PROAR: LEARNING PROSPECTIVE REASONING WITH AUTOREGRESSIVE VIDEO MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Autoregressive (AR) video models excel at causal generation, but their reliance on next-chunk prediction confines them to a short-sighted, reactive paradigm. This limitation is particularly consequential for reasoning-oriented generation, where achieving a target outcome through valid intermediate states matters more than local visual plausibility. To address this challenge, we propose Learning **Pro**spective Reasoning with **A**uto**r**egressive Video Models **(ProAR)**, a novel framework that transforms autoregressive video generation into a goal-oriented reasoning process. ProAR introduces two key components: (1) To anchor generation to the long-range *outcome*, we integrate goal-frame prediction into the autoregressive loop via an asymmetric attention mask, enabling the predicted goal frame to guide the generation of intermediate states without being disrupted by them. (2) To guide short-range *transition*, we introduce future representation self-alignment to encourage current hidden states to anticipate upcoming temporal dynamics. By leveraging teacher-forcing in AR training, we extract clean future representations in a single forward pass and align current representations with them using a lightweight, training-only predictor. Together, these two mechanisms seamlessly combine explicit, sparse target supervision with implicit, dense step-wise guidance, promoting coherent, goal-directed reasoning progress with modest computational cost. Experiments show that ProAR's complementary components consistently improve performance across diverse visual reasoning benchmarks. The framework proves highly training-efficient, surpassing fully trained standard AR baselines using only 25% of the training steps. This paradigm also demonstrates promising applicability to embodied reasoning tasks.

1 INTRODUCTION

Video generation models are moving beyond visual synthesis toward reasoning through visual dynamics. They have been used to solve game-like problems such as maze navigation and visual puzzles (Wang et al., 2026a; Xu et al., 2026b; Zhu et al., 2026b), simulate physical processes (Yao et al., 2026), and support embodied control (Bruce et al., 2024; Zhu et al., 2026c; He et al., 2025; Wang et al., 2026c; Ye et al., 2026; Li et al., 2026). In these settings, the generated video is itself the reasoning process: the model must achieve a target outcome by generating a sequence of valid visual states. This paradigm requires video models to reason about and plan the evolution of visual states, rather than merely produce visually plausible content, opening new possibilities for physical and embodied intelligence (Kataoka et al., 2026). We refer to this emerging paradigm as ***video-native reasoning***, where a model reasons by directly generating how the visual world evolves.

Autoregressive (AR) video models provide a natural foundation for video-native reasoning. Unlike bidirectional video models, AR models generate video sequentially, conditioning each new chunk on the preceding visual history. This causal generation process supports long-horizon generation and low-latency streaming (Teng et al., 2025; Yin et al., 2025; Yang et al., 2025a; Huang et al., 2026; Zhu et al., 2026a), making AR models particularly suitable for real-time and interactive applications (Teng et al., 2025; Yin et al., 2025; Yang et al., 2025a; Huang et al., 2026; Zhu et al., 2026a).

However, there is a fundamental mismatch between the standard AR training objective and reasoning-oriented generation. Next-chunk prediction optimizes only for locally plausible continuations of the observed history without accounting for their impact on eventual task completion, influencing all subsequent predictions. Consequently, short-sighted decisions compound over time,

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

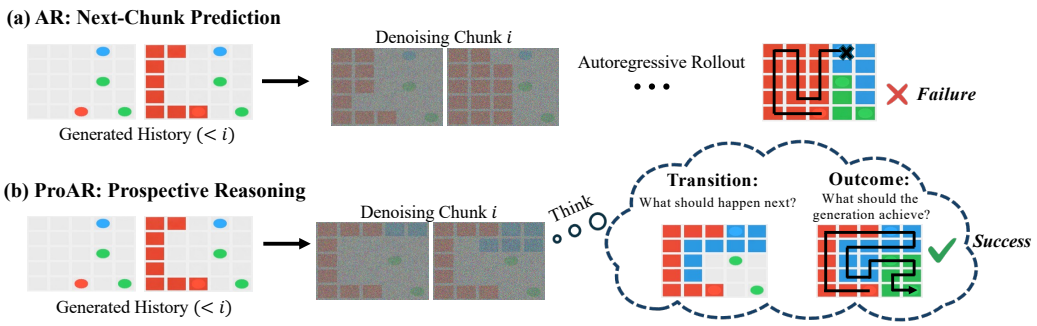


Figure 1: Overview of ProAR. (a) Standard AR relies on local next-chunk prediction, leading to trajectory drift and task failure. (b) ProAR introduces prospective reasoning to anticipate immediate transitions and final outcomes, guiding generation toward task success.

