LOOKING AHEAD: MODELING FUTURE EVOLUTION IN AUTOREGRESSIVE VIDEO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Autoregressive video diffusion models generate videos chunk by chunk, yet producing plausible continuations does not guarantee that they capture how scenes evolve. Under teacher forcing, strong temporal redundancy may allow local visual continuity to serve as a shortcut for current-chunk denoising, reducing the pressure to model visual evolution. We introduce Future Chunk Prediction (FCP), an auxiliary objective jointly optimized with teacher-forced denoising that predicts temporally separated chunks without access to the intervening visual context. Drawing supervision directly from training videos, FCP requires no additional pretrained visual encoder and leaves autoregressive inference unchanged. Our method achieves superior physical plausibility over standard teacher-forced autoregressive training with a lower overall training compute budget, highlighting the effectiveness and efficiency of future-chunk supervision for modeling visual evolution. Intriguingly, our visualizations suggest that FCP fosters a more structured representation space, which may be valuable for building autoregressive world models.

1 INTRODUCTION

Autoregressive video generation (Yang et al., 2025; Huang et al., 2025; Zhu et al., 2026; Yuan et al., 2026) has emerged as an important direction in video generation, offering a flexible paradigm in which visual sequences can be extended continuously. It typically operates through chunk-wise temporal rollout, where each new chunk is generated conditioned on preceding visual context. As an important pathway toward generative world modeling (Bruce et al., 2024; Valevski et al., 2025), such a history-conditioned generator is expected not merely to produce plausible successive chunks, but also to model how visual scenes evolve over time.

However, the ability to extend a video does not by itself guarantee that the model has learned how the scene should evolve. We argue that a key reason is standard teacher forcing, where the current chunk is denoised given ground-truth history. Due to the strong temporal redundancy in videos, neighboring chunks often share highly similar appearance and motion patterns. As a result, the model may rely heavily on local visual continuity to recover the current chunk, creating a shortcut that provides limited pressure to capture the underlying dynamics of scene evolution.

This motivates us to seek supervision that goes beyond what can be readily inferred from local visual continuity. A natural place to look is farther into the future. As the prediction target moves farther from the observed history, local appearance and motion continuity provide less direct support for recovering the target, thereby placing greater pressure on the model to capture temporal progression. Importantly, this supervision comes directly from the training video itself, without requiring external pretrained representations or semantic alignment (Yu et al., 2025; Zhang et al., 2025a).

Motivated by this intuition, we introduce Future Chunk Prediction (FCP), a simple auxiliary objective for autoregressive video diffusion training. In addition to denoising the target chunk under standard teacher forcing, FCP uses the learned intermediate representation to predict a target beyond the current chunk, without exposing the model to the intervening ground-truth visual context. This complementary objective places additional training pressure on the learned representations to capture information relevant to visual evolution, a capability fundamental to effective temporal rollout in autoregressive video generation. Importantly, FCP is introduced only during training and leaves the original autoregressive inference process unchanged.

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.

To assess whether this additional supervision improves the modeling of visual evolution, we focus on physical plausibility as a concrete and challenging evaluation dimension. Our method demonstrates improved physical plausibility over standard teacher-forced autoregressive training under matched data exposure and training budget, highlighting the effectiveness and efficiency of future-chunk supervision. These results highlight the potential of exploiting richer predictive supervision from existing video data, rather than relying solely on increased training resources.

Beyond the improvements in generation quality, we further examine how FCP affects the intermediate video representations. Interestingly, the introduced supervision induces more pronounced spatiotemporal organization in the representation space, suggesting that FCP also meaningfully reshapes the representations learned during training. This finding highlights the potential of future prediction for learning representations better suited to autoregressive video generation, which may further benefit generative world modeling.

2 Related Work

2.1 Autoregressive Video Diffusion

Conventional video diffusion models typically synthesize fixed-duration clips in a single denoising process (Ho et al., 2022; Blattmann et al., 2023; Team, 2025; Wang et al., 2025a). Autoregressive video diffusion instead extends video chunk by chunk, conditioning on preceding visual context (Voleti et al., 2022; Chen et al., 2024; Teng et al., 2025). This formulation supports streaming generation and provides a basis for interactive visual simulation and generative world modeling (Bruce et al., 2024; He et al., 2025; Yang et al., 2025; Valevski et al., 2025).

Prior work addresses rollout quality by adapting bidirectional models into causal few-step generators through distribution matching or adversarial training (Yin et al., 2025; Lin et al., 2025), exposing models to self-generated histories or extended student rollouts (Huang et al., 2025; Cui et al., 2025; Liu et al., 2025), and examining the causal compatibility of teacher supervision with student generation (Zhu et al., 2026; Zheng et al., 2026). Other approaches manage historical context through active windows, cached states, or global anchors (Gao et al., 2025; Yang et al., 2025; Liu et al., 2025), while selective retrieval and cache compression balance long-range consistency against computational and memory costs (Cai et al., 2025; Samuel et al., 2026; Ji et al., 2026). Our approach complements these directions by adding future-chunk supervision to teacher-forced training without prescribing rollout distillation or inference-time context management.

2.2 Representation Alignment for Generation

Aligning intermediate diffusion features with vision foundation models can accelerate training and improve generation quality (Yu et al., 2025). Video extensions use cross-frame and multi-feature alignment (Hwang et al., 2025; Lee et al., 2025), or relational supervision targeting physical plausibility (Zhang et al., 2025a). These methods rely on targets from external pretrained encoders (Oquab et al., 2024; Wang et al., 2023), which may complicate scaling to large video corpora. Our approach instead derives supervision from future chunks in the training videos, without an additional pretrained vision model, offering a data-native path toward scaling.

Our work was developed independently of and concurrently with Next Forcing (Xu et al., 2026), the closest related approach. We include relevant comparisons in our experiments and provide a more detailed discussion in the appendix.

3 Preliminaries

3.1 Flow Matching

Flow matching learns a time-dependent vector field that continuously connects the data distribution to a simple noise distribution. Let $\mathbf{x}$ denote a clean video sample, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ denote Gaussian noise, and $t \sim \mathcal{U}(0, 1)$. We define the intermediate state along a linear probability path as

$$\widetilde{\mathbf{x}}_t = (1 - t)\mathbf{x} + t\boldsymbol{\epsilon}, \tag{1}$$

whose target velocity is $\boldsymbol{\epsilon} - \mathbf{x}$. A conditional velocity model $\mathbf{v}_\theta$ is trained with

$$\mathcal{L}_{\mathrm{FM}}(\theta) = \mathbb{E}_{\mathbf{x},\boldsymbol{\epsilon},t,\mathbf{c}}\left[\left\|\mathbf{v}_\theta(\widetilde{\mathbf{x}}_t, t, \mathbf{c}) - (\boldsymbol{\epsilon} - \mathbf{x})\right\|_2^2\right], \tag{2}$$

where $\mathbf{c}$ denotes optional conditioning information. At inference, a numerical sampler follows the learned velocity field to transform noise at $t = 1$ into a video sample at $t = 0$.

3.2 Autoregressive Video Generation

Autoregressive video generation, also referred to as streaming video generation, generates videos through chunk-wise temporal rollout (Yin et al., 2025). In diffusion-based instantiations, each autoregressive step generates a multi-frame chunk conditioned on the preceding observed or generated context. Repeatedly applying a fixed-length chunk generator supports long-form video generation over a flexible number of chunks. This chunk-wise formulation is distinct from discrete-token autoregressive video generation (Hong et al., 2023; Yu et al., 2024; Kondratyuk et al., 2024; Wang et al., 2024), which sequentially predicts tokenized representations, and from single-shot generation, which produces a fixed-length clip in one invocation.

Let $\mathbf{x}_{1:K} = \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$ denote a video divided into K ordered, non-overlapping chunks of equal length, and let $\mathbf{c}$ denote the text condition. The conditional distribution of the complete video can be factorized as

$$p(\mathbf{x}_{1:K} \mid \mathbf{c}) = \prod_{k=1}^{K} p(\mathbf{x}_k \mid \mathbf{x}_{<k}, \mathbf{c}), \tag{3}$$

where $\mathbf{x}_{<k}$ denotes the chunks preceding $\mathbf{x}_k$.

3.3 Teacher-Forced Training

Different training paradigms for autoregressive video generation construct the historical context using either clean chunks (Voleti et al., 2022; Huang et al., 2025; Zhang et al., 2025b) or noise-perturbed ones (Chen et al., 2024; Ruhe et al., 2024). Here, we adopt the clean-history setting. A common realization of this setting is teacher-forced training, in which the historical context $\mathbf{x}_{<k}$ is taken directly from the ground-truth video. Under the flow-matching formulation introduced in Section 3.1, Gaussian noise $\boldsymbol{\epsilon}_k$ and a flow time t are sampled to perturb the target chunk $\mathbf{x}_k$ along the linear probability path, yielding $\widetilde{\mathbf{x}}_{k,t}$. The model predicts the corresponding velocity while conditioned on the clean preceding chunks $\mathbf{x}_{<k}$ and text condition $\mathbf{c}$:

$$\mathcal{L}_{\mathrm{TF}}(\theta) = \mathbb{E}_{\mathbf{x}_{1:K},\mathbf{c},k,t,\boldsymbol{\epsilon}_k}\left[\left\|\mathbf{v}_\theta(\widetilde{\mathbf{x}}_{k,t}, t, \mathbf{x}_{<k}, \mathbf{c}) - (\boldsymbol{\epsilon}_k - \mathbf{x}_k)\right\|_2^2\right]. \tag{4}$$

Together, this defines a teacher-forced flow-matching objective for chunk-wise autoregressive video generation.

4 Method

In this section, we detail our proposed method. Section 4.1 introduces the general formulation of Future Chunk Prediction (FCP) and formalizes its auxiliary future-prediction objective, while Section 4.2 presents the FCP architecture adopted in this work.

