}, t), t')\right), \tag{2}$$

where the two student predictions are computed without gradient tracking. Independent noise sampling provides two stochastic perturbations of the same underlying latent, allowing us to probe the stability of the resulting teacher-guided consistency predictions. The consensus target aggregates the two independently constructed predictions, reducing its dependence on any individual perturbed path. In parallel, we randomly select one of the two perturbed states, $x_t^{(1)}$ or $x_t^{(2)}$, as the input to the direct student prediction, denoted as x_t. The resulting prediction $f_\phi(x_t, t)$ remains fully trainable and is optimized to match the detached consensus target. This shared input ensures that the direct prediction and one of the teacher-guided predictions originate from the same perturbed state, while the other path provides an independent prediction for consensus estimation. Because the consensus target is formed from the teacher-advanced states while the direct prediction starts from the noisy state, the discrepancy also reflects the sensitivity of the student's endpoint to the teacher's single Euler step, in addition to the injected perturbation.

Uncertainty-Aware Loss. Following prior discrepancy-based uncertainty estimation methods (Zheng & Yang, 2021; Zhang et al., 2025), we use the student–consensus discrepancy as a parameter-free proxy for local uncertainty. Specifically, we compute the element-wise squared discrepancy between the student's direct prediction $f_\phi(x_t, t)$ and the consensus target $\hat{x}_0$:

$$U = \text{sg}\left[(f_\phi(x_t, t) - \hat{x}_0)^2\right], \tag{3}$$

where $\mathrm{sg}[\cdot]$ denotes stop-gradient. The resulting uncertainty has the same spatiotemporal latent structure as the student prediction, providing fine-grained uncertainty estimates over spatiotemporal locations. The uncertainty is treated as fixed when optimizing the student within each update. Large values of U indicate regions where the direct student prediction disagrees with the consensus prediction induced by the two teacher-guided paths, suggesting that the corresponding consistency supervision is less reliable. We normalize U by its mean value to obtain $\tilde{U} = U/(\bar{U} + \delta)$ before applying the exponential weight, where $\bar{U}$ is the per-video mean over all spatiotemporal locations, and δ is a small constant for numerical stability. The uncertainty-aware consistency loss is then defined as:

$$\mathcal{L}_{\text{ua}} = \mathbb{E}\left[e^{-\lambda\tilde{U}} \cdot \sqrt{(f_\phi(x_t, t) - \hat{x}_0)^2 + \eta}\right], \tag{4}$$

where λ controls the uncertainty weighting strength, and η is a small constant for numerical stability. Since $\tilde{U} \geq 0$ and $\lambda > 0$, the exponential weight $e^{-\lambda\tilde{U}} \in (0, 1]$ ensures that low-uncertainty spatiotemporal locations retain relatively stronger gradient contributions, while high-uncertainty locations are suppressed.

Why Not a Learnable Uncertainty Model? A natural alternative is to introduce a learnable uncertainty head to predict local uncertainty from student features. However, without explicit uncertainty supervision, such a head must learn the notion of uncertainty indirectly through the distillation objective, making its estimates sensitive to optimization dynamics and potentially poorly aligned with actual consistency errors. We instead derive uncertainty from the consistency discrepancy, directly linking uncertainty to the reliability of the distillation signal. Our experiments in Sec. 4.2 and App. A show that two perturbed paths are sufficient for uncertainty estimation.

3.3 Adversarial Refinement with Semantic Alignment

Training with the consistency loss alone tends to produce blurry outputs due to regression toward the mean. We complement it with two auxiliary objectives to preserve perceptual quality during aggressive step reduction.

Feature-Space Adversarial Loss. Following prior distillation works (Lv et al., 2025; Lu et al., 2025), we employ a discriminator $\mathcal{D}$ on intermediate transformer features, which provide richer structural and semantic information than pixel space. Specifically, h_{fake} and h_{real} are extracted by the teacher transformer from student-generated samples and corresponding real samples, respectively, and then fed to $\mathcal{D}$. The generator and discriminator losses are:

$$\mathcal{L}_G^{\text{adv}} = \mathbb{E}\left[\max\left(0, 1 - \mathcal{D}\left(h_{fake}\right)\right) + \lambda_{\text{feat}} \parallel h_{real} - h_{fake} \parallel_2^2\right], \tag{5}$$

$$\mathcal{L}_D^{\text{adv}} = \mathbb{E}\left[\max\left(0, 1 - \mathcal{D}\left(h_{real}\right)\right) + \max\left(0, 1 + \mathcal{D}\left(h_{fake}\right)\right)\right], \tag{6}$$

where λ_{feat} balances the feature matching term.

Semantic Alignment with Frozen Vision Embeddings. To provide semantic guidance for adversarial training, we align intermediate discriminator representations with frozen DINOv2 (Oquab et al., 2023) representations extracted from real video frames. Specifically, the selected teacher features are processed by the corresponding discriminator heads and projected by a learnable head $\mathrm{Proj}(\cdot)$ to match the dimension of $f_{\text{vis}}(\cdot)$. The alignment loss is:

$$\mathcal{L}_{\text{align}} = -\mathbb{E}\left[\cos(\mathrm{Proj}(\mathcal{D}(h_{real})), f_{\text{vis}}(y))\right], \tag{7}$$

where y denotes the corresponding ground-truth video frames temporally subsampled to match the latent temporal resolution. The alignment provides semantic guidance to the adversarial signal and helps reduce content drift during few-step generation.

3.4 Training Procedure

As shown in Figure 2, we train f_ϕ with LoRA adapters (Hu et al., 2021) on frozen base weights, while jointly optimizing $\mathcal{D}$ and Proj. The total loss for the student model f_ϕ combines the uncertainty-aware distillation loss and the adversarial generator loss, while the discriminator $\mathcal{D}$ is trained with the adversarial loss augmented by the semantic alignment loss:

$$\mathcal{L}_S = \mathcal{L}_{\text{ua}} + \lambda_{\text{adv}}\mathcal{L}_G^{\text{adv}}, \quad \mathcal{L}_D = \mathcal{L}_D^{\text{adv}} + \lambda_{\text{align}}\,\mathcal{L}_{\text{align}}, \tag{8}$$

where λ_{adv} is the adversarial loss weight, and λ_{align} controls the strength of semantic regularization.

