DreamWorld: Unified World Modeling in Video Generation

Boming Tan[1*], **Xiangdong Zhang**[2*], **Ning Liao**[2*], **Jingtao Zhang**[3], **Yuqing Zhang**[1], **Xiaoqiu Zhong**[2], **Xiaosong Jia**[4], **Xue Yang**[2], **Yanyong Zhang**[1], **Shaofeng Zhang**[1†]
[1]University of Science and Technology of China, [2]Shanghai Jiao Tong University,
[3]Uber, [4]Fudan University
bomingtan@foxmail.com, sfzhang@ustc.edu.cn

Abstract

Despite impressive progress in video generation, existing models remain limited to surface-level plausibility, lacking a coherent and unified understanding of the world. Prior approaches typically incorporate only a single form of world-related knowledge or rely on rigid alignment strategies to introduce additional knowledge. However, aligning the single world knowledge is insufficient to constitute a world model that requires jointly modeling multiple heterogeneous dimensions (e.g., physical commonsense, 3D and temporal consistency). To address this limitation, we introduce **DreamWorld**, a unified framework that integrates complementary world knowledge into video generators via a **Joint World Modeling Paradigm**, jointly predicting video pixels and features from foundation models to capture temporal dynamics, spatial geometry, and semantic consistency. However, naively optimizing these heterogeneous objectives can lead to visual instability and temporal flickering. To mitigate this issue, we propose *Consistent Constraint Annealing (CCA)* to progressively regulate world-level constraints during training, and *Multi-Source Inner-Guidance* to enforce learned world priors at inference. Extensive evaluations show that DreamWorld improves world consistency, outperforming Wan2.1 by 2.26 points on VBench. Code is available at **Github**.

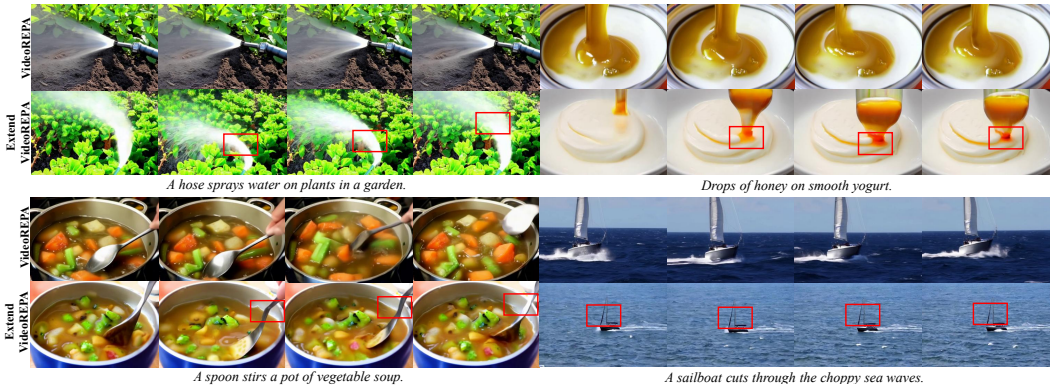


Figure 1: Limitations of extending VideoREPA to multi-source knowledge, leading to structural implausibility and unnatural distortion. The physics score(PC) on the Videophy benchmark dropped from **29.7** to **24.1**.

*Equal contribution. †: Shaofeng Zhang is the corresponding author. This work was supported by the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China and NSFC Youth Project 62606504.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

1 Introduction

The landscape of text-to-video (T2V) generation has been transformed by scalable diffusion transformers [35, 31], pushing the field beyond short visual clips toward the ambition of general-purpose world models [14, 7]. Although state-of-the-art systems such as Lumiere [3], OpenSora [56], and Wan2.1 [43] achieve impressive cinematic fidelity through large-scale data and computation, a fundamental gap remains: these models largely function as visual generators rather than world simulators. Optimized primarily for pixel-level distribution matching [22, 5], they fail to internalize structured and comprehensive world knowledge, which is reflected in their limited performance on world-centric benchmarks such as VBench 2.0 [55].

To mitigate the gap between visual realism and world understanding, prior work [46, 52] has explored injecting external world knowledge into video generation models. A representative direction is Representation Alignment (REPA) [50], which aligns generative models with pretrained experts to transfer structured priors. Building on this paradigm, VideoREPA [52] introduces Token Relation Distillation (TRD) as a soft alignment strategy, demonstrating its effectiveness in distilling knowledge from a single expert model. However, constructing a holistic world model inherently requires the simultaneous integration of *multiple heterogeneous knowledge sources*. Our empirical investigations reveal that naively extending the REPA to concurrently align with semantic, spatial, and dynamic experts leads to a multi-objective optimization dilemma. As illustrated in Figure 1, conflicting relational gradients from distinct teacher models induce optimization instability, suggesting that relational alignment is insufficient for the multi-source integration required by a unified world model (see Appendix A).

In response to these limitations, we propose **DreamWorld**, a unified framework designed to internalize comprehensive world knowledge through a **Joint World Modeling Paradigm**. Inspired by VideoJAM [9], DreamWorld extends standard video latents into a composite feature space, compelling the model to predict video pixels alongside a set of world features. Specifically, DreamWorld integrates temporal dynamics from Optical Flow, spatial geometry from VGGT [44], and semantic understanding from DINOv2 [33].

However, the direct superposition of such heterogeneous optimization objectives frequently induces optimization instability and temporal flickering. To mitigate this, we propose *Consistent Constraint Annealing (CCA)*, a decay mechanism that guarantees convergence by progressively modulating the influence of world knowledge, thereby ensuring high-fidelity visual generation while effectively assimilating world priors. Moreover, a *Multi-Source Inner-Guidance* mechanism is incorporated at inference time, which leverages the model's own predicted knowledge features to steer the generation process, ensuring trajectories that strictly adhere to real-world laws. **Our main contributions are summarized as follows:**

i) We present **DreamWorld**, the first unified video generation framework to integrate multi-source world knowledge, including 3D Semantic consistency, Motion Temporal dynamics, and 2D Spatial geometry.

ii) We propose a novel training strategy, **Consistent Constraint Annealing (CCA)**, which harmonizes knowledge injection with visual quality, ensuring coherent and artifact-free generation.

iii) Extensive evaluations demonstrate that DreamWorld significantly outperforms baselines and VideoJAM, establishing a new standard for world models.

2 Related Works

Video Diffusion Models. The paradigm of video generation has been revolutionized by the adoption of Diffusion Transformer (DiT) architectures [35, 31] Following the scaling laws originally observed in language modeling [23] and validated in video [41], recent state-of-the-art systems such as Wan2.1 [43] and Hunyuan-Video [24] have achieved photorealistic results by training on massive video-text corpora. To enhance training stability and inference efficiency beyond standard diffusion, Flow Matching [29] has emerged as a superior generative framework, as successfully implemented in LTX-Video [15] and Pyramid Flow [21]. Furthermore, the open-source community has democratized access to high-fidelity generation through open-weights models like Mochi 1 [39] and

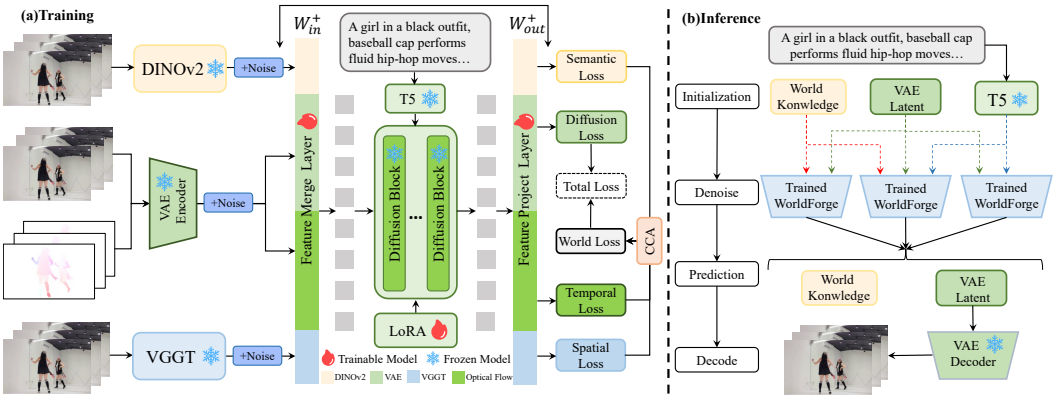


Figure 2: **Overview of DreamWorld. (a) Training:** Expert models extract multimodal features, which are noised, concatenated, and fused via a linear layer $\mathbf{W_{in}^{+}}$. A subsequent layer $\mathbf{W_{out}^{+}}$ jointly predicts appearance and world knowledge, optimized by a World Loss with *Consistency Constraint Annealing (CCA)* to ensure generation fidelity. **(b) Inference:** We introduce *Multi-Source Inner-Guidance*, leveraging inherent noise features to direct the final video generation.

CogVideoX [49]. Despite these achievements, pure diffusion-based approaches often struggle with real-world understanding, lacking the inherent capability for maintaining global consistency [37].