4.1 Future Chunk Prediction

Building upon the teacher-forced formulation in Section 3.3, each target chunk is denoised conditioned on its temporally contiguous ground-truth history. As discussed in Section 1, the strong temporal redundancy of video allows this immediately preceding correct context to provide substantial support for current-chunk denoising, reducing the pressure to capture temporal progression and potentially hindering the learning of representations that capture visual evolution. This observation suggests a simple design principle: we seek a learning signal that is less directly supported by such immediately preceding context, yet remains intrinsic to the training video itself, allowing the supervision to scale naturally with video data without relying on external pretrained representations.

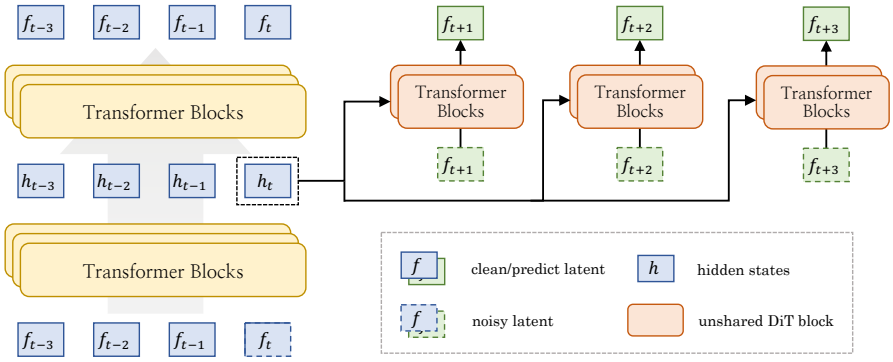


Figure 1: **Overview of Future Chunk Prediction (FCP).** Intermediate representations from the autoregressive backbone condition auxiliary transformer blocks for future-chunk prediction.

To this end, we introduce Future Chunk Prediction (FCP), which supplements next-chunk denoising with an auxiliary task of predicting temporally more distant chunks from the same video. By excluding the intervening chunks from the visual context, this task discourages reliance on local temporal continuity and encourages stronger predictive representations of visual evolution.

Formally, re-indexing a factor in equation 3, standard next-chunk prediction given the observed prefix $\mathbf{x}_{\leq k}$ models

$$p(\mathbf{x}_{k+1} \mid \mathbf{x}_{\leq k}, \mathbf{c}).$$

FCP instead requires the same available context to support prediction of a farther target:

$$p_{\mathrm{FCP}}(\mathbf{x}_{k+\Delta} \mid \mathbf{x}_{\leq k}, \mathbf{c}), \qquad \Delta > 1, \quad k + \Delta \leq K, \tag{5}$$

where Δ is an integer temporal offset measured in chunks. In this formulation, the intervening chunks $\mathbf{x}_{k+1:k+\Delta-1}$ are excluded from the conditioning context, yielding a supervisory signal that is distinct from and complementary to standard teacher forcing.

A natural implementation of FCP uses an auxiliary predictor with dedicated parameters. In preliminary experiments, sharing generator parameters between teacher-forced denoising and masked-context future prediction degraded temporal coherence, possibly because the two objectives place competing demands on shared parameters.

However, training this predictor in isolation would leave the main generator unaffected. We therefore condition the predictor on the generator's intermediate hidden states and backpropagate the auxiliary loss through them (DeepSeek-AI, 2024). This design regularizes these representations to capture visual evolution and better support autoregressive rollout.

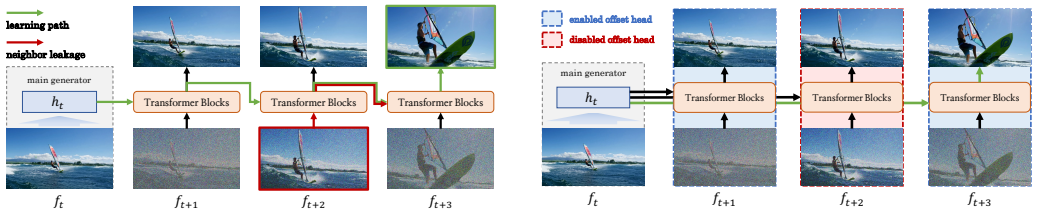
For a practical realization of the above formulation, let $\mathbf{h}_k$ denote the generator's intermediate hidden states corresponding to chunk k. The auxiliary prediction is then expressed as

$$p_{\mathrm{FCP}}(\mathbf{x}_{k+\Delta} \mid \mathbf{h}_k), \qquad \Delta \geq 1. \tag{6}$$

Compared with equation 5, this formulation incorporates two deliberate choices regarding the conditioning information. First, we omit explicit text conditioning, encouraging the auxiliary predictor to derive predictive information from $\mathbf{h}_k$ rather than a separate textual input. Second, we condition the auxiliary predictor on $\mathbf{h}_k$ rather than $\mathbf{h}_{\leq k}$, encouraging a more compact encoding of predictive information while reducing computational cost. We examine the effects of these conditioning choices through ablation studies in Section 5.

4.2 FCP Architecture

Building on the formulation in equation 6, we present one concrete implementation of FCP to evaluate the effectiveness of the proposed approach. For the autoregressive generator, we use a standard video diffusion model (Wang et al., 2025a) whose backbone consists of stacked diffusion transformer (DiT) blocks with chunk-causal attention. The auxiliary predictor adopts a lightweight counterpart of this DiT architecture with a separate set of parameters. Figure 1 illustrates the overall architecture, with component-specific designs described below.



(a) Chained prediction: information from noisy intermediate targets may leak to subsequent predictors along the red arrows.

(b) Parallel prediction: offset-specific predictors can run concurrently across GPUs and be flexibly enabled or disabled during training.

Figure 2: Two variants of future-prediction supervision, illustrated through the green paths from $\mathbf{h}_t$ to f_{t+3}. The chained design passes hidden states between predictors, whereas the parallel design conditions each predictor directly on the source representations.

Source Representations. We extract intermediate features associated with chunk k in both clean and noise-perturbed states, denoted by $\mathbf{h}_k^{\text{clean}}$ and $\mathbf{h}_k^{\text{noisy}}$, respectively. Clean context provides historical information for subsequent chunk generation, while the noisy target is progressively denoised to produce the current chunk. We consider both roles essential to autoregressive generation and therefore regularize both types of representations through future-prediction supervision.

Parallel Future Prediction. Each future offset has a separate predictor conditioned directly on $\mathbf{h}_k$, rather than on hidden states from preceding predictors (DeepSeek-AI, 2024; Xu et al., 2026). In the chained design (Figure 2a), information from the noisy input for f_{t+2} can reach the predictor for f_{t+3} along the red path. Given video's temporal redundancy, this may create a shortcut that weakens the pressure on the source representation to encode future evolution. Our parallel design (Figure 2b) removes inter-offset information flow, preventing this particular shortcut. Its independent branches could also run concurrently across GPUs or be flexibly loaded and unloaded during training.

Feature Conditioning. To capture information at different semantic levels, we extract features from multiple generator layers, concatenate them along the channel dimension at corresponding token positions, and fuse them with an MLP. Inspired by prior AdaLN-based conditioning designs (Wang et al., 2026b; Pan et al., 2026; Wang et al., 2026a), we inject the fused features into every auxiliary DiT block through token-wise AdaLN modulation, preserving fine-grained conditioning across spatiotemporal positions.

Asymmetric Text Conditioning. Unlike the text-conditioned main generator, the auxiliary predictors omit text cross-attention from every DiT block. Consistent with equation 6, this asymmetry blocks a direct pathway through which text prompts could leak cues about future events, encouraging the predictors to rely on the generator's source representations instead. This design also reduces the predictors' parameter count and computational cost.

4.3 Training and Inference

Training. We jointly optimize the autoregressive generator and auxiliary predictors using flow matching. For each offset Δ in a predefined set $\mathcal{D}$, a separate auxiliary predictor follows the flow-matching objective in Section 3.1, conditioned on either clean or noisy source representations. Target noise and flow time are resampled for each auxiliary prediction. Let $\mathcal{L}_\Delta^{\text{clean}}$ and $\mathcal{L}_\Delta^{\text{noisy}}$ denote the corresponding losses, each averaged over all valid supervision pairs $(k, k{+}\Delta)$ satisfying $k{+}\Delta \le K$. The resulting joint objective takes the form:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{TF}} + \sum_{\Delta \in \mathcal{D}} w_\Delta \frac{\mathcal{L}_\Delta^{\text{clean}} + \mathcal{L}_\Delta^{\text{noisy}}}{2}. \tag{7}$$

Here, w_Δ controls the contribution of each offset. Clean and noisy source representations share the predictor for a given offset, and the auxiliary losses backpropagate through these representations into the generator.

Table 1: Physics-IQ Verified results for image-to-video generation (Rädsch et al., 2026). Baseline scores are primarily from the benchmark. IQ is the overall score. Params counts visual-denoiser parameters. **Bold** and underlined mark the best and second-best autoregressive scores.