4 EXPERIMENT

Setting. We use Wan2.1-T2V-1.3B (Wan et al., 2025) as the teacher model for distillation. The student reuses the teacher weights and is equipped with LoRA adapters (rank 128). We perform distillation on 81-frame video sequences at a resolution of 832×480, using a batch size of 4. The student and discriminator are optimized via AdamW (Loshchilov & Hutter, 2019) with learning rates of 4×10^{-5} and 1×10^{-5}, respectively. Semantic alignment uses a frozen DINOv2 ViT-B/14 encoder. The uncertainty weighting strength is set to $\lambda = 1$, and the semantic alignment loss weight is set to $\lambda_{\text{align}} = 1$. We provide ablation studies on these two hyperparameters in App. B and App. C, respectively. We follow the experimental protocol of DCM (Lv et al., 2025) for our distillation framework, setting the adversarial loss weight $\lambda_{\text{adv}} = 0.5$ and the feature-matching loss weight $\lambda_{\text{feat}} = 1$. All experiments are conducted on a single NVIDIA H800 80GB GPU. We distill for approximately 2,000 steps, which takes about 20 hours to complete.

Evaluation Metrics. For video quality evaluation, we adopt VBench 2.0 (Zheng et al., 2025) as our primary metric, which provides a fixed set of prompts and standardized evaluation protocols for assessing video generation quality. It systematically evaluates adherence to real-world physical laws and semantics across 18 fine-grained sub-abilities, aggregated into five key dimensions: Human Fidelity, Controllability, Creativity, Physics, and Commonsense. We follow the benchmark's fixed prompt set and evaluate generated videos under multiple random seeds to ensure robust and reproducible comparisons. To complement automatic metrics, we further conduct a user study on subjective visual quality and semantic coherence.

4.1 COMPARISON WITH COMPETITIVE METHODS

Competitive Methods. We compare against five representative methods spanning both multi-step video generation models and recent distillation approaches. *Wan2.1-1.3B* (Wan et al., 2025) and *CogVideoX-5B* (Yang et al., 2025) are included as multi-step (50-step) baselines that provide reference points for generation quality under substantially more inference steps. For distillation-based methods, we include three state-of-the-art approaches: *DCM* (Lv et al., 2025) employs a staged distillation strategy based on the *Wan2.1-1.3B* teacher and enables complete video generation within 4 inference steps; *CausalForcing* (Zhu et al., 2026) and *OneForcing* (Feng et al., 2026) use *Wan2.1-14B* as the teacher and adopt autoregressive frame-wise distillation. CausalForcing achieves 4-step generation for all frames, while OneForcing applies 4 steps only to the first frame and reduces subsequent frames to single-step inference, trading off initial fidelity for faster sequential generation.

Quantitative Comparison. As shown in Table 1, our method achieves the highest VBench 2.0 mean score of **0.556** with only 4 inference steps using a lightweight 1.3B teacher, surpassing both 50-step baselines and all distillation baselines. Compared to Wan2.1-1.3B and CogVideoX-5B, which require 50 steps, our approach delivers substantial improvements in all metrics while reducing computational cost by over an order of magnitude. Among distillation methods, our method leads in Human Fidelity (0.861), Creativity (0.642), and Physics (0.516). The high Creativity score suggests that the proposed distillation framework can preserve diverse generation under aggressive step reduction. OneForcing and DCM are marginally better in Controllability and Commonsense but worse elsewhere. These results indicate that our consistency distillation framework with dual-path target prediction and discrepancy-based uncertainty estimation enables a 1.3B student to match or exceed the quality of larger models and competitive distillation baselines under highly efficient inference conditions.

Qualitative Comparison. As shown in Figure 3, we compare representative outputs for the prompt "*A horse is grazing in the field, then it suddenly starts running across the meadow*". Wan2.1-1.3B generates two distinct horses for the grazing and running actions, breaking entity continuity even though both motions are present. CogVideoX-5B, CausalForcing, and OneForcing capture the initial grazing scene but omit the subsequent running motion, neglecting the specified temporal transition. CausalForcing further exhibits progressive color drift, which may be associated with error accumulation in autoregressive generation. DCM exhibits noticeable color distortion and less natural textures in our comparison. Notably, our method adopts Wan2.1-1.3B itself as the teacher model, yet generates videos of higher visual quality and stronger prompt alignment than the teacher. Our uncertainty-aware distillation relaxes the supervision on unreliable consistency constraints relative to reliable ones, enabling the student to further improve visual fidelity and temporal-semantic coherence beyond the teacher's output.

Table 1: Quantitative comparison on VBench 2.0 across 5 video quality dimensions. All videos are generated at 832×480 resolution with $\ell = 81$ frames. NFE denotes the number of function evaluations during inference. Our method achieves the highest mean score with 4 NFEs using a 1.3B teacher. DCM also uses a 1.3B teacher, while CausalForcing and OneForcing use a 14B teacher.

Method	NFE	Teacher	Human Fidelity	Creativity	Controllability	Commonsense	Physics	Mean
Wan2.1-1.3B	50	-	0.711	0.551	0.099	0.519	0.382	0.452
CogVideoX-5B	50	-	0.797	0.344	0.199	0.539	0.405	0.457
Distillation Methods								
CausalForcing	$4 + 4 \times (\ell//4)$	Wan2.1-14B	0.855	0.462	0.184	0.565	0.492	0.512
OneForcing	$4 + 1 \times (\ell//4)$	Wan2.1-14B	0.754	0.631	**0.213**	0.571	0.460	0.526
DCM	4	Wan2.1-1.3B	0.828	0.599	0.209	**0.603**	0.394	0.527
UACD(Ours)	4	Wan2.1-1.3B	**0.861**	**0.642**	0.203	0.556	**0.516**	**0.556**



Input: ***A horse is grazing in the field, then it suddenly starts running across the meadow.***

Figure 3: Qualitative comparison with competitive methods. Our method generates videos with superior visual quality, accurate temporal transitions, and full prompt adherence compared to all baselines, highlighting that our uncertainty modeling enables the student to surpass its own teacher. More sample videos are available on our anonymous project website.