Representation Alignment. Representation Alignment (REPA) [51] addresses the lack of structural awareness in pixel-space diffusion by injecting high-level semantic priors from pre-trained foundation models. Building on this, recent advancements have introduced more sophisticated alignment mechanisms [54, 20, 26, 27, 53]. In the video domain, AlignVid [30], VideoREPA [52] and MoAlign [6] extend this mechanism to spatio-temporal alignment, enforcing adherence to a coherent semantic layout in generated frames. Although these methods offer a pathway to more robust video generation, they typically focus on appearance consistency rather than dynamic causal logic.

World Modeling. The concept of World Models [14] transcends video generation, aiming for a rigorous understanding of the environment's underlying laws to predict and simulate the world. Genie [8] and Genie 2 [34] instantiate this as an interactive simulator, learning a latent action space for controlling video rollouts. Alternatively, the Joint-Embedding Predictive Architecture (JEPA), exemplified by V-JEPA [4] and V-JEPA2 [1], utilizes non-generative predictive mechanisms for learning abstract world states. Frame-level [13, 38, 10, 47, 36] context mechanisms introduce frame-level context guidance by adding noise to context frames during training. Additionally, some methods [48] utilize 3D information to enhance spatial coherence. Hybrid approaches such as DriveWorld [32] and UniWorld [28] attempt the unification of generative decoding with state-space modeling. However, existing paradigms lack the capability to synergize knowledge from multiple heterogeneous expert models [46]. Bridging this gap, our work fuses complementary priors, resulting in a world model that surpasses implicit alignment methods in long-horizon consistency and realism.

3 Method

This section outlines the proposed framework. We first introduce the World Knowledge priors and the video diffusion backbone in Section 3.1, followed by the unified preprocessing and alignment protocols in Section 3.2. Section 3.3 then details the Joint World Knowledge Learning (see Figure 2), optimized via the Consistent Constraint Annealing (CCA) strategy to balance physical constraints and generative fidelity. Finally, Section 3.4 presents the Multi-Source Inner-Guidance mechanism for controllable inference.

3.1 Preliminaries

Video Diffusion Transformers. Recent advances in text-to-video generation, exemplified by models such as Wan2.1 [43], utilizes Flow Matching [29] transformers to model the transition

from noise to video as a continuous-time process. The pipeline begins with a 3D Causal VAE that compresses the video $\mathbf{x}$ into latent representations $\mathbf{z}_0 = \mathcal{E}(\mathbf{x})$. The core generative mechanism learns to straighten the probability path between the target data $\mathbf{z}_0$ and the Gaussian prior $\mathbf{z}_1$. Defining the intermediate state as:

$$\mathbf{z}_t = t\mathbf{z}_1 + (1-t)\mathbf{z}_0 \tag{1}$$

The model optimizes the velocity field v_θ by regressing against the constant flow $\mathbf{z}_1 - \mathbf{z}_0$:

$$\mathcal{L}_{\text{flow}} = \mathbb{E}_{t,\mathbf{z}_0,\mathbf{z}_1}\left[\|v_\theta(\mathbf{z}_t, t, \mathbf{c}) - (\mathbf{z}_1 - \mathbf{z}_0)\|^2\right] \tag{2}$$

where $\mathbf{c}$ represents the conditioning text embeddings. During inference, this learned velocity field guides the generation, allowing the model to recover the video latent $\hat{\mathbf{z}}_0$ by numerically solving the ordinary differential equation (ODE) from $t = 1$ to $t = 0$:

$$\hat{\mathbf{z}}_0 = \mathbf{z}_1 + \int_1^0 v_\theta(\mathbf{z}_\tau, \tau, \mathbf{c})\, d\tau \tag{3}$$

WorldKnowledge Priors. We construct a composite feature space, $\mathcal{Z}_{\text{world}}$, encapsulating essential world knowledge by unifying the fundamental *what*, *where*, and *how* of physical reality. Specifically, *DINOv2* [33] extracts robust semantics to define what objects exist, ensuring appearance consistency. *VGGT* [44] models 2D geometric constraints to establish where entities reside spatially. Finally, *Optical Flow* [17] encodes dense pixel-level trajectories to govern how elements evolve over time.

3.2 Preprocessing

We implement a unified preprocessing protocol comprising motion-to-RGB conversion, spatio-temporal alignment, and channel compression, thereby aligning heterogeneous priors with the diffusion backbone.

Motion Representation Transformation. Since standard video encoders require RGB inputs, dense displacement fields $\mathbf{d} \in \mathbb{R}^{H\times W\times 2}$—computed through RAFT [40]—are mapped into a 3-channel space following the VideoJAM protocol [9].

$$m = \min\left(1, \frac{\sqrt{u^2+v^2}}{\sigma\sqrt{H^2+W^2}}\right), \alpha = \arctan2(v, u) \tag{4}$$

where the normalized motion magnitude m and motion direction α are computed from the optical flow components (u, v) to modulate the pixel-wise intensity and hue, respectively, while σ serves as a scaling factor that controls the sensitivity of the motion magnitude normalization. The resulting visualizations are subsequently sampled and encoded by a pre-trained 3D causal VAE [43] as videos, yielding the compressed temporal latent $\mathbf{z}_{\text{temporal}}$.

Unified Alignment and Integration. Raw video foundation model (VFM) outputs are spatially interpolated and temporally pooled to align with the latent dimensions. Given the aligned tensor $\mathbf{F} \in \mathbb{R}^{C\times T\times H\times W}$, we reconcile distributional discrepancies across semantic, temporal, and spatial features via global normalization:

$$\hat{\mathbf{F}} = \frac{\mathbf{F} - \mu}{\sigma + \epsilon} \tag{5}$$

where μ and σ are the scalar mean and standard deviation across all dimensions, and $\epsilon = 10^{-6}$. Next, we apply PCA [25] to compress $\hat{\mathbf{F}}$, filtering redundant noise and distilling features into compact latents: $\mathbf{z}_{\text{temporal}}$, $\mathbf{z}_{\text{semantic}}$, and $\mathbf{z}_{\text{spatial}}$. Finally, we concatenate these priors along the channel dimension to form the consolidated world latent, efficiently conditioning the model without excessive computational overhead:

$$\mathbf{Z}_{\text{world}} = [\mathbf{z}_{\text{temporal}}, \mathbf{z}_{\text{semantic}}, \mathbf{z}_{\text{spatial}}] \tag{6}$$

3.3 Joint World Knowledge Learning

Rather than treating $\mathbf{Z}_{\text{world}}$ merely as a conditioning signal [19] or directly aligned representation [51], we concatenate $\mathbf{Z}_{\text{world}}$ with the latent and embed the joint features into the diffusion block, enabling the learning of mutual information between visual appearance and underlying world knowledge.

Joint Feature Integration. The core architectural modification in our approach focuses on enabling **joint modeling of video appearance and high-level world knowledge** within a unified transformer framework. Concretely, we expand the linear projection layers to accommodate inputs from both the denoising stream and the world knowledge stream. We extend the pre-trained input projection layer $\mathbf{W}_{\text{in}}$ to accommodate the concatenated input $\mathbf{z}_t = [\mathbf{z}_{\text{vae}}, \mathbf{Z}_{\text{world}}]_t$. The resulting projection layer is defined as:

$$\mathbf{W}_{\text{in}}^{+} = [\mathbf{W}_{\text{in}}, \mathbf{0}] \in \mathbb{R}^{(C_{\text{vae}}+C_{\text{world}})\times D} \tag{7}$$

The weights corresponding to $\mathbf{Z}_{\text{world}}$ are initialized to **zero**, ensuring that the initial model behavior exactly matches the original pre-trained Wan2.1. This design stabilizes training by preventing abrupt interference from high-dimensional world features, while allowing their influence to be gradually learned through optimization. Symmetrically, the output projection is expanded to $\mathbf{W}_{\text{out}}^{+} \in \mathbb{R}^{D\times(C_{\text{vae}}+C_{\text{world}})}$ to predict a joint velocity field, which is subsequently decomposed into modality-specific components:

$$\mathbf{z}_{\text{pred}} = \hat{\mathbf{v}}(\mathbf{z}_t, y, t) = \mathbf{M}_{(\mathbf{z}_t \bullet \mathbf{W}_{\text{in}}^{+})} \bullet \mathbf{W}_{\text{out}}^{+} \xrightarrow{\text{split}} [\hat{\mathbf{z}}_{\text{vae}}, \hat{\mathbf{z}}_{\text{temporal}}, \hat{\mathbf{z}}_{\text{semantic}}, \hat{\mathbf{z}}_{\text{spatial}}] \tag{8}$$

where $\mathbf{M}$ denotes the attention blocks and y represents the conditioning prompt.

Training Objective with Consistent Constraint Annealing. The model is optimized using a Flow Matching objective defined over the joint velocity field. The total loss is formulated as a weighted sum of modality-specific flow matching terms:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{vae}} + \lambda_{\text{temp}}(t)\mathcal{L}_{\text{temporal}} + \lambda_{\text{sem}}(t)\mathcal{L}_{\text{semantic}} + \lambda_{\text{spa}}(t)\mathcal{L}_{\text{spatial}} \tag{9}$$

Specifically, each individual loss term $\mathcal{L}_k$, where $k \in \{vae, temporal, semantic, spatial\}$, quantifies the flow matching error by computing the Mean Squared Error (MSE) between the predicted velocity component $\hat{\mathbf{z}}_k$ and the target conditional flow $\mathbf{u}_t^{(k)}$, which denotes the target velocity field pointing to the ground truth feature $\mathbf{z}_k$.