Model	Params	IQ ↑	SP ↑	ST ↑	WS ↑	MSE ↑
Bidirectional						
Wan2.2-5B (Wang et al., 2025a)	5B	24.9	34.2	22.1	21.1	22.1
Sora 2 (OpenAI, 2025)	–	26.5	37.3	27.0	26.9	14.8
Cosmos3-Nano (Aditi et al., 2026)	A8B	29.1	40.4	22.0	24.6	29.5
Wan2.2-14B (Wang et al., 2025a)	A14B	32.2	51.1	20.5	28.5	28.9
HunyuanVideo-1.5 (Team, 2025)	8.3B	33.4	47.1	26.9	29.7	30.0
Grok Imagine Video (xAI, 2026)	–	34.8	52.7	21.4	35.7	29.6
Autoregressive						
Causal Forcing (Zhu et al., 2026)	1.3B	12.1	21.4	8.5	11.4	7.2
Causal Forcing $_{\text{Stage 1}}$ (Zhu et al., 2026)	1.3B	19.2	26.7	15.4	16.9	17.9
Helios-Base (Yuan et al., 2026)	14B	26.9	32.7	31.7	21.8	21.4
MAGI-1-4.5B (Teng et al., 2025)	4.5B	<u>31.1</u>	27.6	**49.5**	20.8	26.6
Wan2.1-1.3B + FCP (Ours)	1.3B	30.9	<u>36.2</u>	29.6	<u>23.2</u>	**34.5**
Wan2.2-5B + FCP (Ours)	5B	**33.8**	**36.8**	<u>41.1</u>	**25.2**	<u>31.9</u>

Inference. At inference, we discard all FCP-specific modules, retaining the same autoregressive generator and rollout procedure as the teacher-forcing baseline without additional overhead. Qualitative visualizations of the auxiliary predictors suggest that they capture plausible future evolution despite limited visual fidelity. We leave their potential use for sampling acceleration through early-stage denoising or generation enhancement through prediction ensembling to future work.

5 Experiments

5.1 Implementation Details

We initialize our autoregressive generators from official Wan checkpoints (Wang et al., 2025a). For our implementation, we choose the image-to-video setting, treating the initial frame as fixed clean context and generating subsequent chunks of three latent frames. Each auxiliary predictor contains three DiT blocks initialized from the corresponding generator; the generators and predictors are then jointly optimized. Further model and training details are provided in Appendix A.

5.2 Benchmarks

Modeling visual evolution involves diverse aspects of temporal dynamics, making its comprehensive evaluation an open challenge (Li et al., 2025b; Ma et al., 2026; Wu et al., 2026). In this work, we focus primarily on physical plausibility (Brooks et al., 2024; Kang et al., 2025; Agarwal et al., 2025), motivated by the availability of dedicated benchmarks and systematic research on its evaluation. This choice is further motivated by the role of physical plausibility in shaping the perceived realism of generated videos. We adopt Physics-IQ Verified (Rädsch et al., 2026) as our primary benchmark because its reference-based numerical metrics assess generated continuations against real-world recordings without relying on VLM-based judgments. We further evaluate on WorldModelBench (Li et al., 2025a), PAI-Bench-G (Zhou et al., 2025), and PhyGround (Lin et al., 2026) as complementary physics-related benchmarks.

5.3 Datasets

We train our models on a mixture of internally sourced videos and WISA-80K (Wang et al., 2025b), using the original text annotations from each source. The training data primarily comprises general-purpose real-world videos, without using data from physics simulators. We select videos longer than five seconds and use approximately the first five seconds, sampled at the default frame rate of the

Table 2: Controlled Physics-IQ Verified (Rädsch et al., 2026) comparisons on Wan2.2-TI2V-5B. The shaded row and columns mark our default configuration and overall IQ scores, respectively.

ID	Method	I2V				V2V			
		IQ ↑	SP ↑	ST ↑	WS ↑	IQ ↑	SP ↑	ST ↑	WS ↑
(a)	Teacher Forcing	30.4	34.8	32.8	24.0	40.6	48.5	40.8	37.8
(b)	+ FCP (3 offsets)	32.1	33.6	39.3	23.2	43.0	50.3	44.6	39.8
(c)	+ FCP (5 offsets)	33.8	36.8	41.1	25.2	44.4	51.1	47.2	40.7
(d)	+ FCP (5 offsets, chained)	31.0	35.6	33.0	23.9	39.9	47.6	38.9	36.6
(e)	+ FCP (offsets 1, 3, 5)	33.1	36.9	38.5	25.5	43.5	51.8	43.8	40.6

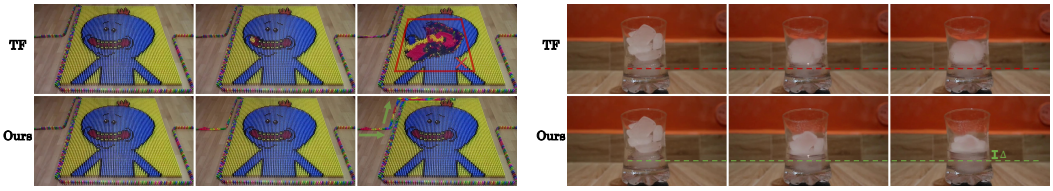


Figure 3: Qualitative comparison of teacher forcing (top) and FCP (bottom). Left: FCP predicts domino toppling along the expected direction (green arrows), whereas teacher forcing produces an anomalous collapse at the center (red box). Right: FCP captures the rise in water level as the ice melts (green Δ), a detail missed by teacher forcing.

corresponding model. Incorporating physics-simulator data (Aditi et al., 2026) may further improve physical plausibility as training scales, but we leave this possibility to future work.

5.4 Main Results

Physics-IQ Verified. Differences in training data and settings limit the attribution of cross-model performance gaps to methodological improvements. We therefore place greater emphasis on the controlled comparisons in Tables 2 and 4, where matched data and experimental settings enable a fairer assessment. Compared with teacher forcing, FCP raises the IQ score from 27.6 to 30.9 for the 1.3B model. For the 5B model, it improves I2V and V2V scores from 30.4 to 33.8 and from 40.6 to 44.4, respectively. These consistent gains support the effectiveness of future-prediction supervision in improving physical plausibility across scales and conditioning settings.

Table 1 provides broader context for these results. Our 5B model outperforms larger bidirectional models, including Wan2.2-14B (Wang et al., 2025a), HunyuanVideo-1.5 (Team, 2025), and Cosmos3-Nano (Aditi et al., 2026), as well as autoregressive pre-training results from comparably sized MAGI-1-4.5B (Teng et al., 2025) and larger Helios-Base (Yuan et al., 2026).

Qualitative Analysis of Physical Dynamics. Figure 3 complements the quantitative results with two examples of temporal evolution. FCP preserves the direction of domino toppling and captures the water-level rise visible in the ice-melting sequence, whereas the teacher-forcing baseline exhibits an anomalous central collapse in the former and misses the subtle level change in the latter. Together, these examples suggest that FCP better captures physically plausible visual evolution, which is central to the perceived realism of generated videos.

Other Physics-Related Benchmarks. We further evaluate on physics-related benchmarks, including WorldModelBench (Li et al., 2025a), PAI-Bench-G (Zhou et al., 2025), and PhyGround (Lin et al., 2026). As shown in Table 3, FCP generally outperforms teacher forcing under matched settings, supporting the generalizability of our approach. These evaluations rely on VLM-based judgments of specified criteria, which may be less sensitive to subtle differences in the timing and magnitude of physical changes than direct comparison with real-world reference videos. This limited sensitivity may partly explain the modest gains on some metrics.

Table 3: Results on WorldModelBench, PAI-Bench-G, and PhyGround. All variants are based on Wan2.2-TI2V-5B and use matched training settings.

Benchmark	Metric	Teacher Forcing	+ FCP (Ours)
WorldModelBench (Li et al., 2025a)	Instruction ↑	1.77	**1.81** (+0.04)
	Physical Laws ↑	4.28	**4.47** (+0.19)
	Common Sense ↑	1.17	**1.36** (+0.19)
	Total ↑	7.22	**7.64** (+0.42)
PAI-Bench-G (Zhou et al., 2025)	AV ↑	63.54	**64.81** (+1.27)
	Common Sense ↑	90.90	**92.10** (+1.20)
	Human ↑	81.62	**82.46** (+0.84)
	Industry ↑	85.65	**87.24** (+1.59)
	Physics ↑	90.09	**90.29** (+0.20)
	Robot ↑	70.56	**72.95** (+2.39)
	Domain Score ↑	81.13	**82.38** (+1.25)
PhyGround (Lin et al., 2026)	General ↑	2.63	**2.66** (+0.03)
	Physics ↑	2.75	**2.80** (+0.05)

Table 4: Ablation of the main FCP design components. The shaded row and column mark the default setting and overall IQ score, respectively.

ID	Clean	Noisy	Text	Chain	IQ ↑	SP ↑
(a)					27.6	32.8
(b)	✓				28.8	33.0
(c)		✓			29.9	34.7
(d)	✓	✓			30.9	36.2
(e)	✓	✓	✓		29.8	34.8
(f)	✓	✓		✓	30.1	34.3

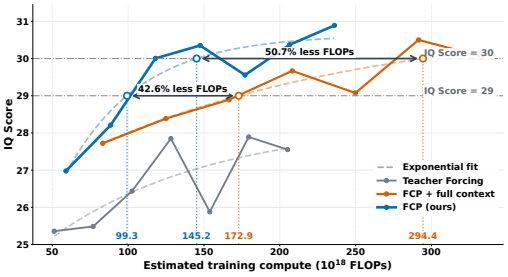


Figure 4: Training convergence on Physics-IQ Verified versus estimated training compute.

5.5 Ablation Studies

We primarily conduct ablation studies on Wan2.1-1.3B under identical settings across variants; detailed settings and results are provided in Appendices A.2 and B.1, respectively.

Source Representations. While prior representation-alignment methods (Yu et al., 2025) primarily regularize intermediate features along the denoising process, clean historical context in autoregressive video generation also serves a predictive role, carrying information from the immediate past to support anticipation of future visual evolution. Results in Table 4 support this view: regularizing clean-context or noisy-target representations improves IQ over teacher forcing by 1.2 and 2.3 points, respectively, while their combination yields a 3.3-point gain.