Efficiency. As shown in Figure 5, our method achieves a favorable balance between generation quality and inference speed among all evaluated approaches. Notably, with only four inference steps, our student model surpasses the generation quality of its teacher (Wan2.1-1.3B) while remaining highly efficient for long video synthesis. Although OneForcing enables one-step generation at the frame level, its autoregressive design still requires sequential generation across 81 frames, whereas our video-level approach requires only four sequential steps for the entire video, avoiding the error accumulation of frame-wise autoregression.

User Study. Following the human evaluation protocol of prior works (Lv et al., 2025; Kong et al., 2024; Zheng et al., 2025), we randomly sample 30 videos per model. Each trial presents a text prompt and two videos, one from our method and one from a competing distillation approach, with randomized order. Raters select the preferred video based on text alignment, motion quality, and visual quality. Each sample is evaluated by 30 independent raters, with aggregated results shown in Figure 4. The voting results indicate that our method is more preferred.

4.2 Ablation Studies and Further Discussion

To isolate the contribution of each component, we conduct controlled experiments with six configurations: *Variant 1* retains only plain consistency distillation together with basic adversarial training, without semantic alignment or uncertainty modeling; *Variant 2* builds upon Variant 1 by incorporating semantic alignment into the adversarial objective; *Variant 3* uses a single teacher-guided path to

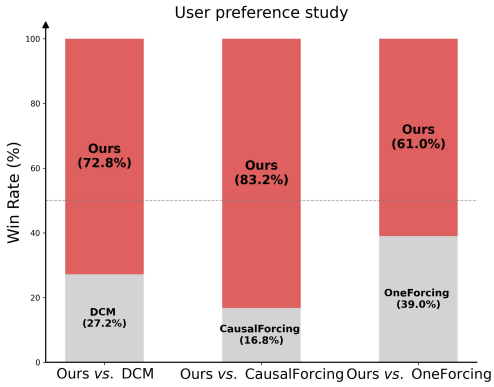


Figure 4: User preference study results. Our approach is favored by a clear majority of users across all pairwise comparisons, reflecting its superior perceptual quality.

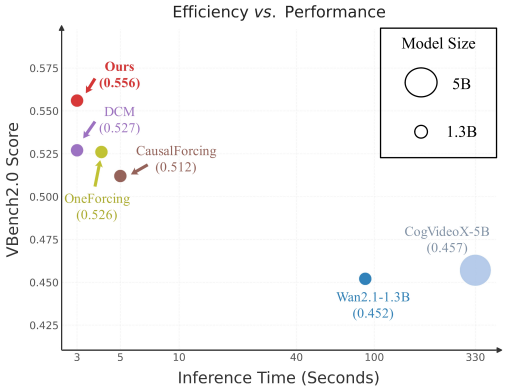


Figure 5: Efficiency comparison across methods. Methods closer to the top-left achieve superior performance, indicating higher generation quality with reduced inference time.

Table 2: Ablation study on primary components. **CD**: Consistency Distillation baseline with adversarial training. **DP**: Dual-path Prediction, where two independently perturbed teacher-guided paths are constructed and the student predicts the endpoint from each path. **UT**: Uncertainty Type. **ED**: Uncertainty Estimation via Discrepancy, where the discrepancy between the student's direct endpoint prediction and the consensus consistency target is used as a parameter-free uncertainty estimate. **LT**: Uncertainty Estimation via a Learnable Token. **DSA**: Discriminator Semantic Alignment loss applied during adversarial training. Variant 1 achieves near-perfect Human Fidelity but a substantially lower mean, since Human Fidelity rewards frame-level realism but not semantic or physical correctness. Our full model (Variant 6) achieves the highest mean with balanced performance, confirming that dual-path uncertainty and semantic alignment are complementary.

	Variants				VBench 2.0					
	CD	DP	UT	DSA	Human Fidelity	Creativity	Controllability	Commonsense	Physics	Mean
(1)	✓				**0.998**	0.396	0.146	0.509	0.169	0.444
(2)	✓			✓	0.597	0.632	0.108	0.551	0.422	0.462
(3)	✓		ED	✓	0.840	0.584	0.197	0.487	0.456	0.513
(4)	✓	✓	LT	✓	0.812	0.605	0.153	0.516	0.458	0.509
(5)	✓	✓	ED		0.854	**0.682**	0.142	0.551	0.422	0.530
(6)	✓	✓	ED	✓	0.861	0.642	**0.203**	**0.556**	**0.516**	**0.556**

construct one teacher-guided supervision target, and estimates uncertainty from the discrepancy between this target and the student's direct prediction; *Variant 4* keeps the dual-path construction but replaces the discrepancy-based uncertainty proxy with a freely learnable uncertainty token optimized end-to-end; *Variant 5* keeps the complete pipeline but removes the semantic alignment from the adversarial objective; and *Variant 6* is our full model.