However, a static weighting scheme often exacerbates the inherent tension between maintaining high-fidelity **generative quality** and enforcing effective **knowledge learning** from the introduced features [12]. To reconcile this conflict, we introduce *Consistent Constraint Annealing (CCA)*, Unlike static strategies, CCA gradually relaxes these weights to zero, prioritizing high-fidelity, artifact-free visual reconstruction in the final phase. Mathematically, the weight $\lambda(t)$ at current training step t is formulated as:

$$\lambda(t) = \lambda_{\text{base}} \cdot \frac{1}{2}\left[1 + \cos\left(\frac{t}{T_{\text{total}}}\pi\right)\right] \tag{10}$$

where $\lambda_{\text{base}} = 0.2$ denotes the initial constraint intensity and T_{total} is the total training duration.

3.4 Multi-Source Inner-Guidance

To effectively guide video generation with predicted world knowledge during inference, we introduce Multi-Source Inner-Guidance. This approach extends standard classifier-free guidance [42, 16] from a typical two-branch (conditional versus unconditional) framework into differential guidance across multiple conditional subspaces. During inference, the joint latent $[\mathbf{z}_{\text{vae}}, \mathbf{Z}_{\text{world}}]$ is initialized from a standard Gaussian distribution. While training uses expert-derived targets for $\mathbf{Z}_{\text{world}}$, inference relies on self-predicted $\mathbf{Z}_{\text{world}}$ maintained along the denoising trajectory. Dynamically generating $\mathbf{Z}_{\text{world}}$ ensures these world priors remain synchronized and aligned with the intermediate state $\mathbf{z}_t$ at each timestep, which intrinsically mitigates train-test distribution shift and error accumulation caused by mismatched external predictors. Formulated within the Flow Matching velocity domain, the rectified velocity field $\mathbf{z}_{\text{pred}}$ is computed by aggregating the fully conditioned prediction with the differential updates from each masked feature subspace:

$$\begin{aligned}\mathbf{z}_{\text{pred}} \leftarrow (1 + w_{\text{txt}} + w_{\text{temp}} + w_{\text{sem}} + w_{\text{spa}})\,\hat{\mathbf{v}}(\mathbf{z}_t, y, t) \\ - \underbrace{w_{\text{txt}}\hat{\mathbf{v}}(\mathbf{z}_t, \emptyset, t)}_{\text{Text Guidance}} - \underbrace{w_{\text{temp}}\hat{\mathbf{v}}(\mathbf{z}_t^{\neg\text{temp}}, y, t)}_{\text{Motion Guidance}} - \underbrace{w_{\text{sem}}\hat{\mathbf{v}}(\mathbf{z}_t^{\neg\text{sem}}, y, t)}_{\text{Semantic Guidance}} - \underbrace{w_{\text{spa}}\hat{\mathbf{v}}(\mathbf{z}_t^{\neg\text{spa}}, y, t)}_{\text{Spatial Guidance}}\end{aligned} \tag{11}$$

where $\mathbf{z}_t^{\neg k}$ denotes masking the corresponding channel in the World Knowledge tensor. Empirically, we prioritize prompt adherence with $w_{\text{txt}} = 5$, while assigning moderate weights $w_{\text{temp}} = w_{\text{sem}} = w_{\text{spa}} = 1$ to the priors. A detailed Inference Cost Analysis is provided in Appendix B.

Table 1: **Quantitative comparison on VBench.FT.** denotes the fine-tuned version, and **Reimpl.** indicates our re-implementation of the method.

Method	Temporal					Semantic					Spatial	Summary		Overall Score
	Subject Consistency	Background Consistency	Temporal Flickering	Motion Smoothness	Dynamic Degree	Object Class	Color	Human Action	Multiple Objects	Scene	Relationship	Quality Score	Semantic Score	
Wan2.1-T2V-1.3B	91.83	94.71	99.17	96.51	65.00	76.09	89.93	74.60	53.66	20.03	62.37	79.81	65.43	76.93
Wan2.1-T2V-1.3B(FT.)	93.59	95.81	**99.36**	97.21	54.08	79.90	88.57	78.98	58.20	28.55	63.31	81.26	68.47	78.71
VideoJAM(Reimpl.)	91.51	**96.01**	99.13	96.05	73.88	79.22	88.92	79.20	59.66	28.34	66.17	81.18	69.08	78.76
DreamWorld (Ours)	**93.62**	94.95	98.81	**98.07**	**79.16**	**81.32**	**92.61**	**81.20**	**65.03**	**29.71**	**70.47**	**83.49**	**70.89**	**80.97**

Table 2: **VBench-2.0 quantitative results. Control.**: Controllability; **Hum. Fid.**: Human Fidelity. Bold and underline denote the best and second-best results.

Method	VBench 2.0 Dimensions				Overall
	Creativity	Control.	Hum. Fid.	Physics	
Wan2.1-T2V-1.3B	45.92	16.81	76.09	**55.85**	50.77
Wan2.1-T2V-1.3B(FT.)	43.13	**18.41**	77.09	54.51	51.18
VideoJAM	49.32	16.33	78.68	52.42	52.33
DreamWorld (Ours)	**50.89**	16.95	**80.11**	55.07	**52.97**

Table 3: **Quantitative comparison on VideoPhy benchmark.** We report Semantic Adherence (SA) and Physical Commonsense (PC) scores. The best results are highlighted in **bold**.

Method	Solid-Solid		Solid-Fluid		Fluid-Fluid		Overall	
	SA	PC	SA	PC	SA	PC	SA	PC
Wan2.1-T2V-1.3B	51.1	22.3	45.2	19.1	45.4	23.6	47.7	21.2
Wan2.1-T2V-1.3B(FT.)	46.2	18.2	47.8	22.6	50.9	23.6	45.1	20.9
VideoJAM	51.7	29.4	38.4	23.3	47.3	20.1	45.3	25.3
DreamWorld (Ours)	**54.5**	**24.5**	**48.6**	**25.4**	**60.1**	**32.7**	**52.9**	**26.2**

4 Experimental Results

4.1 Implementation Details

Model Configuration. The proposed framework is established upon the pre-trained **Wan2.1-T2V-1.3B** [43] architecture. The model is configured to synthesize video sequences comprising 81 frames with a spatial resolution of 480×832. To encapsulate multi-modal world knowledge priors, the joint feature stack employs the pre-trained RAFT model for optical flow estimation, while semantic and spatial representations are extracted via DINOv2 [33] and VGGT [44], respectively.

Training Details. For training protocol, the WISA [45] open-source dataset is utilized. All video samples are uniformly sampled for 81 frames, and then center-cropped and resized to align with the target resolution of 480×832. The model was efficiently fine-tuned using LoRA on a subset of 32k WISA videos, with a total of 2,000 optimization steps. The experiments are conducted on a cluster of 8 NVIDIA A100 GPUs with a total batch size of 16. For further elaboration on hyperparameters and architectural specifications, please refer to Appendix B.

4.2 Quantitative Results

For a comprehensive evaluation of generated video quality, we use the industry-standard VBench, its latest iteration VBench 2.0, VideoPhy, and the unified world generation benchmark, WorldScore. Additionally, all test prompts in our experiments are strictly based on the official prompt suites from the respective benchmark repositories.

VBench. The generative capability is systematically evaluated using VBench [18], a comprehensive benchmark assessing video quality and semantic consistency. As detailed in Table 1, DreamWorld achieves an overall score of 80.97, outperforming both the fine-tuned Wan2.1-T2V-1.3B baseline and VideoJAM, which aligns directly with our expectations for the injection of world knowledge. In the **temporal** dimension, kinematic priors significantly boost Motion Smoothness to 98.07 and Dynamic Degree to 79.16, indicating a better grasp of physical motion. Within the **semantic** and **spatial** domains, robust semantic and geometric constraints directly elevate the scores for Multiple Objects, Scene, and spatial Relationship to 65.03, 29.71, and 70.47, respectively. This confirms that incorporating explicit world knowledge successfully grounds the model's understanding of complex entities and spatial interactions, refining visual details and coherence rather than introducing artifacts. For a detailed breakdown across all 16 sub-dimensions, please refer to Appendix C.

VBench 2.0. Evaluations are extended to the rigorous VBench 2.0 framework [55], which employs a refined protocol designed to mirror human perceptual preferences across complex motion and compositional tasks. As quantitative results in Table 2 demonstrate, **DreamWorld** secures the leading position with an aggregate score of **53.64**, outperforming both the large-scale Wan2.1 baselines and

Table 4: **Quantitative comparison on WorldScore**. Designed to assess the capabilities of world models, this benchmark evaluates simulation logic through static and dynamic scores. The best and second-best results are highlighted in **bold** and underlined.