Parallel Future Prediction. Chained prediction provides a natural alternative (DeepSeek-AI, 2024; Xu et al., 2026) to our parallel design; however, as illustrated in Figure 2a, it may exploit intermediate future information as a shortcut. Consistent with this concern, Table 4 shows that parallel prediction improves the IQ score by 0.8, from 30.1 to 30.9. The gap is more pronounced in the controlled 5B comparison in Table 2, where parallel prediction achieves IQ-score gains of 2.8 for I2V and 4.5 for V2V over the corresponding five-offset chained variant. This larger degradation may arise because the longer chain provides more stages for propagating intermediate future information, while the higher-capacity model may more readily exploit the resulting shortcut. Together, these results show that our parallel design consistently outperforms the chained alternative across model scales and conditioning settings. It also supports nonconsecutive future offsets, as evaluated in row (e) of Table 2, through the independent branches shown in Figure 2b.

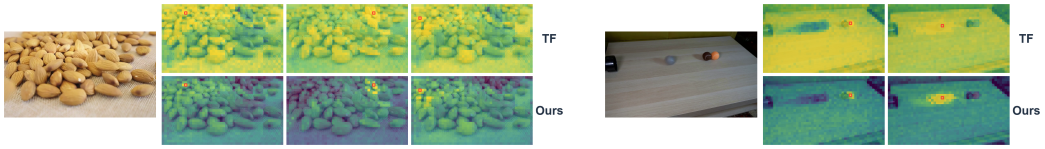


Figure 5: **Spatial Feature Similarity Maps.** Each map shows within-frame similarity to a selected query token. The top and bottom rows correspond to teacher forcing and FCP, respectively.

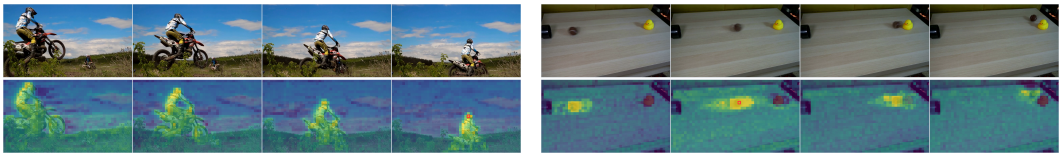


Figure 6: **Spatiotemporal Feature Similarity Maps.** The top row shows video frames, while the bottom row visualizes similarity to a selected query token across space and time.

Training Efficiency. The closely related approach Next Forcing (Xu et al., 2026) conditions future prediction on the full historical representations $\mathbf{h}_{\leq k}$, motivating a full-context FCP variant for comparison. Figure 4 compares teacher forcing, this full-context variant, and our default configuration. Compared with standard teacher forcing, our configuration achieves higher IQ scores at comparable estimated training compute, indicating that its gains do not arise simply from additional computation. Furthermore, our configuration is estimated to require 42.6% and 50.7% less training compute than the full-context variant to reach the same IQ scores of 29 and 30, respectively. These results indicate that extending the conditioning context from $\mathbf{h}_k$ to $\mathbf{h}_{\leq k}$ provides no additional benefit for modeling video evolution, while the lower computational cost of $\mathbf{h}_k$ makes our formulation better suited to scaling autoregressive video models.

Text-Free Future Prediction. Comparing rows (d) and (e) in Table 4, adding explicit text conditioning to the auxiliary predictors reduces the IQ score by 1.1. This supports withholding text from future predictors, as direct textual cues may reveal future evolution and weaken their reliance on visual representations.

5.6 Representation Space Analysis

Beyond physical-plausibility evaluation, we visualize the generator's intermediate representations using query-to-token similarity maps. As shown in Figure 5, teacher forcing produces diffuse responses extending into unrelated regions, whereas FCP yields more selective responses aligned with the queried object and visually related instances. Figure 6 further shows that high-similarity regions remain aligned with corresponding content as it moves across frames. Together, these visualizations suggest that FCP-trained representations capture clearer spatial structure and meaningful temporal correspondence. We believe that such semantically structured representations may provide a useful foundation for future world models.

6 Conclusion

We introduced Future Chunk Prediction (FCP), an auxiliary objective that predicts temporally separated chunks from the same training video in autoregressive video diffusion. Optimized jointly with the standard teacher-forced denoising objective, FCP provides complementary supervision to the generator's intermediate representations without requiring an additional pretrained encoder. The auxiliary predictors are removed at inference, leaving the autoregressive generation process unchanged. Controlled comparisons show improved physical plausibility across model scales and conditioning settings. More surprisingly, our visualizations suggest that FCP also induces a more structured representation space, which we believe may be important for autoregressive world models. Together, these findings support future prediction as a promising training principle for autoregressive video generation.

REPRODUCIBILITY STATEMENT

Sections 4 and 5 describe our method, training setup, datasets, and evaluation protocols, with further implementation details in Appendix A. Code and model checkpoints are not included in the submission, but we plan to release them after preparation. For the public portion of our training data, we will provide source links and selection criteria; we will also provide representative examples of the internally sourced data. The complete internal corpus cannot be released, so exact reproduction of training with identical data may not be possible; nevertheless, we will make every effort to support reproducibility.

AI USE STATEMENT

Generative AI tools were used to assist with implementing FCP and the associated training, inference, and evaluation code, as well as with translation. We did not use these tools to generate datasets, develop the conceptual framework, formulate mathematical claims, propose or refine research hypotheses, design or provide feedback on methods or experiments, clean or reformat datasets, conduct qualitative data analysis, or interpret results; this work contains no mathematical proofs. We also used generative AI for drafting and editing manuscript text, searching for and summarizing relevant literature, writing-related brainstorming, formatting in-text citations, and suggesting paper structure, titles, section headings, and keywords. The authors reviewed and tested AI-assisted code, checked AI-assisted text and literature recommendations, and take full responsibility for the final content of this work.

REFERENCES

Aditi, Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson G. Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, Aarti Basant, Mukesh Beladiya, Mohammad Qazim Bhat, Zaid Pervaiz Bhat, Dan Blick, Vanni Brighella, Han Cai, Tiffany Cai, Eric Cameracci, Jiaxin Cao, Yulong Cao, Mark Carlson, Carlos Casanova, Ting-Yun Chang, Yan Chang, Yu-Wei Chao, Prithvijit Chattopadhyay, Roshan Chaudhari, Chieh-Yun Chen, Junyu Chen, Ke Chen, Qizhi Chen, Wenkai Chen, Xiaotong Chen, Yu Chen, An-Chieh Cheng, Click Cheng, Xiu Chia, Jeana Choi, Chaeyeon Chung, Wenyan Cong, Yin Cui, Magdalena Dadela, Nalin Dadhich, Wenliang Dai, Joyjit Daw, Alperen Degirmenci, Rodrigo Vieira Del Monte, Robert Denomme, Sameer Dharur, Marco Di Lucca, Ke Ding, Wenhao Ding, Yifan Ding, Yuzhu Dong, Nicole Drumheller, Yilun Du, Aigul Dzhumamuratova, Aleksandr Efitorov, Hamid Eghbalzadeh, Naomi Eigbe, Imad El Hanafi, Hassan Eslami, Benedikt Falk, Jiaojiao Fan, Jim Fan, Amol Fasale, Sergiy Fefilatyev, Liang Feng, Francesco Ferroni, Sanja Fidler, Xiao Fu, Vikram Fugro, Prashant Gaikwad, TJ Galda, Katelyn Gao, Yihuai Gao, Wenhang Ge, Sreyan Ghosh, Arushi Goel, Vivek Goel, Akash Gokul, Rama Govindaraju, Jinwei Gu, Miguel Guerrero, Elfie Guo, Aryaman Gupta, Siddharth Gururani, Hugo Hadfield, Song Han, Ankur Handa, Zekun Hao, Mohammad Harrim, Ali Hassani, Nathan Hayes-Roth, Yufan He, Chris Helvig, and Cyrus Hogg. Cosmos 3: Omnimodal world models for physical AI. *CoRR*, abs/2606.02800, 2026.

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaojiao Fan, Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Samuel Huffman, Pooya Jannaty, Jingyi Jin, Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura Leal-Taixé, Anqi Li, Zhaoshuo Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma, Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page, Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum Reda, Xiaowei Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Yen-Chen Lin, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao, and Artur Zólkowski. Cosmos world foundation model platform for physical AI. *CoRR*, abs/2501.03575, 2025.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *CoRR*, abs/2311.15127, 2023.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. OpenAI technical report, 2024. URL `https://openai.com/index/video-generation-models-as-world-simulators/`.

Jake Bruce, Michael D. Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal M. P. Behbahani, Stephanie C. Y. Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott E. Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In *ICML*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4603–4623. PMLR / OpenReview.net, 2024.

Shengqu Cai, Ceyuan Yang, Lvmin Zhang, Yuwei Guo, Junfei Xiao, Ziyan Yang, Yinghao Xu, Zhenheng Yang, Alan L. Yuille, Leonidas J. Guibas, Maneesh Agrawala, Lu Jiang, and Gordon Wetzstein. Mixture of contexts for long video generation. *CoRR*, abs/2508.21058, 2025.

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

Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Self-forcing++: Towards minute-scale high-quality video generation. *CoRR*, abs/2510.02283, 2025.

DeepSeek-AI. Deepseek-v3 technical report. *CoRR*, abs/2412.19437, 2024.

Kaifeng Gao, Jiaxin Shi, Hanwang Zhang, Chunping Wang, Jun Xiao, and Long Chen. Ca2-vdm: Efficient autoregressive video diffusion model with causal generation and cache sharing. In *ICML*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, Baixin Xu, Hao-Xiang Guo, Kaixiong Gong, Cyrus Wu, Wei Li, Xuchen Song, Yang Liu, Eric Li, and Yahui Zhou. Matrix-game 2.0: An open-source, real-time, and streaming interactive world model. *CoRR*, abs/2508.13009, 2025.

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

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

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

Sungwon Hwang, Hyojin Jang, Kinam Kim, Minho Park, and Jaegul Choo. Cross-frame representation alignment for fine-tuning video diffusion models. *CoRR*, abs/2506.09229, 2025.