Quantitative Effect of Primary Components. As shown in Table 2, Variant 1 achieves near-perfect Human Fidelity (0.998) but lower scores on Physics (0.169) and Controllability (0.146). Without semantic alignment or uncertainty reweighting, the student produces clean but semantically and physically wrong frames, which Human Fidelity does not penalize since it rewards frame-level realism rather than semantic or physical correctness. The configuration without uncertainty modeling (Variant 2) achieves a mean VBench 2.0 score of only 0.462, whereas introducing the single-path uncertainty proxy (Variant 3) raises it to 0.513, indicating that uncertainty-aware supervision helps to stabilize consistency distillation. Replacing the discrepancy-based proxy with a freely learnable uncertainty token (Variant 4) results in a lower mean score of 0.509, suggesting that directly grounding uncertainty in prediction discrepancies provides a more effective signal than learning uncertainty weights solely from the distillation objective. These results suggest that explicit uncertainty estimation is beneficial for consistency distillation. Moreover, combining the proposed dual-path construction with discrepancy-based uncertainty estimation yields a more effective uncertainty-aware supervision signal than either the single-path proxy or the learnable uncertainty token. We further investigate the number of target paths in the supplementary material. Extending the dual-path construction to three and four paths yields comparable performance at higher computational cost, validating our dual-path design as an efficient choice.

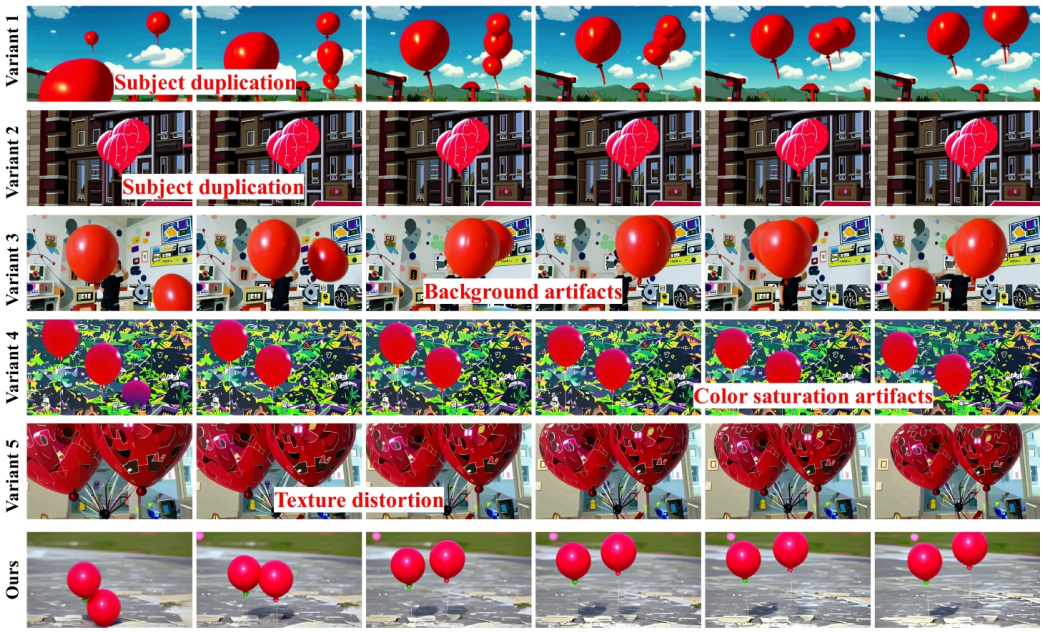


Figure 6: Qualitative comparison of ablation variants. Our full model generates videos that faithfully align with the text prompt and closely resemble real-world video content in motion dynamics and visual fidelity.

Qualitative Effect of Primary Components. As shown in Figure 6, we visualize ablation results under the prompt "*Two red balloons are floating in the air.*" to examine how each component affects generation quality. Variants 1 and 2 produce an incorrect number of balloons, and Variant 1 further yields color tones that deviate markedly from real-world videos, indicating that plain consistency distillation without semantic alignment struggles with both object counting and photorealism. Introducing uncertainty estimation via single-path target (Variant 3) or a learnable token (Variant 4) leads to cluttered background scenes, with Variant 4 additionally exhibiting excessively saturated background colors, suggesting that these alternative uncertainty estimation strategies are less effective at preserving clean spatial structure. Removing semantic alignment from the adversarial objective (Variant 5) yields distorted balloon textures, confirming that semantic-aligned adversarial training is essential for preserving fine-grained structural fidelity. Only our full model simultaneously achieves accurate prompt adherence, natural color reproduction, clean background composition, and realistic motion, suggesting that all proposed components are complementary and jointly beneficial for high-quality few-step video generation.

Limitations. Our method achieves high-quality video generation in 4 inference steps, but extending consistency distillation to 1-step remains difficult. As discussed in OneForcing (Feng et al., 2026), the sharply concentrated curvature of video teacher trajectories near the high-noise endpoint degrades trajectory-based consistency objectives when the trajectory is compressed into a single step.

5 Conclusion

We introduce Uncertainty-Aware Consistency Distillation for high-quality few-step video generation. Our approach recognizes that consistency supervision is not equally reliable across spatiotemporal locations, with its reliability varying according to how much the content changes over time. We derive a parameter-free local uncertainty signal from independently perturbed teacher-guided paths to adapt the strength of distillation supervision. Combined with semantic-aligned adversarial training, our method enables effective distillation into four inference steps, surpassing both the teacher and existing distillation methods on VBench 2.0 and human evaluations. These results show the effectiveness of uncertainty-aware supervision for improving the fidelity and semantic coherence of few-step video generation.

6 AI USE STATEMENT

In this work, we used generative AI tools (*e.g.*, ChatGPT) to edit this paper to improve readability, specifically for grammar and spelling checks. We have not used generative AI tools for creating scientific figures or for any content that would constitute plagiarism. We have carefully reviewed all AI-assisted text revisions to ensure their accuracy. We take full responsibility for the final content of this work.

REFERENCES

Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22563–22575. IEEE, 2023.

Guanjie Chen, Shirui Huang, Yifu Sun, Kai Liu, Jianchen Zhu, Xiaoye Qu, Yu Cheng, and Peng Chen. Flash-dmd: Towards high-fidelity few-step image generation with efficient distillation and joint reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6010–6020, June 2026.

Yiyang Chen, Zhedong Zheng, Wei Ji, Leigang Qu, and Tat-Seng Chua. Composed image retrieval with text feedback via multi-grained uncertainty regularization. In *International Conference on Learning Representations*, volume 2024, pp. 54915–54927, 2024.