easily steering a locally plausible trajectory away from the intended goal (Fig. 1(a)). Existing methods primarily focus on improving visual quality, temporal consistency, or prediction horizon (Yu et al., 2026; Xu et al., 2026a; Qian et al., 2026; Liu et al., 2026; Luo et al., 2026), rather than explicitly addressing the short-sightedness of AR generation in goal-directed reasoning tasks.

We argue that autoregressive video-native reasoning requires a prospective view of generation. We define ***prospective reasoning*** as the capacity to anticipate possible future outcomes and state transitions based on observed history, and to use this foresight to guide ongoing generation. This perspective exposes two complementary requirements (Fig. 1(b)). At the long-range level, the model should anticipate its target outcome. At the short-range level, it should anticipate upcoming transitions that drive meaningful progress. Long-range outcome awareness provides global direction, while short-range transition awareness encourages coherent and steady progress toward the goal. Together, they enable future expectations to guide present generation, rather than solely reacting to the past.

Motivated by these requirements, we propose Learning **Pro**spective Reasoning with **A**uto**r**egressive Video Models (**ProAR**). ProAR introduces two complementary mechanisms: (1) *Outcome guidance* provides explicit but sparse long-range supervision. At each autoregressive step, ProAR jointly predicts the current chunk and a goal frame representing the anticipated final state. An asymmetric attention mask prevents the current chunk from influencing the goal belief, while allowing the predicted goal to guide current generation. As the video unfolds, the model continuously updates its goal belief based on the growing history. (2) *Transition guidance* provides implicit and dense short-range supervision through future representation self-alignment. Leveraging teacher forcing, ProAR reuses the clean next-chunk representation computed within the same forward pass to supervise the current states to anticipate upcoming temporal dynamics. This design requires neither an external representation encoder nor an additional backbone forward pass, and the lightweight alignment predictor is removed entirely during inference. At inference, ProAR generates autoregressively without access to ground-truth future states, relying on its dynamically updated goal-frame predictions and learned anticipatory representations to guide generation.

We evaluate ProAR on 13 visual reasoning tasks, including a 10-task subset of VBVR (Wang et al., 2026a) covering perceptual and spatial reasoning, and 3 abstract reasoning tasks from VideoRLVR (Zhu et al., 2026b). Compared with standard AR generation, ProAR raises the mean score on the 10-task VBVR subset from 0.664 to 0.801, and the average success rate on VideoRLVR from 50.97 to 52.97. Ablation studies further demonstrate that Outcome and Transition guidance provide complementary gains across these tasks. ProAR is also highly training-efficient: on the 10-task VBVR subset, it surpasses the fully trained AR baseline using only 25% of the training steps. Moreover, ProAR demonstrates promising generalization to embodied reasoning, delivering consistent improvements on the WorldArena simulation benchmark (Shang et al., 2026).

Our contributions are threefold: (1) We revisit autoregressive video generation from a reasoning perspective, identifying the lack of prospective guidance as a key limitation of next-chunk prediction. (2) We propose ProAR, which combines goal belief and future representation self-alignment to provide *outcome* and *transition* guidance with minimal overhead. (3) Experiments across diverse video reasoning tasks demonstrate the complementary benefits of outcome and transition, driving consistent improvements in both reasoning performance and training efficiency, with promising generalization to embodied reasoning.