Yicheng Ji, Zhizhou Zhong, Jun Zhang, Qin Yang, XiTai Jin, Ying Qin, Wenhan Luo, Shuiyang Mao, Wei Liu, and Huan Li. Forcing-kv: Hybrid KV cache compression for efficient autoregressive video diffusion models. *CoRR*, abs/2605.09681, 2026.

Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. In *ICML*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025.

Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, Krishna Somandepalli, Hassan Akbari, Yair Alon, Yong Cheng, Joshua V. Dillon, Agrim Gupta, Meera Hahn, Anja Hauth, David Hendon, Alonso Martinez, David Minnen, Mikhail Sirotenko, Kihyuk Sohn, Xuan Yang, Hartwig

Adam, Ming-Hsuan Yang, Irfan Essa, Huisheng Wang, David A. Ross, Bryan Seybold, and Lu Jiang. Videopoet: A large language model for zero-shot video generation. In *ICML*, volume 235 of *Proceedings of Machine Learning Research*, pp. 25105–25124. PMLR / OpenReview.net, 2024.

Dohun Lee, Hyeonho Jeong, Jiwook Kim, Duygu Ceylan, and Jong Chul Ye. Improving video diffusion transformer training by multi-feature fusion and alignment from self-supervised vision encoders. *CoRR*, abs/2509.09547, 2025.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph E. Gonzalez, Ion Stoica, Song Han, and Yao Lu. Worldmodelbench: Judging video generation models as world models. In *NeurIPS*, 2025a.

Yumeng Li, William H. Beluch, Margret Keuper, Dan Zhang, and Anna Khoreva. VSTAR: generative temporal nursing for longer dynamic video synthesis. In *ICLR*. OpenReview.net, 2025b.

Juyi Lin, Arash Akbari, Yumei He, Lin Zhao, Haichao Zhang, Arman Akbari, Xingchen Xu, Zoe Y. Lu, Enfu Nan, Hokin Deng, Edmund Yeh, Sarah Ostadabbas, Yun Fu, Jennifer G. Dy, Pu Zhao, and Yanzhi Wang. Phyground: Benchmarking physical reasoning in generative world models. *CoRR*, abs/2605.10806, 2026.

Shanchuan Lin, Ceyuan Yang, Hao He, Jianwen Jiang, Yuxi Ren, Xin Xia, Yang Zhao, Xuefeng Xiao, and Lu Jiang. Autoregressive adversarial post-training for real-time interactive video generation. In *NeurIPS*, 2025.

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. *CoRR*, abs/2509.25161, 2025.

Ziqi Ma, Mengzhan Liufu, and Georgia Gkioxari. Out of sight, out of mind? evaluating state evolution in video world models. *CoRR*, abs/2603.13215, 2026.

OpenAI. Sora 2 system card. Technical report, OpenAI, September 2025. URL `https://cdn.openai.com/pdf/50d5973c-c4ff-4c2d-986f-c72b5d0ff069/sora_2_system_card.pdf`.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024, 2024.

Xichen Pan, Aashu Singh, Satya Narayan Shukla, Xiangjun Fan, Shlok Kumar Mishra, and Saining Xie. Repfusion: Leveraging multimodal priors for denoising in representation space. *CoRR*, abs/2606.14700, 2026.

Tim Rädsch, Yuki M. Asano, Hilde Kuehne, Stefan Bauer, Priyank Jaini, Robert Geirhos, and Carsten T. Lüth. Physics-iq verified. *CoRR*, abs/2606.18943, 2026.

David Ruhe, Jonathan Heek, Tim Salimans, and Emiel Hoogeboom. Rolling diffusion models. In *ICML*, volume 235 of *Proceedings of Machine Learning Research*, pp. 42818–42835. PMLR / OpenReview.net, 2024.

Dvir Samuel, Issar Tzachor, Matan Levy, Micahel Green, Gal Chechik, and Rami Ben-Ari. Fast autoregressive video diffusion and world models with temporal cache compression and sparse attention. *CoRR*, abs/2602.01801, 2026.

Tencent Hunyuan Foundation Model Team. Hunyuanvideo 1.5 technical report. *CoRR*, abs/2511.18870, 2025.

Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, W. Q. Zhang, Weifeng Luo, Xiaoyang Kang, Yuchen Sun, Yue Cao, Yunpeng Huang, Yutong Lin, Yuxin Fang, Zewei Tao, Zheng Zhang, Zhongshu Wang, Zixun Liu, Dai Shi, Guoli Su, Hanwen Sun, Hong Pan, Jie Wang, Jiexin Sheng, Min Cui, Min Hu, Ming Yan, Shucheng Yin, Siran Zhang, Tingting Liu, Xianping Yin, Xiaoyu Yang, Xin Song, Xuan Hu, Yankai Zhang, and Yuqiao Li. MAGI-1: autoregressive video generation at scale. *CoRR*, abs/2505.13211, 2025.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. In *ICLR*. OpenReview.net, 2025.

Vikram Voleti, Alexia Jolicoeur-Martineau, and Chris Pal. MCVD - masked conditional video diffusion for prediction, generation, and interpolation. In *NeurIPS*, 2022.

Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Xiaofeng Meng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *CoRR*, abs/2503.20314, 2025a.

Jing Wang, Ao Ma, Ke Cao, Jun Zheng, Zhanjie Zhang, Jiasong Feng, Shanyuan Liu, Yuhang Ma, Bo Cheng, Dawei Leng, et al. Wisa: World simulator assistant for physics-aware text-to-video generation. *arXiv preprint arXiv:2503.08153*, 2025b.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae V2: scaling video masked autoencoders with dual masking. In *CVPR*, pp. 14549–14560. IEEE, 2023.

Shuai Wang, Liang Li, Yang Chen, Ruopeng Gao, Yao Teng, and Limin Wang. Uniddt: Unifying multimodal understanding and generation with decoupled diffusion transformer. *CoRR*, abs/2606.16255, 2026a.

Shuai Wang, Zhi Tian, Weilin Huang, and Limin Wang. DDT: decoupled diffusion transformer. In *CVPR*, pp. 40633–40642. Computer Vision Foundation, 2026b.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-token prediction is all you need. *CoRR*, abs/2409.18869, 2024.

Keming Wu, Yijing Cui, Wenhan Xue, Qijie Wang, Xuan Luo, Zhiyuan Feng, Zuhao Yang, Sudong Wang, Sicong Jiang, Haowei Zhu, Zihan Wang, Ping Nie, Wenhu Chen, and Bin Wang. Worldreasonbench: Human-aligned stress testing of video generators as future world-state predictors. *CoRR*, abs/2605.10434, 2026.

xAI. Grok imagine api: State-of-the-art video generation across quality, cost, and latency, 2026. URL `https://x.ai/news/grok-imagine-api`. Accessed: 2026-08-28.

Gangwei Xu, Qihang Zhang, Jiaming Zhou, Xing Zhu, Yujun Shen, Xin Yang, and Yinghao Xu. Next forcing: Causal world modeling with multi-chunk prediction. *CoRR*, abs/2606.11187, 2026.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, Song Han, and Yukang Chen. Longlive: Real-time interactive long video generation. *CoRR*, abs/2509.22622, 2025.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Frédo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *CVPR*, pp. 22963–22974. Computer Vision Foundation / IEEE, 2025.

Lijun Yu, José Lezama, Nitesh Bharadwaj Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G. Hauptmann, Boqing Gong, Ming-Hsuan Yang, Irfan Essa, David A. Ross, and Lu Jiang. Language model beats diffusion - tokenizer is key to visual generation. In *ICLR*. OpenReview.net, 2024.

Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *ICLR*. OpenReview.net, 2025.

Shenghai Yuan, Yuanyang Yin, Zongjian Li, Xinwei Huang, Xiao Yang, and Li Yuan. Helios: Real real-time long video generation model. *arXiv preprint arXiv:2603.04379*, 2026.

Xiangdong Zhang, Jiaqi Liao, Shaofeng Zhang, Fanqing Meng, Xiangpeng Wan, Junchi Yan, and Yu Cheng. Videorepa: Learning physics for video generation through relational alignment with foundation models. In *NeurIPS*, 2025a.

Yuan Zhang, Jiacheng Jiang, Guoqing Ma, Zhiying Lu, Bo Wang, Haoyang Huang, Jianlong Yuan, and Nan Duan. Generative pre-trained autoregressive diffusion transformer. In *NeurIPS*, 2025b.

Kaiwen Zheng, Guande He, Min Zhao, Jintao Zhang, Huayu Chen, Jianfei Chen, Chen-Hsuan Lin, Ming-Yu Liu, Jun Zhu, and Qianli Ma. Causal-rcm: A unified teacher-forcing and self-forcing open recipe for autoregressive diffusion distillation in streaming video generation and interactive world models. *CoRR*, abs/2606.25473, 2026.

Fengzhe Zhou, Jiannan Huang, Jialuo Li, Deva Ramanan, and Humphrey Shi. Pai-bench: A comprehensive benchmark for physical AI. *CoRR*, abs/2512.01989, 2025.

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. *CoRR*, abs/2602.02214, 2026.

A Additional Experimental Details

A.1 Model Configurations

We initialize the generators from the official Wan (Wang et al., 2025a) checkpoints listed in Table 5. For all checkpoints, we implement image-to-video generation by keeping the initial image's latent representation as fixed clean context and generating only subsequent chunks. Source features are extracted from the outputs of four selected DiT blocks and concatenated at corresponding token positions. A two-layer MLP with dimensions $4d \to 2d \to d$ and a GELU activation fuses these features, where d is the backbone hidden dimension. Each future offset has an independently parameterized auxiliary branch, including its fusion MLP, patch embedding, DiT blocks, and output head. Within each offset, all these parameters are shared between the clean and noisy source pathways. Auxiliary DiT blocks inherit the hidden and FFN dimensions of the corresponding backbone and are initialized from the layers specified in Table 5. All layer indices are one-based.