Method	Static			Dynamic		Overall Score
	3D Consistency	Photometric Consistency	Subjective Quality	Motion Accuracy	Motion Smoothness	
Wan2.1-T2V-1.3B	72.31	74.41	52.21	32.36	68.18	51.03
Wan2.1-T2V-1.3B(FT.)	71.73	74.62	**53.83**	**35.49**	64.76	50.95
VideoJAM	71.53	75.39	44.03	33.38	**71.95**	49.38
DreamWorld (Ours)	**73.16**	**77.55**	52.45	34.75	69.02	**51.48**

Table 5: **Ablation studies.** We analyze our method's effectiveness through(1) different **feature integration strategies**, and (2)ablating the **Multi-Source Inner-Guidance** mechanism.

Method	Quality	Semantic	Overall
VGGT Only	81.76	71.36	79.68
VGGT + DINOv2	82.08	71.58	79.98
w/o Text Guidance	77.09	47.41	71.15
w/o VGGT Guidance	81.84	70.87	79.65
w/o DINOv2 Guidance	81.98	70.95	79.78
w/o Optical Flow	81.87	**71.06**	79.71
DreamWorld (Ours)	**83.49**	70.89	**80.97**

the recent VideoJAM. The model proves particularly adept at high-fidelity synthesis highlighting its capability to generate diverse, semantically rich content. Crucially, this aesthetic superiority does not come at the expense of dynamics; our method maintains a highly competitive score, significantly ahead of VideoJAM and Baseline, ultimately striking a superior balance between generative freedom and controllability.

VideoPhy. We further extend our evaluation to VideoPhy [2], a rigorous benchmark designed to assess physical commonsense in generated videos. Unlike traditional metrics, VideoPhy specifically scrutinizes the model's ability to adhere to physical laws through Semantic Adherence (SA) and Physical Commonsense (PC), which evaluates compliance with real-world physics. Following WISA [45], when the values of SA and PC are greater than or equal to 0.5, we set them as SA = 1 and PC =1. Values less than 0.5 are set as SA = 0 and PC = 0.

As presented in Table 3, **DreamWorld** outperforms both the standard Baseline and the motion-prior-based VideoJAM across aggregated metrics. Specifically, our model achieves a state-of-the-art SA of 52.9% and PC of 26.2%, significantly surpassing other leading method. This improvement validates that our world-aware constraints not only enhance the realism of dynamic interactions but also preserve the semantic fidelity of the generation.

WorldScore. To comprehensively assess holistic simulation capabilities, we benchmark DreamWorld using WorldScore [11], a unified benchmark tailored for the rigorous assessment of world simulators. As presented in Table 4, our framework yields the highest overall score of 51.48. Beyond the aggregate score, the fine-grained evaluation explicitly demonstrates the substantial benefits of integrating world knowledge. By grounding the generation process in semantic and geometric priors, DreamWorld achieves remarkable static performance, highlighted by its leading *3D Consistency* of 73.16 and *Photometric Consistency* of 77.55, reflecting a profound understanding of structural and appearance preservation. Furthermore, rather than compromising temporal evolution for spatial quality, the embedded kinematic constraints ensure robust *Motion Accuracy* and *Motion Smoothness*. This seamlessly balanced performance across both static and dynamic axes underscores the efficacy of our multidimensional approach in simulating a realistic world.

4.3 Qualitative Results

We compare DreamWorld against VideoJAM and Wan2.1 to evaluate world modeling capabilities. As illustrated in Figure 3, our method exhibits superior world consistency. Semantically, in the space station scenario, our model accurately follows the prompt to carefully tilt the cup, allowing the liquid to flow down naturally; in contrast, both VideoJAM and Baseline fail to initiate the critical tilting action. Spatially, the happy dog demonstrates excellent 3D occlusion, ensuring the ears and turtleneck sweater are correctly positioned without physically impossible penetration. Temporally, as seen in the Gwen Stacy reading example, our model generates smooth and natural motion consistent and the character's facial identity remains stable, avoiding the temporal deformations often observed in baseline methods. These results demonstrate that injecting multi-modal world knowledge effectively constrains the diffusion process, ensuring strict alignment with real-world logic. Please refer to Appendix G for further qualitative comparisons.

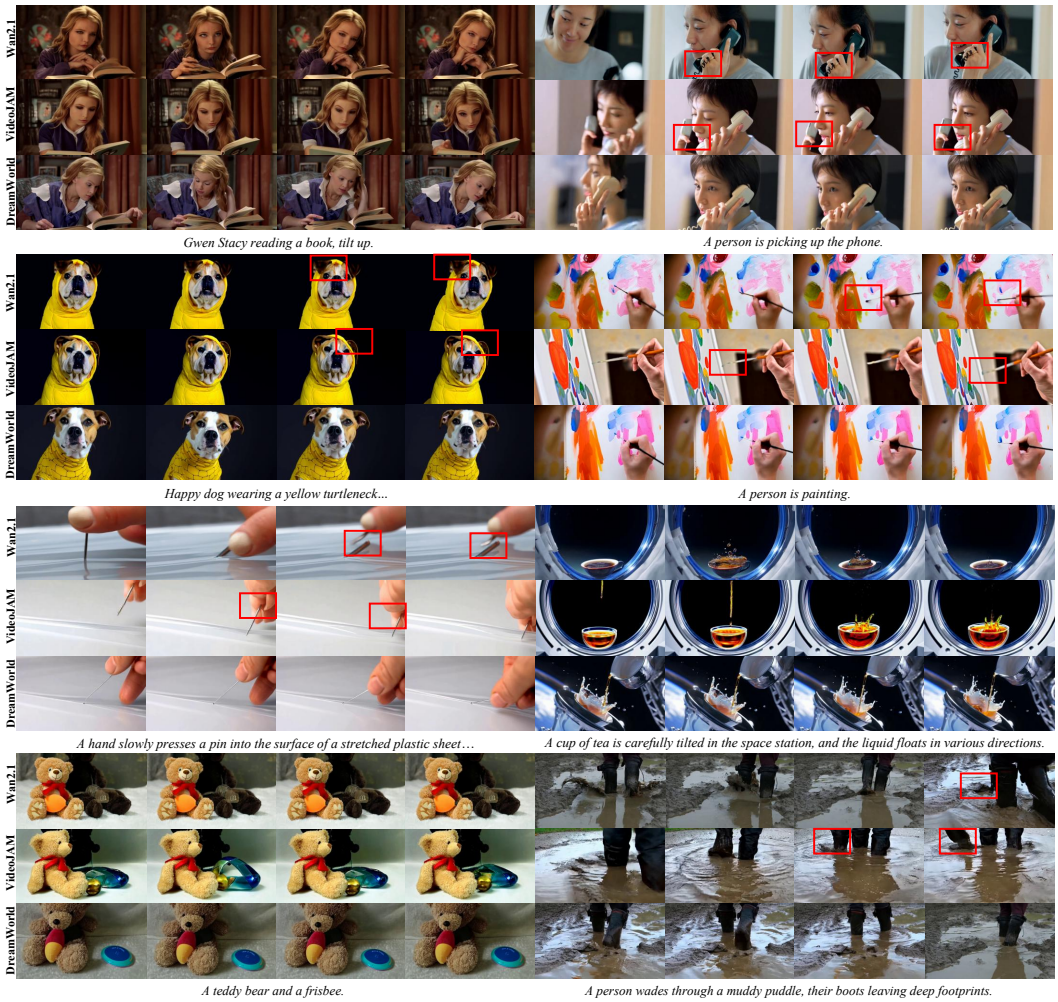


Figure 3: **Qualitative comparison.** DreamWorld outperforms baselines by maintaining semantic realism, spatial integrity, and temporal precision. In contrast, competitor models frequently exhibit geometric penetrations and uncanny distortions

4.4 Ablation Studies

The efficacy of DreamWorld is validated through a three-dimensional analysis, with VBench consistently employed as the default benchmark for ablation experiments, barring any specific exceptions.

Effectiveness of World Knowledge The necessity of the comprehensive WorldKnowledge stack is rigorously validated by progressively evaluating its constituent feature subsets. As shown in Table 5, relying on the spatial prior (**VGGT Only**) provides basic geometric grounding but lacks semantic coherence. The integration of the semantic prior(**VGGT + DINOv2**) significantly bolsters the model's capabilities, particularly in maintaining object permanence and ensuring precise text-video alignment. Nevertheless, the optimal performance is achieved only by the full model, confirming that the synergy of spatial guidance, semantic context, and temporal consistency is essential for robust and realistic world simulation.

Effectiveness of CCA. The qualitative efficacy of Consistent Constraint Annealing (CCA) is evidenced by the visual artifacts observed in the static weighting baseline, as shown in Figure 4. Without the dynamic schedule, optimization conflicts manifest as severe abnormal highlighting and exposure anomalies, particularly visible in the Balcony scenes where lighting becomes unnaturally intense. Furthermore, temporal instability is evident in the Moving Sheep, which suffers from



Figure 4: **Effectiveness of CCA.** Without CCA, the generated videos suffer from severe flickering and abnormal highlighting artifacts, leading to degraded visual quality.