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.

Zhaopeng Dou, Zhongdao Wang, Weihua Chen, Yali Li, and Shengjin Wang. Reliability-aware prediction via uncertainty learning for person image retrieval. In *European Conference on Computer Vision*, pp. 588–605. Springer, 2022.

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.

Kevin Frans, Danijar Hafner, Sergey Levine, and Pieter Abbeel. One step diffusion via shortcut models. In *International Conference on Learning Representations*, volume 2025, pp. 34668–34684, 2025.

Jiatao Gu, Shuangfei Zhai, Yizhe Zhang, Lingjie Liu, and Joshua M Susskind. Boot: Data-free distillation of denoising diffusion models with bootstrapping. In *ICML 2023 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2023.

Paris Her, Logan Manderle, Philipe A Dias, Henry Medeiros, and Francesca Odone. Uncertainty-aware gaze tracking for assisted living environments. *IEEE Transactions on Image Processing*, 32: 2335–2347, 2023.

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

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Wen Jiang, Boshu Lei, and Kostas Daniilidis. Fisherrf: Active view selection and mapping with radiance fields using fisher information. In *European conference on computer vision*, pp. 422–440. Springer, 2024.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in neural information processing systems*, 30, 2017.

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

Jinsol Lee and Ghassan AlRegib. Gradients as a measure of uncertainty in neural networks. In *2020 IEEE International Conference on Image Processing (ICIP)*, pp. 2416–2420. IEEE, 2020.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Mattia Litrico, Alessio Del Bue, and Pietro Morerio. Guiding pseudo-labels with uncertainty estimation for source-free unsupervised domain adaptation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7640–7650. IEEE, 2023.

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.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Yanzuo Lu, Yuxi Ren, Xin Xia, Shanchuan Lin, Xing Wang, Xuefeng Xiao, Andy J Ma, Xiaohua Xie, and Jian-Huang Lai. Adversarial distribution matching for diffusion distillation towards efficient image and video synthesis. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16818–16829. IEEE, 2025.

Eric Luhman and Troy Luhman. Knowledge distillation in iterative generative models for improved sampling speed. *arXiv preprint arXiv:2101.02388*, 2021.

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

Zhengyao Lv, Chenyang Si, Tianlin Pan, Zhaoxi Chen, Kwan-Yee K Wong, Yu Qiao, and Ziwei Liu. Dual-expert consistency model for efficient and high-quality video generation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 14983–14993. IEEE, 2025.

Amir M Mansourian, Rozhan Ahmadi, Masoud Ghafouri, Amir Mohammad Babaei, Elaheh Badali Golezani, Zeynab Yasamani Ghamchi, Vida Ramezanian, Alireza Taherian, Kimia Dinashi, Amirali Miri, et al. A comprehensive survey on knowledge distillation. *arXiv preprint arXiv:2503.12067*, 2025.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14297–14306. IEEE, 2023.

Jay Nandy, Wynne Hsu, and Mong Li Lee. Towards maximizing the representation gap between in-domain & out-of-distribution examples. *Advances in neural information processing systems*, 33: 9239–9250, 2020.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Maithra Raghu, Katy Blumer, Rory Sayres, Ziad Obermeyer, Bobby Kleinberg, Sendhil Mullainathan, and Jon Kleinberg. Direct uncertainty prediction for medical second opinions. In *International conference on machine learning*, pp. 5281–5290. PMLR, 2019.

Sara Sabour, Suhani Vora, Daniel Duckworth, Ivan Krasin, David J Fleet, and Andrea Tagliasacchi. Robustnerf: Ignoring distractors with robust losses. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20626–20636. IEEE, 2023.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=TIdIXIpzhoI.

Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pp. 87–103. Springer, 2024.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

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

Fu-Yun Wang, Zhaoyang Huang, Alexander W Bergman, Dazhong Shen, Peng Gao, Michael Lingelbach, Keqiang Sun, Weikang Bian, Guanglu Song, Yu Liu, et al. Phased consistency models. *Advances in neural information processing systems*, 37:83951–84009, 2024.

Jiacheng Wang, Zhedong Zheng, Wei Xu, and Ping Liu. Rigi: Rectifying image-to-3d generation inconsistency via uncertainty-aware learning. *IEEE Transactions on Image Processing*, 2026.

Zhendong Wang, Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. Diffusion-GAN: Training GANs with diffusion. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=HZf7UbpWHuA.

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.

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, 2025.

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, 2024.

Guiyu Zhang, Huan-ang Gao, Zijian Jiang, Hao Zhao, and Zhedong Zheng. Ctrl-u: Robust conditional image generation via uncertainty-aware reward modeling. In *International Conference on Learning Representations*, volume 2025, pp. 86480–86499, 2025.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *2018 IEEE/CVF conference on computer vision and pattern recognition*, pp. 586–595. IEEE, 2018.

Ruiyang Zhang, Hu Zhang, and Zhedong Zheng. Vl-uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation. *arXiv preprint arXiv:2411.11919*, 2024.

Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. Uncertainty-aware blind image quality assessment in the laboratory and wild. *IEEE Transactions on Image Processing*, 30:3474–3486, 2021.

Xinyu Zhang, Dongdong Li, Zhigang Wang, Jian Wang, Errui Ding, Javen Qinfeng Shi, Zhaoxiang Zhang, and Jingdong Wang. Implicit sample extension for unsupervised person re-identification. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7359–7368. IEEE, 2022.

Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Lulu Gu, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, et al. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness. *arXiv preprint arXiv:2503.21755*, 2025.

Hongkai Zheng, Weili Nie, Arash Vahdat, Kamyar Azizzadenesheli, and Anima Anandkumar. Fast sampling of diffusion models via operator learning. In *International conference on machine learning*, pp. 42390–42402. PMLR, 2023.