Table 5: Model configurations. All layer indices are one-based. Hidden and FFN dimensions apply to both the backbone and its auxiliary DiT blocks.

Configuration	**Wan2.1 T2V-1.3B**	**Wan2.2 TI2V-5B**
Backbone depth	30	30
Hidden dimension d	1536	3072
FFN intermediate dimension	8960	14336
Source feature layers	9, 15, 21, 27	3, 12, 21, 30
Auxiliary DiT depth	3	3
Auxiliary initialization layers	17, 18, 19	28, 29, 30
Future offsets $\mathcal{D}$	$\{1, 2, 3\}$	$\{1, 2, 3, 4, 5\}$

The 1.3B and 5B configurations may appear to follow different design principles. These differences arise from choices made at different stages of development.

For source-layer selection, the 1.3B configuration evolved from an early single-layer variant using layer 15, subsequently expanded to include both shallower and deeper features. For the 5B model, we select approximately evenly spaced layers backward from the final block to provide direct auxiliary supervision across the generator's depth, while excluding the earliest blocks whose features may not yet be sufficiently semantically informative. Computational constraints have prevented a systematic comparison of these source-layer selection strategies, which we may revisit in future scaling efforts.

For auxiliary initialization, the 1.3B configuration follows the early variant using source features from layer 15, with blocks 17–19 used to initialize the auxiliary predictor. For the 5B model, we instead initialize from the final backbone blocks, decoupling this choice from the source-layer configuration and simplifying subsequent scale-up. These blocks are also closer to the output and are chosen to better match the output head, which is likewise initialized from the original generator. Preliminary experiments suggest that this initialization strategy may offer a modest improvement in early-stage convergence.

A.2 Training Configurations

Wan2.1-1.3B. We use Wan2.1-1.3B for early-stage exploration and controlled ablation studies. Training runs for 20,000 steps on eight GPUs using 85-frame clips at a resolution of 768×432 and 16 FPS, with a per-GPU batch size of one and a global batch size of eight. We use AdamW with $(\beta_1, \beta_2) = (0, 0.999)$, a weight decay of 0.01, and a gradient clipping threshold of 0.1. The generator and auxiliary branches share the same learning rate, which is linearly warmed up to 1×10^{-5} over the first 1,000 steps and held constant thereafter. Both use uniformly sampled flow times followed by a timestep shift of 5.0. We assign a weight of 1 to the main teacher-forcing loss and set $w_\Delta = 0.5$ for every future offset in equation 7. During training, we replace the text prompt with an empty string with a probability of 0.1. We maintain an exponential moving average of the

generator parameters with a decay rate of 0.9995, starting at step 2,000, and use the EMA weights corresponding to each evaluated training step. All 1.3B ablation experiments follow these training settings. Unless otherwise specified, the 5B experiments follow the same training settings as the 1.3B model.

Wan2.2-5B. The first stage runs for 25,000 steps on 32 GPUs using 121-frame clips at a resolution of 1024×576 and 24 FPS, with a per-GPU batch size of one and a global batch size of 32. All other training settings follow the 1.3B configuration.

A.3 Evaluation Settings

Sampling Settings. We use UniPC with 50 denoising steps per chunk, a classifier-free guidance scale of 3.0, and a timestep shift of 5.0 for all models. Sampling uses the corresponding EMA weights, with all FCP-specific modules removed. When the inference sequence exceeds the training sequence length, we retain the first three latent frames as attention sinks during KV-cache eviction.

Physics-IQ Verified (Rädsch et al., 2026). We evaluate both image-to-video and video-to-video generation. For I2V, we condition on the benchmark-provided initial image. For V2V, we use the longest suffix of the three-second conditioning video compatible with the model's latent and chunk-length constraints. Both settings generate five-second continuations, and evaluation excludes the conditioning video. Our models use 16:9 resolutions and require no spatial cropping.

Unless otherwise specified, we use best-practice prompts (BPP) and evaluate all 198 benchmark samples using the official Physics-IQ Verified implementation without modifying its metrics. Evaluation uses each model's output frame rate specified in Section A.2, with reference videos at the corresponding frame rate. We perform four independent runs with seeds 42–45, compute benchmark scores separately for each run, and report their mean.

Other Physics-Related Benchmarks (Li et al., 2025a; Zhou et al., 2025; Lin et al., 2026). We evaluate the teacher-forced and FCP-trained Wan2.2 5B models on all samples of WorldModel-Bench, PAI-Bench-G, and PhyGround in the I2V setting, using otherwise identical training settings. We use the benchmark-provided initial images and text prompts, resize images while preserving their aspect ratio, and crop them to the 1024×576 target resolution when needed. All three evaluations generate five-second videos at 24 FPS. For each benchmark, we use its official evaluation code and unmodified metrics, compute scores separately for seeds 42–45, and report their mean in Table 3.

A.4 Evaluation of Public Models

Since Physics-IQ Verified serves as our primary benchmark, we supplement its reported comparisons with additional evaluations of relevant public models and settings that are not fully covered in the benchmark paper. Unless otherwise specified, all evaluations in this subsection follow the common protocol described in Section A.3.

Wan2.2-TI2V-5B (Wang et al., 2025a). For the official Wan2.2-TI2V-5B checkpoint, we use a resolution of 1280×704 and consistently crop and align the conditioning images, reference videos, and evaluation masks. We also tested this checkpoint at the 1024×576 resolution used by our 5B models, but observed lower scores, potentially reflecting a mismatch with its intended resolution settings.

Causal Forcing (Zhu et al., 2026). We evaluate both the official final checkpoint and the Stage 1 checkpoint of Causal Forcing using the repository's inference code and default sampling settings. We additionally evaluate Stage 1 because it uses standard teacher-forced training, providing a comparison more closely aligned with our training setting before rollout distillation. Both checkpoints follow the same preprocessing and evaluation protocol. For Physics-IQ Verified, we condition on the benchmark-provided initial image and BPP prompt, generating 81 frames at 16 FPS. We resize while preserving aspect ratio and consistently crop the conditioning images, reference videos, and evaluation masks to 832×480. We evaluate all 198 samples with the official metrics, compute scores separately for seeds 42–45, and report their mean.

Helios Base (Yuan et al., 2026). We evaluate the official Helios Base checkpoint using the unmodified inference code and repository-default sampling configuration. We use the officially recommended resolution of 640×384, applying aspect-ratio-preserving resizing followed by center cropping to the conditioning inputs, reference videos, and evaluation masks with identical transformation parameters. Helios Base generates 120-frame, five-second continuations at 24 FPS.

MAGI-1-4.5B (Teng et al., 2025). We evaluate the official MAGI-1-4.5B checkpoint using its unmodified inference code and default sampling parameters. For Physics-IQ Verified, we directly use the benchmark-provided initial images and BPP prompts. We resize the conditioning images, reference videos, and evaluation masks consistently to 768×432 without cropping, and generate five-second videos at 24 FPS. Following the common protocol above, we evaluate all 198 samples with the official Physics-IQ Verified metrics, compute scores separately for seeds 42–45, and report their mean.

A.5 Representation Visualization Protocol

We use the final 20,000-step EMA checkpoints of the teacher-forced and FCP-trained Wan2.1-1.3B models to visualize their representations on real videos. Both models receive the same clean video latents, and we extract intermediate features from a single selected source layer of the main generator for each comparison. The layer and manually selected query-token position are matched between the two models, with the query vector extracted separately from each model's features. We compute cosine similarity between the query and other tokens in the same frame for spatial maps; for temporal maps, a query from one frame is compared with tokens across the remaining frames. Similarity values are displayed directly on the token grid without spatial interpolation. Within each comparison, the teacher-forced and FCP maps use the same color map and a shared value range determined from that comparison's similarity values.

A.6 Auxiliary Prediction Visualization Protocol

For Figure 10, we use the final 25,000-step EMA checkpoint of our Wan2.2-5B FCP model and visualize predictions on real videos. The main generator processes clean ground-truth history with its corresponding text prompt, and we extract the hidden states $\mathbf{h}_k$ of the final source chunk. Conditioned on $\mathbf{h}_k$, each offset-specific auxiliary branch independently samples a future chunk from random noise using the same UniPC settings as standard inference: 50 denoising steps, a classifier-free guidance scale of 3.0, and a timestep shift of 5.0. The branches receive neither intervening ground-truth or generated chunks nor a separate text input. For display, we sample decoded frames uniformly and align them temporally with the ground-truth frames. The auxiliary branches are retained only for this analysis and are otherwise removed during standard inference.

B More Results

B.1 Detailed Component Ablations

Table 6 extends the condensed component ablations in Table 4 by reporting the complete breakdown of Physics-IQ Verified metrics, providing a more comprehensive comparison of different component choices.

B.2 Feature Conditioning

Table 7 compares different strategies for injecting source features into the auxiliary predictors. Concatenating source features along the channel dimension followed by linear fusion is a common and straightforward implementation. Even with this simple design, FCP improves the IQ score from 27.6 to 29.5 over teacher forcing. Motivated by prior work demonstrating the effectiveness of AdaLN for feature injection (Wang et al., 2026b; Pan et al., 2026), we apply token-wise AdaLN throughout the auxiliary DiT blocks, further increasing the IQ score to 30.9. We hypothesize that this improvement arises from its layer-wise modulation and learned scaling factors, which provide fine-grained conditioning pathways and strengthen gradient propagation through the source representations into the generator. Notably, directly concatenating source and future tokens within self-attention is less

Table 6: Ablation study of the main components on Physics-IQ Verified (Rädsch et al., 2026). All variants are implemented based on Wan2.1 1.3B. The shaded row denotes our default configuration.

clean source	noisy source	text condition	chained connection	IQ ↑	SP ↑	ST ↑	WS ↑	MSE ↑
				27.6	32.8	25.5	20.4	31.6
✓				28.8	33.0	29.5	21.0	31.9
	✓			29.9	34.7	28.2	21.6	35.0
✓	✓			30.9	36.2	29.6	23.2	34.5
✓	✓	✓		29.8	34.8	28.4	22.0	34.2
✓	✓		✓	30.1	34.3	29.8	21.7	34.4

effective, reaching an IQ score of 28.8. We hypothesize that joint self-attention weakens the fine-grained correspondence between source and future representations.