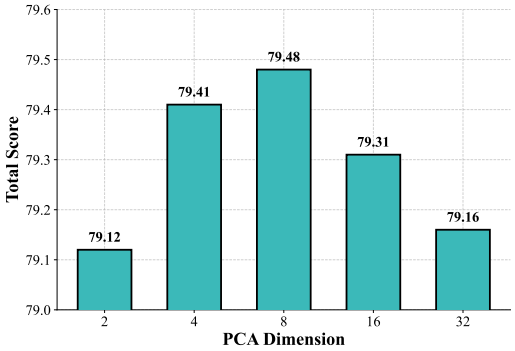


Figure 5: **Influence of Dimensionality reduction.** Ablations on the PCA dimensionality using a 10k dataset demonstrate that a reduction to 8 dimensions yields the optimal performance.

high-frequency flickering and texture jittering. The proposed CCA strategy effectively eliminates these artifacts, allowing the model to refine visual details without interference, thus ensuring both photorealism and temporal smoothness.More ablation experiments are provided in Appendix D.

Effectiveness of Multi-Source Inner-Guidance. We investigate the necessity of integrating diverse priors through an ablation study on the Multi-Source Inner-Guidance mechanism (Table 5). The results indicate that removing any individual component leads to consistent performance degradation. Notably, excluding Temporal Priors causes the most significant drop in Quality and Overall scores, the absence of Text Guidance or Semantic Priors severely impairs the Semantic score. Ultimately, DreamWorld achieves superior performance, confirming that the synergistic integration of text, semantic, spatial, and temporal cues is essential for a robust physics-aware world model.

Influence of Dimensionality reduction. We investigate the sensitivity of our model to the PCA dimensionality, which controls the information bottleneck of world knowledge priors. To quantify this impact, we report the VBench Total Score across varying dimensions.As illustrated in Figure 5, the performance peaks at a dimensionality of 8, achieving a Total Score of 79.48. We observe that lower dimensions (e.g., 2 or 4) result in over-compression, failing to retain critical semantic and geometric details necessary for effective structural guidance. Conversely, excessively high dimensions retain redundant information, which complicates the joint feature integration within the diffusion backbone and leads to a degradation in overall synthesis quality. Consequently, we adopt a PCA dimension of 8 as the optimal configuration for our expert feature representations.

5 Conclusion

In this work, we presented **DreamWorld**, a unified framework designed to bridge the prevailing gap between high-fidelity video synthesis and intrinsic world coherence. We identify that previous approaches often focus on distilling a single expert model. However, directly enforcing simultaneous alignment across multiple knowledge sources is prone to collapsing into trivial solutions due to conflicting optimization objectives. We introduced a softer **Joint World Modeling Paradigm** that facilitates the synergistic integration of multi-source expert knowledge. This gentle alignment approach, complemented by our *Consistent Constraint Annealing (CCA)* strategy and *Multi-Source Inner-Guidance*, effectively harmonizes the intricate interplay between structural logic and generative freedom. DreamWorld establishes a new state-of-the-art on standard benchmarks, validating its potential as a robust foundation for next-generation general-purpose world simulators.

Limitations. Despite these advancements, our approach is currently constrained by the computational resources and the diversity of the training datasets. Future research could explore optimizing the efficiency of this multi-source integration and incorporating more diverse data curation to further enhance the universality of the simulated world.

References

[1] M. Assran, A. Bardes, D. Fan, Q. Garrido, R. Howes, Mojtaba, Komeili, M. Muckley, A. Rizvi, C. Roberts, K. Sinha, A. Zholus, S. Arnaud, A. Gejji, A. Martin, F. R. Hogan, D. Dugas, P. Bojanowski, V. Khalidov, P. Labatut, F. Massa, M. Szafraniec, K. Krishnakumar, Y. Li, X. Ma, S. Chandar, F. Meier, Y. LeCun, M. Rabbat, and N. Ballas. V-jepa 2: Self-supervised video models enable understanding, prediction and planning, 2025.

[2] H. Bansal, Z. Lin, T. Xie, Z. Zong, M. Yarom, Y. Bitton, C. Jiang, Y. Sun, K.-W. Chang, and A. Grover. Videophy: Evaluating physical commonsense for video generation, 2024.

[3] O. Bar-Tal, H. Chefer, O. Tov, C. Herrmann, R. Paiss, S. Zada, A. Ephrat, J. Hur, G. Liu, A. Raj, Y. Li, M. Rubinstein, T. Michaeli, O. Wang, D. Sun, T. Dekel, and I. Mosseri. Lumiere: A space-time diffusion model for video generation, 2024.

[4] A. Bardes, Q. Garrido, J. Ponce, M. Rabbat, Y. LeCun, M. Assran, and N. Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv:2404.08471*, 2024.

[5] J. Berg, C. Zhu, Y. Bao, I. Durugkar, and A. Gupta. Semantic world models, 2025.

[6] A. Bhowmik, D. Korzhenkov, C. G. M. Snoek, A. Habibian, and M. Ghafoorian. Moalign: Motion-centric representation alignment for video diffusion models, 2025.

[7] T. Brooks, B. Peebles, C. Holmes, W. DePue, Y. Guo, L. Jing, D. Schnurr, J. Taylor, T. Luhman, E. Luhman, C. Ng, R. Wang, and A. Ramesh. Video generation models as world simulators. 2024.

[8] J. Bruce, M. Dennis, A. Edwards, J. Parker-Holder, Y. Shi, E. Hughes, M. Lai, A. Mavalankar, R. Steigerwald, C. Apps, Y. Aytar, S. Bechtle, F. Behbahani, S. Chan, N. Heess, L. Gonzalez, S. Osindero, S. Ozair, S. Reed, J. Zhang, K. Zolna, J. Clune, N. de Freitas, S. Singh, and T. Rocktäschel. Genie: Generative interactive environments, 2024.

[9] H. Chefer, U. Singer, A. Zohar, Y. Kirstain, A. Polyak, Y. Taigman, L. Wolf, and S. Sheynin. Videojam: Joint appearance-motion representations for enhanced motion generation in video models, 2025.

[10] B. Chen, D. M. Monso, Y. Du, M. Simchowitz, R. Tedrake, and V. Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion, 2024.

[11] H. Duan, H.-X. Yu, S. Chen, L. Fei-Fei, and J. Wu. Worldscore: A unified evaluation benchmark for world generation, 2025.

[12] H. Fu, C. Li, X. Liu, J. Gao, A. Celikyilmaz, and L. Carin. Cyclical annealing schedule: A simple approach to mitigating kl vanishing, 2019.

[13] M. Fuest, V. T. Hu, and B. Ommer. Maskflow: Discrete flows for flexible and efficient long video generation, 2025.

[14] D. Ha and J. Schmidhuber. World models. 2018.

[15] Y. HaCohen, B. Brazowski, N. Chiprut, Y. Bitterman, A. Kvochko, A. Berkowitz, D. Shalem, D. Lifschitz, D. Moshe, E. Porat, E. Richardson, G. Shiran, I. Chachy, J. Chetboun, M. Finkelson, M. Kupchick, N. Zabari, N. Guetta, N. Kotler, O. Bibi, O. Gordon, P. Panet, R. Benita, S. Armon, V. Kulikov, Y. Inger, Y. Shiftan, Z. Melumian, and Z. Farbman. Ltx-2: Efficient joint audio-visual foundation model, 2026.

[16] J. Ho and T. Salimans. Classifier-free diffusion guidance, 2022.

[17] B. K. Horn and B. G. Schunck. Determining optical flow. *Artificial Intelligence*, 17(1):185–203, 1981.

[18] Z. Huang, Y. He, J. Yu, F. Zhang, C. Si, Y. Jiang, Y. Zhang, T. Wu, Q. Jin, N. Chanpaisit, Y. Wang, X. Chen, L. Wang, D. Lin, Y. Qiao, and Z. Liu. Vbench: Comprehensive benchmark suite for video generative models, 2023.

[19] J. Jang, J. Kim, K. Baek, and N. Kwak. Multi-dimensional preference alignment by conditioning reward itself, 2025.

[20] D. Jiang, M. Wang, L. Li, L. Zhang, H. Wang, W. Wei, G. Dai, Y. Zhang, and J. Wang. No other representation component is needed: Diffusion transformers can provide representation guidance by themselves, 2025.

[21] Y. Jin, Z. Sun, N. Li, K. Xu, K. Xu, H. Jiang, N. Zhuang, Q. Huang, Y. Song, Y. Mu, and Z. Lin. Pyramidal flow matching for efficient video generative modeling, 2025.

[22] B. Kang, Y. Yue, R. Lu, Z. Lin, Y. Zhao, K. Wang, G. Huang, and J. Feng. How far is video generation from world model: A physical law perspective, 2025.

[23] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models, 2020.