Zhedong Zheng and Yi Yang. Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation. *International Journal of Computer Vision*, 129(4):1106–1120, 2021.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=BYInOck3gr.

CONTENTS

A Ablation on the Number of Perturbed Paths 14
B Ablation on Uncertainty Weighting Strength 14
C Ablation on Semantic Alignment Strength 14
D Ablation on Semantic Alignment Depth 14
E Additional Visual Comparison Results 15

A ABLATION ON THE NUMBER OF PERTURBED PATHS

In our Dual-Path Target Construction, the consistency target $\hat{x}_0$ and the associated uncertainty proxy U are computed from $K = 2$ independently perturbed paths. To investigate the effect of the number of paths, we generalize the target construction to K paths, where $\hat{x}_0 = \frac{1}{K}\sum_{i=1}^{K} f_\phi(\cdot)$, and evaluate different choices of $K \in \{1, 2, 3, 4\}$. This experiment examines the trade-off between the stability of the averaged consistency target and the additional computational cost introduced by multiple paths. The corresponding VBench 2.0 scores are reported in Table 3. The results show that using two paths provides an effective balance between target stability and computational overhead, motivating our default choice of $K = 2$.

B ABLATION ON UNCERTAINTY WEIGHTING STRENGTH

To investigate the effect of the uncertainty weighting strength λ, we evaluate six values, $\lambda \in \{0, 0.1, 0.5, 1, 5, 10\}$, while keeping all other settings fixed. When $\lambda = 0$, the weighting term reduces to $e^{-\lambda\tilde{U}} = 1$, recovering the standard consistency distillation objective without uncertainty-aware reweighting. As shown in Table 4, introducing uncertainty-aware weighting substantially improves performance, with the mean VBench 2.0 score increasing from 0.462 at $\lambda = 0$ to 0.556 at $\lambda = 1$. Larger values further improve the overall score, with $\lambda = 5$ achieving the highest mean of 0.568, while $\lambda = 10$ remains competitive at 0.562. These results indicate that the proposed uncertainty-aware supervision is beneficial over a broad range of weighting strengths. We use $\lambda = 1$ as the default setting in our main experiments. This value is fixed a priori rather than selected on the benchmark: although $\lambda = 5$ attains a higher mean, choosing the weighting strength on the same benchmark that is used for comparison would inflate the reported score, so Table 4 is provided to characterize the sensitivity rather than to select the operating point. We also note that, because $\tilde{U}$ is normalized by its per-video mean, a location with average uncertainty receives weight $e^{-\lambda}$, so λ scales the overall magnitude of the distillation term relative to the adversarial term in the student objective in addition to controlling the relative reweighting.

C ABLATION ON SEMANTIC ALIGNMENT STRENGTH

The semantic alignment loss $\mathcal{L}_{\text{align}}$ anchors the discriminator's intermediate features to frozen vision embeddings, encouraging semantic consistency during few-step generation. We investigate the sensitivity of our framework to the alignment weight $\lambda_{\text{align}} \in \{0.1, 0.5, 1, 5, 10\}$. The corresponding VBench 2.0 scores are reported in Table 5. We use $\lambda_{\text{align}} = 1$ in all main experiments, which provides a favorable balance among the evaluated configurations.

D ABLATION ON SEMANTIC ALIGNMENT DEPTH

Our semantic alignment loss $\mathcal{L}_{\text{align}}$ computes the cosine similarity between frozen vision embeddings and the discriminator's projected features. The projection heads take intermediate features extracted from selected layers of the Wan teacher model. To investigate the effect of feature extraction depth, we evaluate four configurations: the last layer only, the last 5 layers, the last 10 layers, and all layers. The corresponding VBench 2.0 scores are reported in Table 6. Based on these results, we use features from all layers for semantic alignment in the main experiments.

Table 3: Ablation on the number of perturbed paths.

K	Human Fidelity	Creativity	Controllability	Commonsense	Physics	Mean
1	0.840	0.584	0.197	0.487	0.456	0.513
2	**0.861**	0.642	**0.203**	0.556	0.516	**0.556**
3	0.860	0.645	0.192	0.545	**0.527**	0.554
4	0.826	**0.674**	0.179	**0.582**	0.519	**0.556**

Table 4: Ablation on uncertainty weighting strength.

λ	Human Fidelity	Creativity	Controllability	Commonsense	Physics	Mean
0	0.597	0.632	0.108	0.551	0.422	0.462
0.1	0.747	0.637	0.160	**0.574**	0.502	0.524
0.5	0.794	0.630	0.154	0.571	0.484	0.526
1	0.861	0.642	**0.203**	0.556	0.516	0.556
5	**0.862**	0.692	0.195	**0.574**	0.518	**0.568**
10	0.824	**0.711**	0.180	0.568	**0.530**	0.562

Table 5: Ablation on semantic alignment strength.

λ_{align}	Human Fidelity	Creativity	Controllability	Commonsense	Physics	Mean
0.1	0.807	0.641	0.164	0.541	0.477	0.526
0.5	0.842	0.622	0.166	0.552	0.495	0.535
1	**0.861**	**0.642**	**0.203**	**0.556**	**0.516**	**0.556**
5	0.838	0.640	0.189	0.528	0.493	0.538
10	0.812	0.611	0.158	0.553	0.488	0.524

Table 6: Ablation on semantic alignment depth.

Depth	Human Fidelity	Creativity	Controllability	Commonsense	Physics	Mean
1	0.860	0.627	0.166	0.555	0.490	0.540
5	0.853	**0.647**	0.185	0.548	0.514	0.549
10	0.844	0.613	0.190	**0.591**	0.506	0.549
all	**0.861**	0.642	**0.203**	0.556	**0.516**	**0.556**

E Additional Visual Comparison Results

More visual comparison results are presented in Figures 7, 8, 9. We observe that our method consistently generates videos that align well with the given prompts while preserving temporal consistency, coherent motion, and stable subject appearance. These results further demonstrate the effectiveness of our uncertainty-aware supervision in improving video quality and maintaining spatiotemporal consistency under few-step generation. More video samples are available on our anonymous project website: `https://uacd.github.io/UACD/`.