Table 7: Ablation study of feature-fusion strategies for injecting features into the future-prediction branch on Physics-IQ Verified (Rädsch et al., 2026). All variants are implemented based on Wan2.1 1.3B. The shaded row denotes our default configuration.

Fusion strategy	IQ ↑	SP ↑	ST ↑	WS ↑	MSE ↑
Token Concat	28.8	34.1	28.1	21.0	32.0
Dim Concat	29.5	32.2	31.2	20.6	34.1
AdaLN	30.9	36.2	29.6	23.2	34.5

B.3 Representation Space Visualization

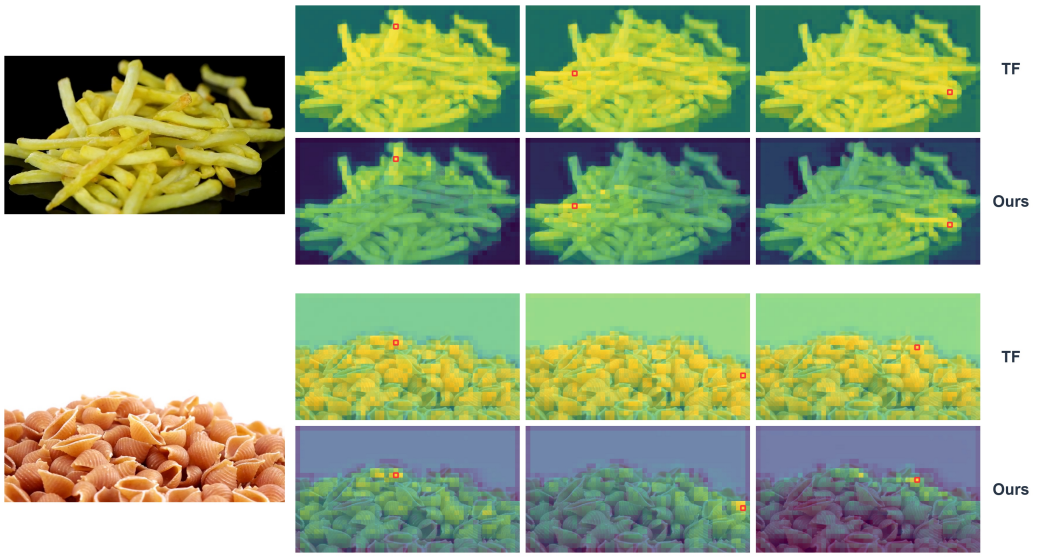


Figure 7: **Additional Spatial Feature Similarity Maps.** Each map shows within-frame similarity to a selected query token; two further examples continue in the next panel. Within each example, the top and bottom rows correspond to teacher forcing and FCP, respectively.

Figure 7 supplements Figure 5 with four additional scenes. Across these examples, FCP-trained representations tend to produce more selective similarity responses around the queried content than

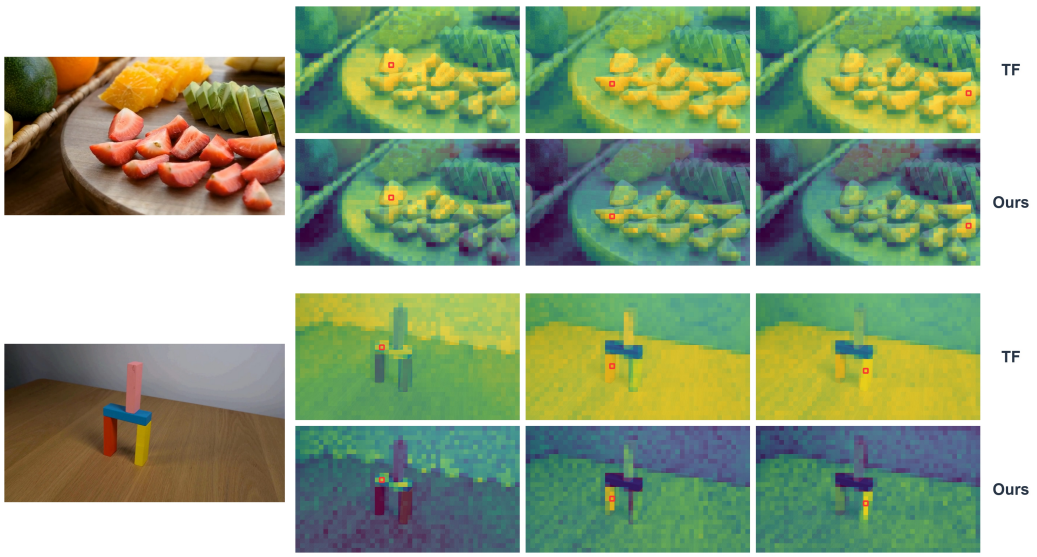


Figure 7: **Additional Spatial Feature Similarity Maps (continued).** Two further examples using the same query-to-token similarity visualization.

the teacher-forced baseline, providing further qualitative support for the spatial structure discussed in Section 5.6.

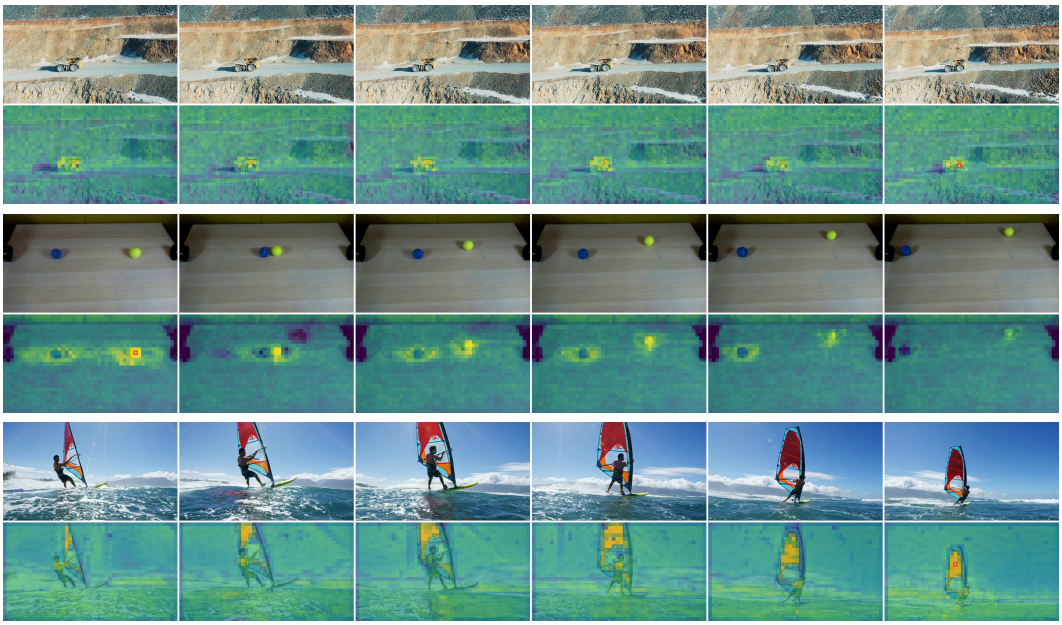


Figure 8: **Additional Spatiotemporal Feature Similarity Maps.** In each example, the top row shows video frames, while the bottom row visualizes similarity to a selected query token across space and time.

Complementing the temporal correspondence discussed in Section 5.6, Figure 8 shows three additional scenes in which high-similarity regions remain aligned with corresponding visual content across frames.

Figure 9 extends Figure 8 to a more challenging small-object setting. FCP maintains localized similarity responses around the queried vehicle across frames, whereas teacher forcing produces