[24] W. Kong, Q. Tian, Z. Zhang, R. Min, Z. Dai, J. Zhou, J. Xiong, X. Li, B. Wu, J. Zhang, K. Wu, Q. Lin, J. Yuan, Y. Long, A. Wang, A. Wang, C. Li, D. Huang, F. Yang, H. Tan, H. Wang, J. Song, J. Bai, J. Wu, J. Xue, J. Wang, K. Wang, M. Liu, P. Li, S. Li, W. Wang, W. Yu, X. Deng, Y. Li, Y. Chen, Y. Cui, Y. Peng, Z. Yu, Z. He, Z. Xu, Z. Zhou, Z. Xu, Y. Tao, Q. Lu, S. Liu, D. Zhou, H. Wang, Y. Yang, D. Wang, Y. Liu, J. Jiang, and C. Zhong. Hunyuanvideo: A systematic framework for large video generative models, 2025.

[25] T. Kouzelis, E. Karypidis, I. Kakogeorgiou, S. Gidaris, and N. Komodakis. Boosting generative image modeling via joint image-feature synthesis, 2026.

[26] J.-Y. Lee, B. Cha, J. Kim, and J. C. Ye. Aligning text to image in diffusion models is easier than you think, 2025.

[27] Z. Li, L. Meng, G. Chao, W. Wu, X. Yan, Y. Yang, Z. Qi, and X. Meng. Semantic-space-intervened diffusive alignment for visual classification, 2025.

[28] B. Lin, Z. Li, X. Cheng, Y. Niu, Y. Ye, X. He, S. Yuan, W. Yu, S. Wang, Y. Ge, et al. Uniworld: High-resolution semantic encoders for unified visual understanding and generation. *arXiv preprint arXiv:2506.03147*, 2025.

[29] Y. Lipman, R. T. Q. Chen, H. Ben-Hamu, M. Nickel, and M. Le. Flow matching for generative modeling, 2023.

[30] Y. Liu, W.-J. Shu, Z. Huang, H. Zheng, Y. Wang, M. Zhang, S.-N. Lim, and H. Yang. Alignvid: Training-free attention scaling for semantic fidelity in text-guided image-to-video generation, 2025.

[31] X. Ma, Y. Wang, X. Chen, G. Jia, Z. Liu, Y.-F. Li, C. Chen, and Y. Qiao. Latte: Latent diffusion transformer for video generation, 2025.

[32] C. Min, D. Zhao, L. Xiao, J. Zhao, X. Xu, Z. Zhu, L. Jin, J. Li, Y. Guo, J. Xing, L. Jing, Y. Nie, and B. Dai. Driveworld: 4d pre-trained scene understanding via world models for autonomous driving, 2024.

[33] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski. Dinov2: Learning robust visual features without supervision, 2024.

[34] J. Parker-Holder, P. Ball, J. Bruce, V. Dasagi, K. Holsheimer, C. Kaplanis, A. Moufarek, G. Scully, J. Shar, J. Shi, S. Spencer, J. Yung, M. Dennis, S. Kenjeyev, S. Long, V. Mnih, H. Chan, M. Gazeau, B. Li, F. Pardo, L. Wang, L. Zhang, F. Besse, T. Harley, A. Mitenkova, J. Wang, J. Clune, D. Hassabis, R. Hadsell, A. Bolton, S. Singh, and T. Rocktäschel. Genie 2: A large-scale foundation world model. 2024.

[35] W. Peebles and S. Xie. Scalable diffusion models with transformers, 2023.

[36] R. Po, Y. Nitzan, R. Zhang, B. Chen, T. Dao, E. Shechtman, G. Wetzstein, and X. Huang. Long-context state-space video world models, 2025.

[37] Y. Qin, Z. Shi, J. Yu, X. Wang, E. Zhou, L. Li, Z. Yin, X. Liu, L. Sheng, J. Shao, L. Bai, W. Ouyang, and R. Zhang. Worldsimbench: Towards video generation models as world simulators, 2024.

[38] K. Song, B. Chen, M. Simchowitz, Y. Du, R. Tedrake, and V. Sitzmann. History-guided video diffusion, 2025.

[39] G. Team. Mochi 1. `https://github.com/genmoai/models`, 2024.

[40] Z. Teed and J. Deng. Raft: Recurrent all-pairs field transforms for optical flow, 2020.

[41] J. Tong, Y. Mou, H. Li, M. Li, Y. Yang, M. Zhang, Q. Chen, T. Liang, X. Hu, Y. Zheng, X. Chen, J. Zhao, X. Huang, and X. Qiu. Thinking with video: Video generation as a promising multimodal reasoning paradigm, 2025.

[42] P. Vincent. A connection between score matching and denoising autoencoders. *Neural computation*, 23(7):1661–1674, 2011.

[43] T. Wan, A. Wang, B. Ai, B. Wen, C. Mao, C.-W. Xie, D. Chen, F. Yu, H. Zhao, J. Yang, J. Zeng, J. Wang, J. Zhang, J. Zhou, J. Wang, J. Chen, K. Zhu, K. Zhao, K. Yan, L. Huang, M. Feng, N. Zhang, P. Li, P. Wu, R. Chu, R. Feng, S. Zhang, S. Sun, T. Fang, T. Wang, T. Gui, T. Weng, T. Shen, W. Lin, W. Wang, W. Wang, W. Zhou, W. Wang, W. Shen, W. Yu, X. Shi, X. Huang, X. Xu, Y. Kou, Y. Lv, Y. Li, Y. Liu, Y. Wang, Y. Zhang, Y. Huang, Y. Li, Y. Wu, Y. Liu, Y. Pan, Y. Zheng, Y. Hong, Y. Shi, Y. Feng, Z. Jiang, Z. Han, Z.-F. Wu, and Z. Liu. Wan: Open and advanced large-scale video generative models, 2025.

[44] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotny. Vggt: Visual geometry grounded transformer, 2025.

[45] J. Wang, A. Ma, K. Cao, J. Zheng, Z. Zhang, J. Feng, S. Liu, Y. Ma, B. Cheng, D. Leng, Y. Yin, and X. Liang. Wisa: World simulator assistant for physics-aware text-to-video generation, 2025.

[46] H. Wu, D. Wu, T. He, J. Guo, Y. Ye, Y. Duan, and J. Bian. Geometry forcing: Marrying video diffusion and 3d representation for consistent world modeling, 2025.

[47] T. Wu, S. Yang, R. Po, Y. Xu, Z. Liu, D. Lin, and G. Wetzstein. Video world models with long-term spatial memory, 2025.

[48] Z. Xiao, Y. Lan, Y. Zhou, W. Ouyang, S. Yang, Y. Zeng, and X. Pan. Worldmem: Long-term consistent world simulation with memory, 2026.

[49] Z. Yang, J. Teng, W. Zheng, M. Ding, S. Huang, J. Xu, Y. Yang, W. Hong, X. Zhang, G. Feng, D. Yin, Y. Zhang, W. Wang, Y. Cheng, B. Xu, X. Gu, Y. Dong, and J. Tang. Cogvideox: Text-to-video diffusion models with an expert transformer, 2025.

[50] S. Yu, S. Kwak, H. Jang, J. Jeong, J. Huang, J. Shin, and S. Xie. Representation alignment for generation: Training diffusion transformers is easier than you think, 2025.

[51] S. Yu, S. Kwak, H. Jang, J. Jeong, J. Huang, J. Shin, and S. Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025.

[52] X. Zhang, J. Liao, S. Zhang, F. Meng, X. Wan, J. Yan, and Y. Cheng. Videorepa: Learning physics for video generation through relational alignment with foundation models, 2025.

[53] P. Zhao, R. Luan, W. Zhang, P. Wu, and S. He. Guiding cross-modal representations with mllm priors via preference alignment, 2025.

[54] B. Zheng, N. Ma, S. Tong, and S. Xie. Diffusion transformers with representation autoencoders, 2025.

[55] D. Zheng, Z. Huang, H. Liu, K. Zou, Y. He, F. Zhang, L. Gu, Y. Zhang, J. He, W.-S. Zheng, Y. Qiao, and Z. Liu. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness, 2025.

[56] Z. Zheng, X. Peng, T. Yang, C. Shen, S. Li, H. Liu, Y. Zhou, T. Li, and Y. You. Open-sora: Democratizing efficient video production for all, 2024.

A Motivation Experiments

Investigating the feasibility of simultaneously assimilating comprehensive physical and semantic knowledge, we implemented an extended experimental scheme labeled **Extend-VideoREPA**, incorporating three independent projection layers—configured with output dimensions of 768, 2048, and 768—to map the VAE latent into the distinct embedding spaces of **DINOv2**, **VGGT**, and **VideoMAEv2**.

By formulating the total training objective as the summation of the standard diffusion loss and the independent Token Relation Distillation (TRD) terms from each expert branch, we compel the generative backbone to concurrently reconstruct the internal relational topology of these diverse feature spaces, thereby internalizing multi-source knowledge through a unified soft alignment paradigm.

B Implementation Details

Training Configuration. We initialize the generative backbone with the pre-trained weights of Wan2.1, while the newly introduced expanded projection layers are initialized to zero to strictly preserve the initial generative distribution. We employ the AdamW optimizer with hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.99$, and a weight decay of $\lambda = 0.2$. The learning rate is set to $1e-5$ with seed 42, utilizing a linear warmup for the first 400 steps followed by a consistency-constrained annealing strategy.To optimize memory efficiency and throughput, we utilize Mixed Precision Training (BF16) alongside gradient checkpointing.