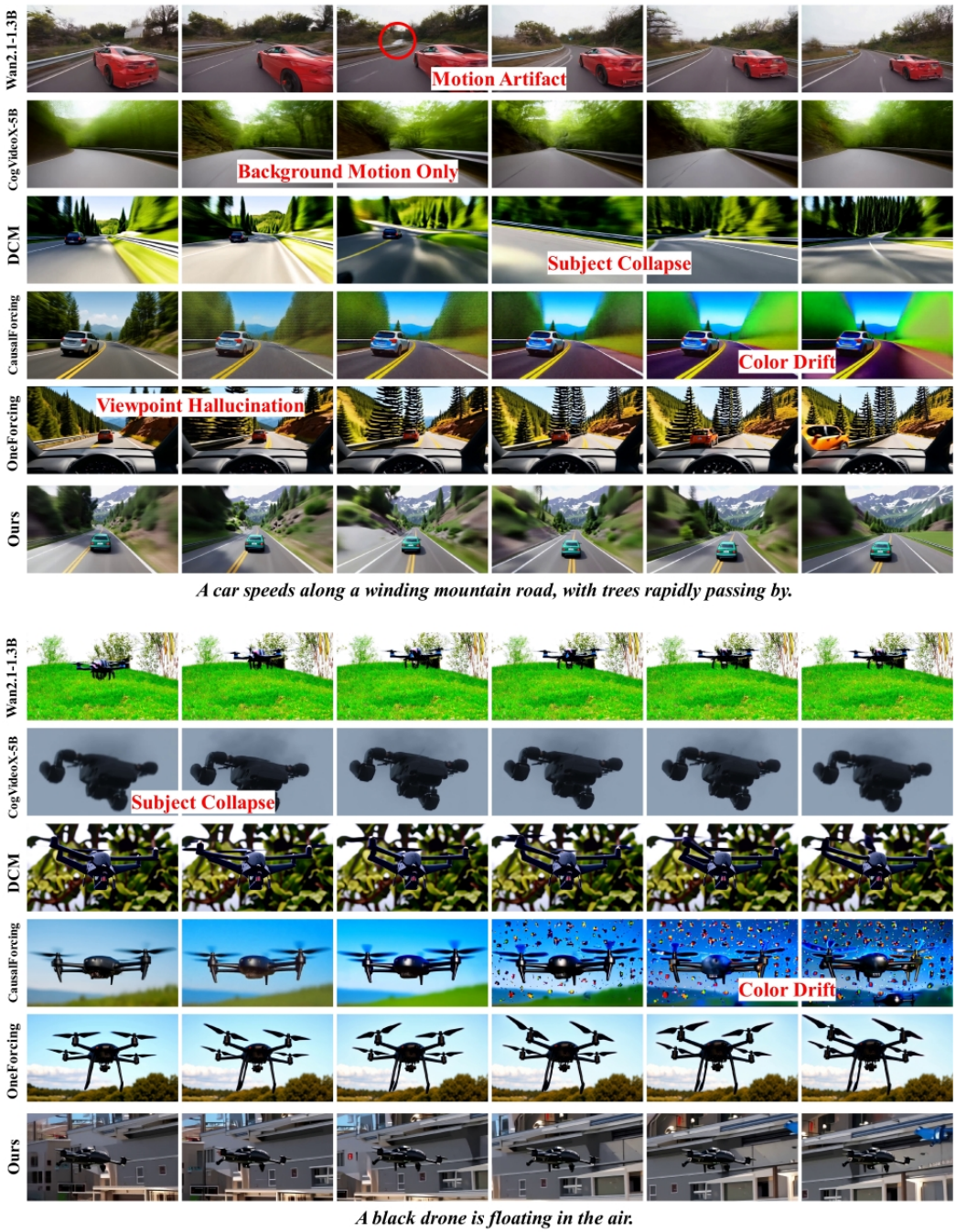


Figure 7: Additional qualitative comparison with competitive methods.