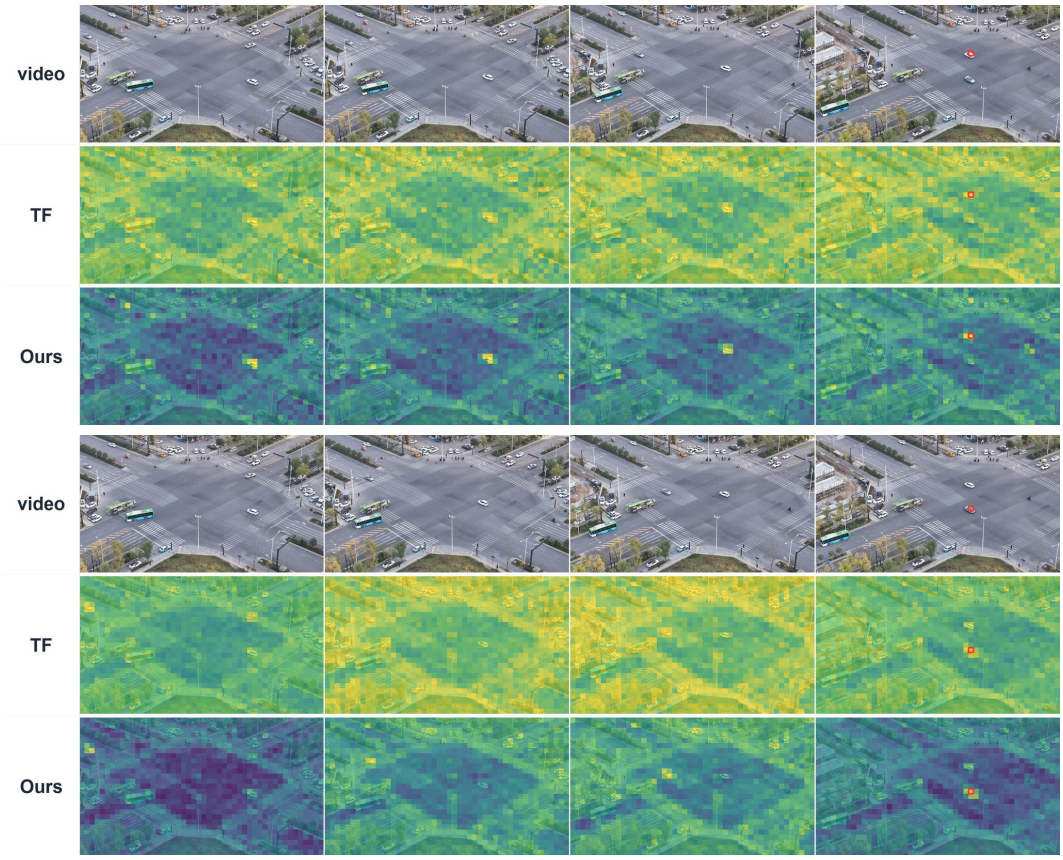


Figure 9: **Detailed Spatiotemporal Feature Similarity Maps.** In each example, the top row shows video frames, while the middle and bottom rows show similarity to a selected query token under teacher forcing and FCP, respectively.

more diffuse responses that are less able to distinguish it from the background and other objects. This contrast suggests that FCP learns representations with stronger spatiotemporal discrimination and more meaningful semantic structure.

B.4 Auxiliary Future Prediction Visualizations

Figure 10 visualizes future chunks sampled independently by the auxiliary branches from the same source representation. Across diverse scenes, the predictions capture broad visual changes without access to the intervening chunks. Visual fidelity tends to decline at farther offsets, yet the predictions often retain motion trends consistent with the ground truth. These examples suggest that $\mathbf{h}_k$ contains information useful for anticipating future evolution, without implying exact reconstruction of the future frames.

C Limitations and Discussions

C.1 Comparison with Next Forcing

As discussed in Section 2.2, Next Forcing (Xu et al., 2026) explores a closely related future-prediction idea. Importantly, our work was developed independently of and concurrently with Next Forcing (Xu et al., 2026). The initial FCP prototype design was completed before the Next Forcing paper became publicly available. While both works explore future prediction for autoregressive video generation, we target different tasks and adopt distinct methodological and experimental designs. First, Next Forcing is primarily framed around world-action modeling and robotic action

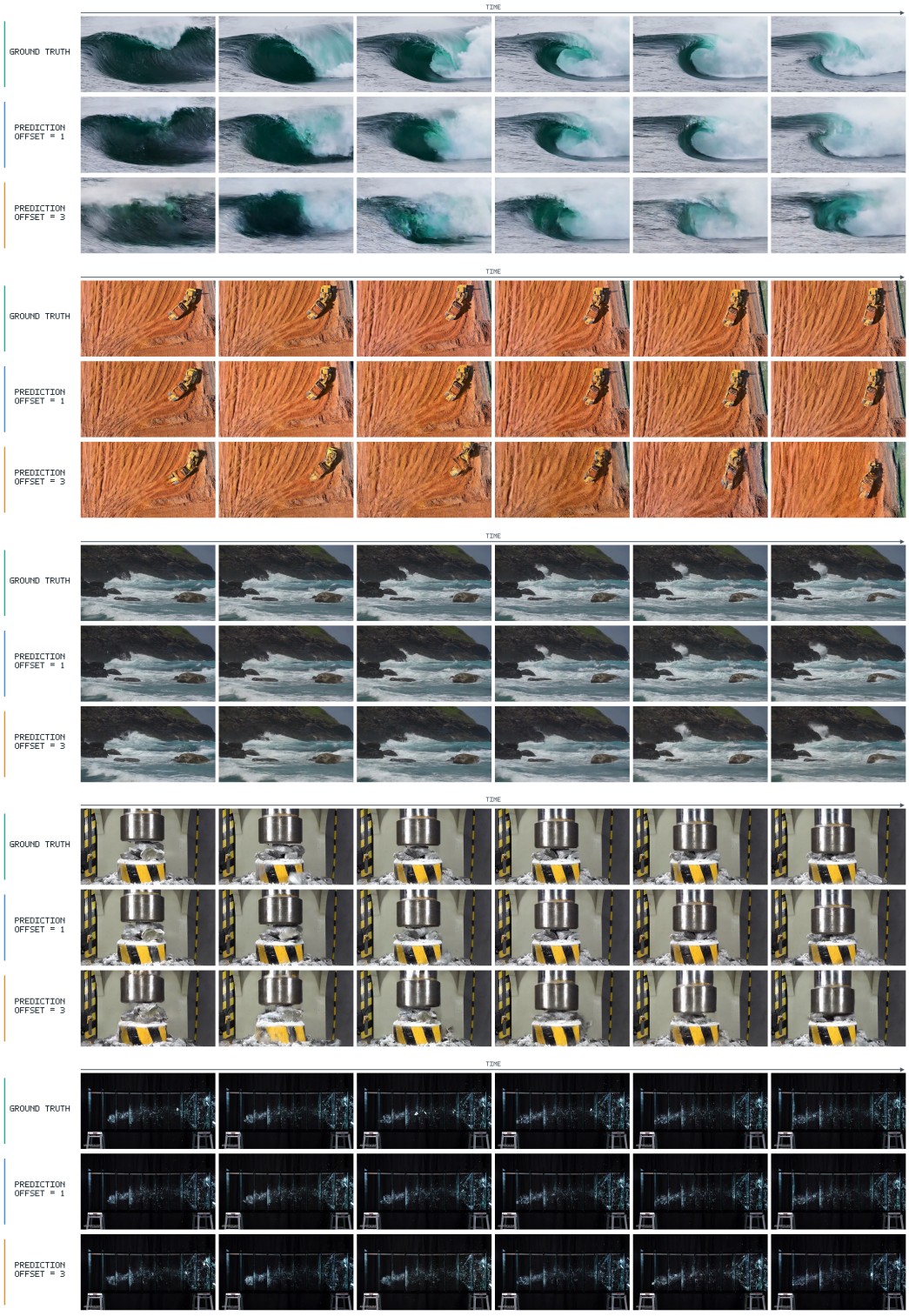


Figure 10: **Auxiliary Future Predictions.** In each example, the top row shows the ground-truth video, while the middle and bottom rows show outputs from auxiliary branches at the future offsets indicated on the left.

learning, whereas we focus on autoregressive video generation as a pathway toward generative world modeling.

Conditioning Context and Efficiency. Second, Next Forcing conditions its auxiliary predictors on the full visible history $\mathbf{h}_{\leq k}$, whereas ours use only $\mathbf{h}_k$. As discussed in Section 4.1, this choice encourages the source-chunk representation itself to compactly encode information predictive of future evolution. Beyond this design rationale, we test the context choice with a full-context FCP variant shown in Figure 4. At matched training steps, the two FCP variants achieve comparable IQ scores, but ours requires roughly 40–50% less training compute to attain the same performance. Our source-chunk-only design thus offers a more efficient and scalable form of future-prediction supervision, an important consideration when training larger video generators.

Parallel versus Chained Prediction. Third, Next Forcing adopts a chained prediction architecture inspired by language-model MTP (DeepSeek-AI, 2024), whereas our offset-specific predictors operate in parallel from $\mathbf{h}_k$. Unlike discrete-token prediction, diffusion-based video prediction gives each auxiliary predictor a partially noised future latent. Chaining may therefore pass information from an intermediate future chunk to later predictors along the red path in Figure 2a, weakening the intended predictive supervision on $\mathbf{h}_k$.

Consistent with this concern, the chained variant retains a gain at 1.3B with three offsets, but offers only a marginal I2V gain and no V2V gain at 5B with five offsets. This decline may reflect both the greater capacity of the model to exploit such shortcuts and the longer chain through which intermediate information can propagate. Parallel prediction, by contrast, remains effective in both settings (Tables 4 and 2), supporting its use for visual future prediction.

Other Differences. Our different task emphasis also shapes several design and evaluation choices. We supervise both clean-context and noisy-target representations, and use token-wise AdaLN to inject source features into the auxiliary predictors, drawing on prior experience in visual generation (Wang et al., 2026b; Pan et al., 2026). Our evaluation focuses on the physical plausibility of generated videos, reflecting our interest in visual evolution rather than action-planning performance.

C.2 Potential Inference-Time Applications

Figure 10 suggests that the auxiliary predictors can anticipate future dynamics, although their outputs do not yet reach the visual fidelity expected of final video generation. Replacing the main generator with these branches throughout denoising may therefore compromise output quality. One possibility is to use them during early, high-noise denoising before handing the sample to the main generator for refinement, potentially reducing sampling cost. Alternatively, combining their predictions with those of the main generator may improve the reliability of the resulting future evolution. These inference-time applications remain untested and are left for future work.

C.3 From-Scratch Pretraining and Further Scaling

Due to computational constraints, our experiments do not cover larger model scales or training on substantially longer videos. We therefore leave broader scaling and evaluation across longer temporal horizons to future work.

All current models are initialized from pretrained bidirectional video diffusion checkpoints rather than pretrained from scratch. In our experiments, adapting these checkpoints for autoregressive training led to early-stage performance degradation. We suspect that bidirectional and autoregressive models may organize their representations differently, so this initialization could limit the extent to which FCP's benefits emerge. We aim to investigate large-scale from-scratch pretraining with FCP as a native objective for autoregressive world models, together with more diverse, higher-quality data tailored to this setting.