Feature Extraction and Alignment. For the World Knowledge priors, we employ visual foundation models to extract offline features, ensuring high-quality guidance signals. Specifically, dense optical flow fields are first computed using RAFT [40]. Crucially, to align with the generative backbone, these 2D displacement fields are converted into RGB visualizations (as detailed in Section 3.2) and subsequently encoded by the frozen Wan2.1 VAE [43]. This maps the motion dynamics into the native latent space, resulting in a temporal feature dimension of $C_{temporal} = 16$. For semantic and spatial representations, we utilize the frozen DINOv2 [33] and VGGT [44] encoders. After spatial alignment and temporal pooling, we apply PCA to compress the channel dimensions of these features to $C_{semantic} = 8$ and $C_{spatial} = 8$, respectively. Consequently, the final composite input fed into the expanded projection layer consists of the noisy video latents $C_{vae} = 16$ concatenated with the WorldKnowledge priors, yielding a total input channel dimensionality of $C_{total} = 48$.

Inference Setup. During inference, videos are generated using the Flow Matching Euler Discrete Scheduler with 50 denoising steps. We apply the proposed Multi-Faceted Self-Guidance with a text guidance scale of $w_{text} = 5$ to ensure prompt fidelity. The auxiliary guidance scales for motion, semantics, and geometry are empirically set to $w_{temporal} = w_{semantic} = w_{spatial} = 1$. This configuration was chosen based on ablation studies to provide subtle yet effective structural rectification without introducing high-frequency artifacts.

Inference Cost Analysis Integrating World Knowledge introduces computational overhead during inference. Specifically, the joint latent space increases peak VRAM by approximately 10.3%. Additionally, inference time doubles from 2 to 4 minutes per video. This latency stems from the Multi-Source Inner-Guidance (Section 3.4), requiring additional forward passes through the network. Despite the increased computational and memory footprints, the substantial improvements in trajectory control and generation quality justify these trade-offs.

C Detailed Quantitative Comparison

C.1 Results of WorldScore

Detailed multidimensional metrics for WorldScore are presented in TableC.1 and TableC.2.

Table C.1: Quantitative comparison on WorldScore (Static Attributes). The best and second-best results are highlighted in **bold** and underlined.

Method	Camera Control	Object Control	Content Alignment	3D Consistency	Photometric Consistency	Style Consistency	Subjective Quality
Wan2.1-T2V-1.3B	23.50	52.34	**66.83**	72.31	74.40	34.08	52.21
Wan2.1-T2V-1.3B(FT.)	**23.66**	49.66	65.35	71.73	74.62	**39.04**	**53.83**
VideoJAM	22.48	**57.34**	59.32	71.53	75.39	25.90	44.03
DreamWorld (Ours)	22.81	51.94	64.95	**73.16**	**77.55**	35.75	52.45

Table C.2: Quantitative comparison on WorldScore (Dynamic Attributes, Summary and Overall Score).The best and second-best results are highlighted in **bold** and underlined.

Method	Dynamic Attributes			Summary		Overall Score
	Motion Accuracy	Motion Magnitude	Motion Smoothness	Static Score	Dynamic Score	
Wan2.1-T2V-1.3B	32.36	**34.14**	68.18	53.66	44.89	51.03
Wan2.1-T2V-1.3B(FT.)	**35.49**	31.35	64.76	53.98	43.86	50.95
VideoJAM	33.38	32.51	**71.95**	50.86	**45.94**	49.38
DreamWorld (Ours)	34.75	32.44	69.02	**54.09**	45.41	**51.48**

Table C.3: **Results of VBench** (Part 1/2): Consistency, Motion, and Object-level Metrics.

Method	Subject Consistency	Background Consistency	Temporal Flickering	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality	Object Class	Multiple Objects	Human Action
Wan2.1-T2V-1.3B	91.83	94.71	99.17	96.51	65.00	57.24	59.14	76.09	53.66	74.60
Wan2.1-T2V-1.3B(FT.)	93.59	95.81	**99.36**	97.21	54.08	**61.13**	63.27	79.90	58.20	78.98
VideoJAM	91.51	**96.01**	99.13	96.05	73.88	58.41	62.68	79.22	59.66	79.20
DreamWorld (Ours)	**93.62**	94.95	98.81	**98.07**	**79.16**	59.59	**66.76**	**81.32**	**65.03**	**81.20**

Table C.4: **Results of VBench** (Part 2/2): Appearance, Style, and Overall Scores.

Method	Color	Spatial Relationship	Scene	Appearance Style	Temporal Style	Overall Consistency	Quality Score	Semantic Score	Overall Score
Wan2.1-T2V-1.3B	89.93	62.37	20.03	**22.91**	22.23	24.25	79.81	65.43	76.93
Wan2.1-T2V-1.3B(FT.)	88.57	63.31	28.55	22.38	23.93	24.94	81.26	68.47	78.71
VideoJAM	88.92	66.17	28.34	22.67	**24.03**	**25.01**	81.18	69.08	78.76
DreamWorld (Ours)	**92.61**	**70.47**	**29.71**	22.14	23.82	24.87	**83.49**	**70.89**	**80.97**

Table C.5: **VBench-2.0 results** (part 1/3).

Method	Human Identity	Dynamic Spatial Relationship	Complex Landscape	Instance Preservation	Multi-View Consistency	Human Clothes	Dynamic Attribute	Complex Plot
Wan2.1-T2V-1.3B	64.21	26.57	16.22	80.41	**14.63**	**99.55**	22.34	10.53
Wan2.1-T2V-1.3B (FT.)	**72.43**	**28.50**	**21.66**	88.24	8.24	90.13	22.71	**11.77**
VideoJAM	72.14	24.08	15.66	**90.41**	6.79	97.74	24.37	10.33
DreamWorld (Ours)	71.68	24.15	14.66	86.14	12.41	99.11	**26.73**	10.44

Table C.6: **VBench-2.0 results** (part 2/3).

Method	Mechanics	Human Anatomy	Composition	Human Interaction	Motion Rationality	Material	Diversity	Motion Order Understanding
Wan2.1-T2V-1.3B	85.71	64.51	40.19	15.00	37.93	57.77	51.65	7.31
Wan2.1-T2V-1.3B (FT.)	**88.63**	68.71	39.35	14.66	37.35	**60.67**	46.91	6.42
VideoJAM	80.31	66.51	44.81	17.33	**39.37**	56.62	53.84	7.53
DreamWorld (Ours)	86.30	**69.35**	**45.01**	**19.00**	38.51	55.17	**56.76**	**8.88**

Table C.7: **VBench-2.0 results** (part 3/3).

Method	Camera Motion	Thermotics	Creativity Score	Commonsense Score	Controllability Score	Human Fidelity Score	Physics Score	Total Score
Wan2.1-T2V-1.3B	19.75	65.27	45.92	59.17	16.81	76.09	**55.85**	50.77
Wan2.1-T2V-1.3B (FT.)	**23.14**	60.49	43.13	62.80	**18.41**	77.09	54.51	51.18
VideoJAM	15.04	61.97	49.32	**64.89**	16.33	78.68	52.42	52.33
DreamWorld (Ours)	14.81	**66.43**	**50.89**	61.82	16.95	**80.11**	55.07	**52.97**



Figure 6: **Effectiveness of CCA.** Without CCA, the generated videos suffer from severe flickering and abnormal highlighting artifacts, leading to degraded visual quality.

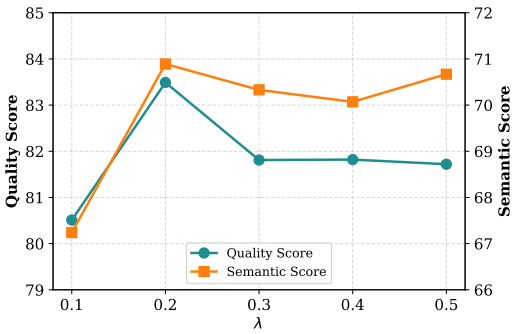


Figure 7: **Influence of loss weight λ.** Quantitative results show that $\lambda = 0.2$ achieves the best trade-off between generation quality and semantic alignment.

Table E.1: **Quantitative ablation of the CCA schedule.** We compare different strategies for scheduling the world loss weight.

Configuration	Quality	Semantic	Overall
No World Loss (Baseline)	81.26	68.47	78.71
Fixed World-Loss Weight	80.76	67.85	78.19
Increasing Schedule	81.13	68.02	78.51
Decreasing Schedule (Ours)	**83.49**	**70.89**	**80.97**

Table E.2: **Comparison of guidance configurations.** Different guidance configurations are compared for cost and quality.

Configuration	Latency	VRAM	Overall
No Guidance	48	25.55	68.84
Text Guidance Only	100	26.55	78.13
DINOv2	144	28.03	78.82
DINOv2 + VGGT	193	30.73	79.71
Multi-Source (Ours)	240	33.42	**80.97**

C.2 VBench and VBench2.0 Results

We report the detailed breakdown of metrics on both VBench and VBench-2.0 benchmarks, providing a granular analysis of model performance. The comprehensive evaluations for VBench are presented in Table C.3 and Table C.4, while the VBench-2.0 results are detailed in Table C.5, Table C.6, and Table C.7.