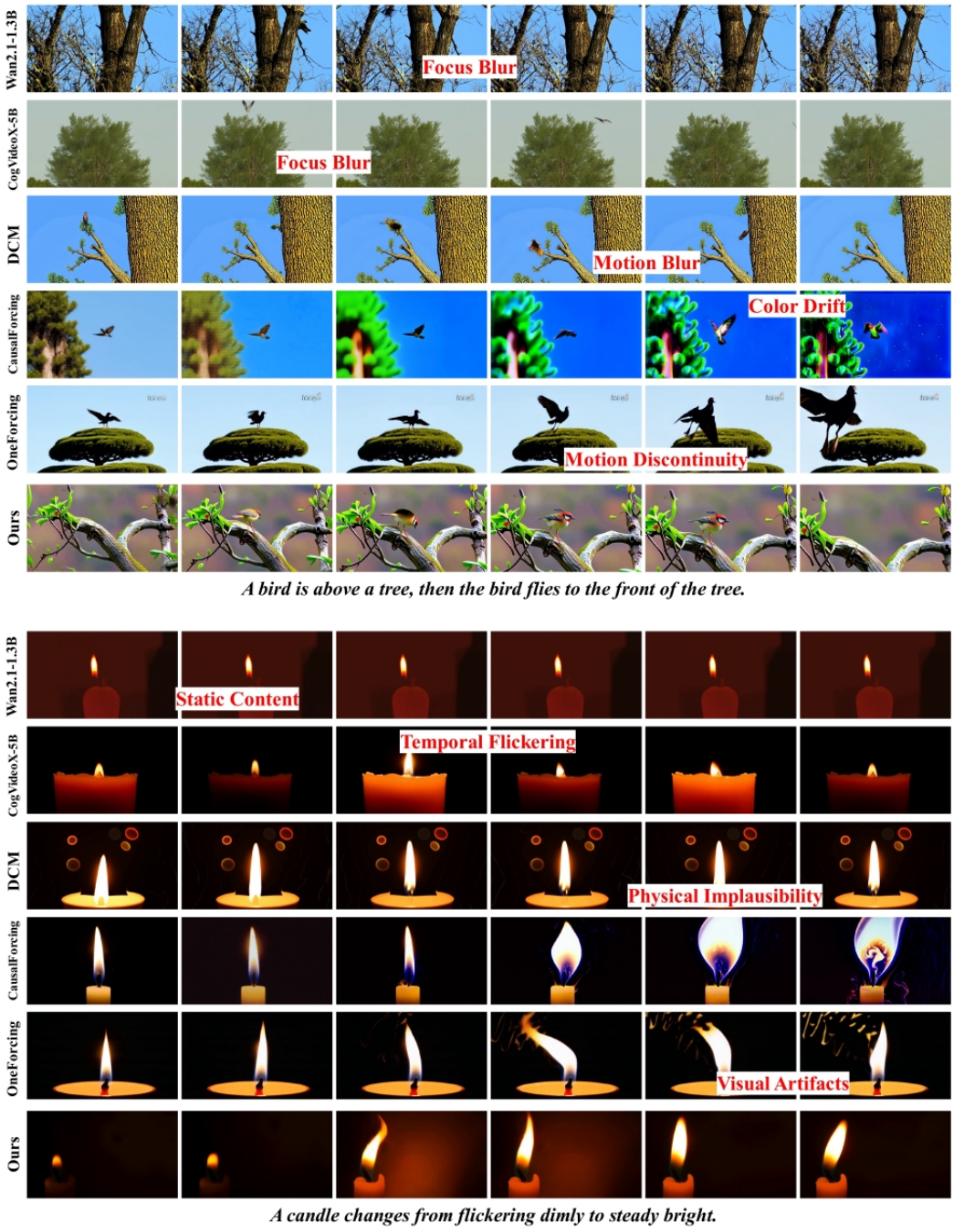


Figure 8: Additional qualitative comparison with competitive methods.

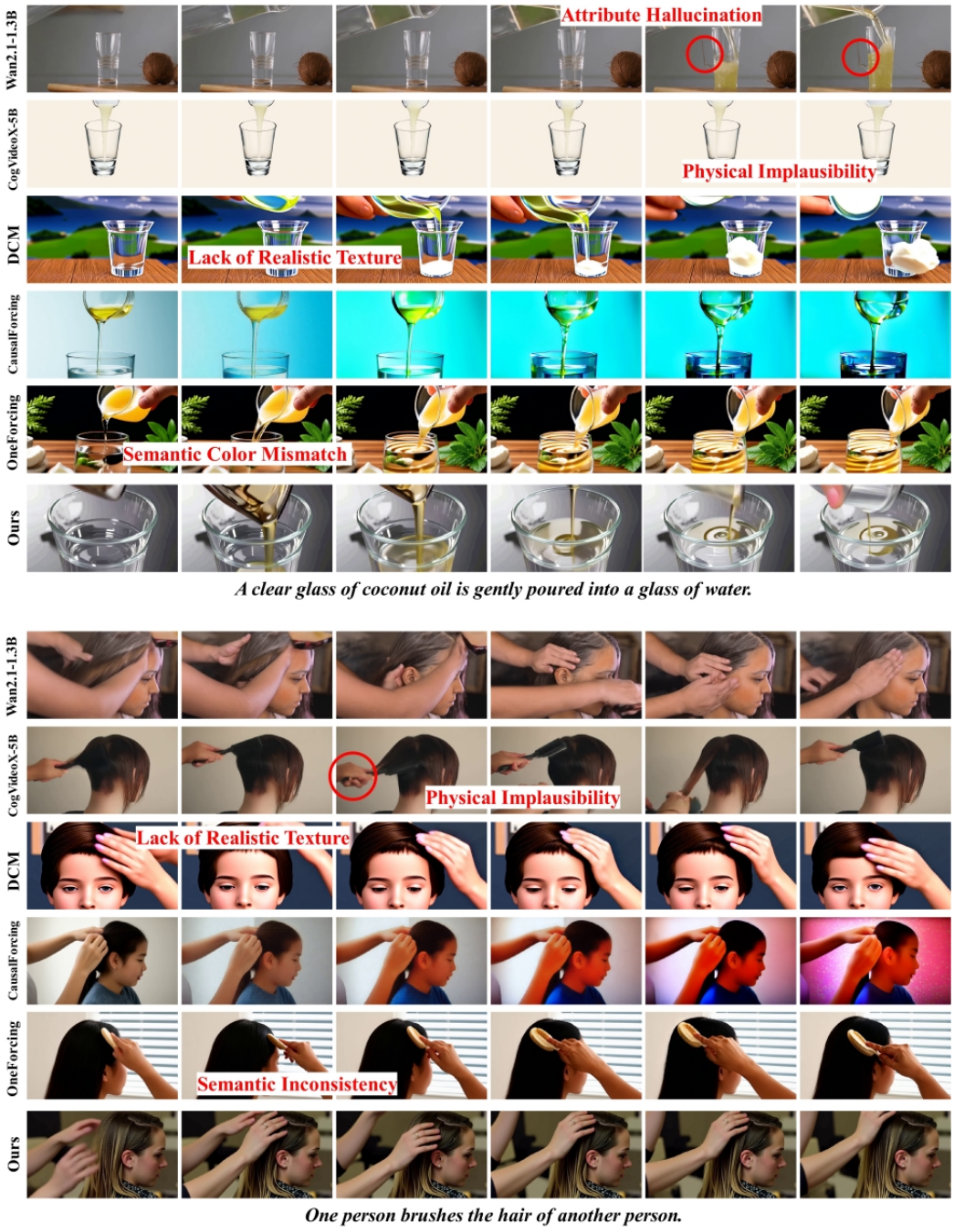


Figure 9: Additional qualitative comparison with competitive methods.