As evidenced by the quantitative data, **DreamWorld** demonstrates robust superiority over the baselines. In VBench, our method achieves significant margins in Spatial Relationship and Dynamic Degree, which contributes to a leading Overall Score of 80.97. Furthermore, on the more challenging VBench-2.0, DreamWorld secures the highest Total Score (52.97) and excels in complex categories such as Human Fidelity and Motion Order Understanding. These results validate that our joint world modeling paradigm effectively internalizes multi-modal priors, resulting in generated content with real-world fidelity.

D More Ablation Study

Effectiveness of CCA. Further ablation studies demonstrating the significant impact of the CCA mechanism on overall model performance are visually presented in Figure6.To further quantitatively investigate the effect of the CCA schedule, we compare different world-loss weighting strategies, including a fixed weight, an increasing schedule, and our decreasing schedule. As shown in Table E.1, the decreasing schedule consistently achieves the best performance across all three metrics. This result suggests that applying stronger structural constraints at earlier training stages while gradually relaxing them at later stages provides a more effective balance between world knowledge learning and visual generation quality.

Guidance Efficiency. We further investigate the computational overhead and performance gains of different guidance configurations. Specifically, the single World-Guidance branch uses only DINOv2 features, while the two World-Guidance branches jointly use DINOv2 and VGGT features. As shown

in Table E.2, introducing additional guidance sources improves generation quality at the cost of increased inference latency and GPU memory consumption. Our Multi-Source Guidance achieves the highest VBench overall score of 80.97, compared with 78.82 using DINOv2 alone and 79.71 using DINOv2 and VGGT.

Influence of λ. We investigate the sensitivity of our model to the loss weight λ, which controls the strength of the world knowledge priors. To quantify the trade-off between generative fidelity and condition adherence, we report both the Quality Score and Semantic Score on VBench across varying λ values. As illustrated in Figure. 7, the performance of both metrics peaks at $\lambda = 0.2$. We observe that lower values fail to effectively enforce the physical and semantic constraints, whereas excessive weights interfere with the diffusion backbone's distribution, leading to a degradation in visual quality. Consequently, we adopt $\lambda = 0.2$ as the optimal configuration for our full model.

E Broader Impacts

Our DreamWorld demonstrates a marked improvement over existing video generation baselines, establishing new state-of-the-art benchmarks in world consistency and physical commonsense. This enhancement is pivotal in advancing general-purpose world simulators and high-fidelity video synthesis technologies. While our contributions are essentially within the scope of methodology and fundamental theory, we are fully aware of the ethical challenges facing generative media, such as the potential misuse for creating deceptive deepfakes. Although this work does not introduce new specific ethical risks beyond those inherent in existing video synthesis techniques, we recognize that highly realistic generation can be exploited for disinformation. We share the position of the broader academic community in firmly supporting responsible and sustainable technology development to ensure the safety and fairness of generative tools, including the adoption of appropriate mitigation strategies.

F Safeguards

To mitigate the potential dual-use risks and misuse of our video generation architecture, we release the model and code under a strict Acceptable Use Policy (AUP). Users who wish to access and utilize our model are required to explicitly agree to this license, which strictly prohibits the generation of harmful, deceptive, or non-consensual content. This licensing approach serves as a primary safeguard to promote responsible downstream usage while still supporting the open research community.

G More qualitative results

We provide additional qualitative comparisons in Figure 8 to further substantiate the superiority of DreamWorld over the baseline Wan2.1. These visual results underscore our model's robustness across semantic, spatial, and temporal dimensions. Specifically, The baseline frequently succumbs to semantic hallucinations and spatial conflicts, where objects unnaturally fuse or distort. Conversely, DreamWorld maintains clear object boundaries and logical interaction physics.

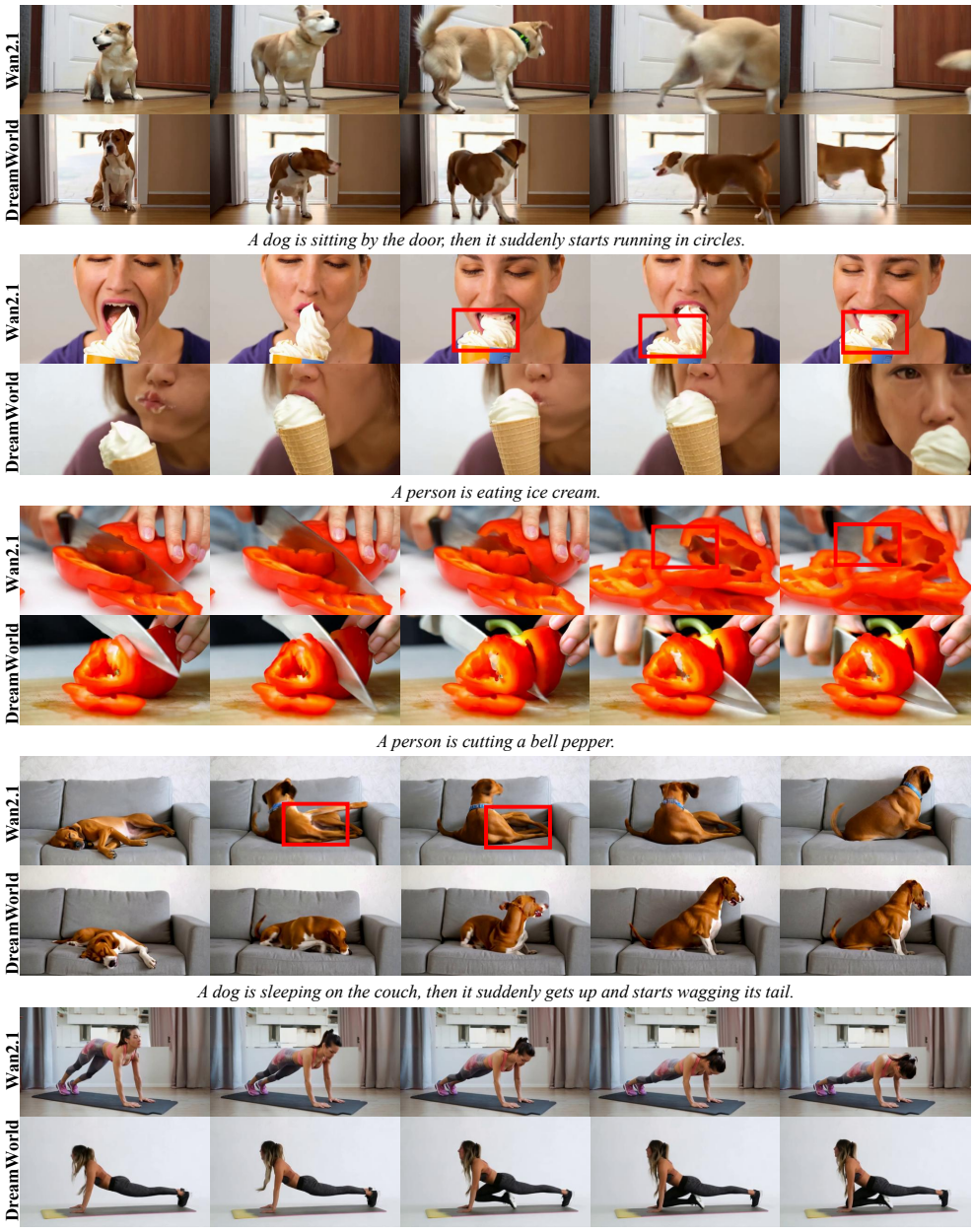


Figure 8: **Additional qualitative comparisons.**The red rectangles highlight anomalies(e.g., object penetration, unnatural disappearance, and limb distortion).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The Abstract and Introduction sections accurately reflect the paper's contribution and scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are shown in Section 5.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: Driven by a discovery in our experiments, we proposed a new method that significantly beats the baseline. It is a purely practical approach, so there are no theoretical assumptions or proofs involved.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We present all the necessary details for reproducing the experiments are provided in Section 4.1 and Appendix B. This includes information on the model, dataset, training parameters, and other relevant aspects.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All code, data, and instructions for reproduction are available at `https://anonymous.4open.science/r/Code-86B0`

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide comprehensive details on the training and testing setup, including data splits, hyperparameter selection, and the optimizer used, in Section 4.1 and Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: Our evaluation relies on standardized benchmark protocols that focus on fixed test sets and deterministic metrics, where reporting single-point scores is the conventional standard for performance comparison.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments in this paper were conducted using GPU resources. Detailed information about the GPU configuration can be found in Section4.1.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research did not violate any NeurIPS ethical guidelines.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Appendix E for the broader impacts of our work.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: Our model is released under a strict Acceptable Use Policy to prevent misuse, as detailed in Appendix F

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We give proper credit to the creators of the models and datasets referenced in our paper.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: We do not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: Our experiments focus on video generation and do not involve any crowdsourcing or human subject studies. Therefore, there are no instructions, screenshots, or compensation details to include.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The research presented in this paper does not include any human subject studies or crowdsourcing experiments.

    Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLM does not impact the core methodology in our paper.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.