2 Related Work

Autoregressive Video Generation. Autoregressive video models generate frames or chunks causally, making them suitable for long, streaming, and interactive generation. Recent methods improve scalability (Teng et al., 2025), efficiency (Yang et al., 2025a; Chen et al., 2026c), and training stability (Yin et al., 2025; Huang et al., 2026; Cui et al., 2026; Zhu et al., 2026a) during autoregressive rollout. Several studies have explored richer future information beyond next-chunk prediction through foresight alignment (Yu et al., 2026), hierarchical denoising (Qian et al., 2026), and future-aware training (Xu et al., 2026a; Liu et al., 2026). These advances have also enabled interactive world models and world action models for embodied agents (Bruce et al., 2024; Zhu et al., 2026c; Wang et al., 2026c; Ye et al., 2026; Li et al., 2026). However, extending the prediction horizon alone can improve visual quality, but is insufficient for task-oriented generation which requires both a clear target and effective intermediate transitions. We address these needs with complementary outcome and transition guidance.

Native Reasoning with Video Generation Models. Recent studies have begun to treat video generation as a paradigm for reasoning rather than only visual synthesis (Tong et al., 2026; Wiedemer et al., 2025; He et al., 2026; Qi et al., 2026). Video generators can perform visual reasoning by generating intermediate solution steps, such as navigating a maze or moving pieces to solve a puzzle (Wang et al., 2026a; Xu et al., 2026b). Existing approaches improve reasoning capabilities through visual prompting (Geirhos et al., 2026), test-time training (Cheng et al., 2026; Jun et al., 2026), explicit chains of visual thought (Huang et al., 2025), and reinforcement learning (Zhu et al., 2026b; Xu et al., 2026b). Other studies further investigate how reasoning capabilities emerge in video generation models (Wang et al., 2026b; Chen et al., 2026a). However, most of these methods build on bidirectional generators that jointly denoise all frames and allow dense interactions across time. Autoregressive models instead strictly enforce causal generation, committing to each chunk without seeing future frames or revising past errors. The closest work in this setting (Qian et al., 2026) introduces hierarchical coarse-to-fine denoising for autoregressive video reasoning. In contrast, ProAR provides lightweight outcome and transition guidance without changing the basic chunk-wise generation process.

Representation Learning in Diffusion Models. Diffusion models learn rich semantic and structural features in their intermediate layers (Tang et al., 2023; Luo et al., 2023). One line of work aligns generative representations with strong pretrained feature spaces (Oquab et al., 2023; He et al., 2022; Assran et al., 2023) to improve training efficiency and generation quality (Yu et al., 2024; Tian et al., 2026; Leng et al., 2025). Another line derives supervision from the generative model itself, using stronger internal representations to guide weaker ones (Jiang et al., 2025; Haghighi et al., 2026; Chefer et al., 2026). Recently, representation alignment has also been explored in video post-training (Zhang et al., 2026; Yu et al., 2026). Most relatedly, Video-Mirai (Yu et al., 2026) distills future bidirectional representations into autoregressive causal features but requires an additional forward pass. In contrast, our Transition module exploit autoregressive training properties to obtain clean future representations within the model, requiring neither an external representation encoder nor an additional model pass.

3 ProAR

In this section, we present ProAR, a novel framework for learning prospective reasoning in autoregressive video generation. We first investigate why local next-chunk supervision is insufficient to ensure goal completion and valid progression (Sec. 3.1). ProAR addresses this gap through two complementary mechanisms. *Outcome guidance* provides explicit, sparse supervision by anchoring generation with an evolving goal belief (Sec. 3.2), while *Transition guidance* provides implicit, dense supervision by encouraging every current representation to anticipate the next stage (Sec. 3.3).

3.1 Preliminaries and Motivation

Autoregressive Video Diffusion. In this paper, we consider autoregressive (AR) video diffusion models. Let $\mathbf{z}^{1:N} = \{\mathbf{z}^1, \dots, \mathbf{z}^N\}$ denote a video divided into N latent chunks, where $\mathbf{z}^{<i}$ denotes all chunks preceding $\mathbf{z}^i$. Given a generation condition $\mathbf{c}$, such as an input image and a text prompt, the joint distribution is factorized as:

$$p_\theta(\mathbf{z}^{1:N} \mid \mathbf{c}) = \prod_{i=1}^{N} p_\theta(\mathbf{z}^i \mid \mathbf{z}^{<i}, \mathbf{c}). \tag{1}$$

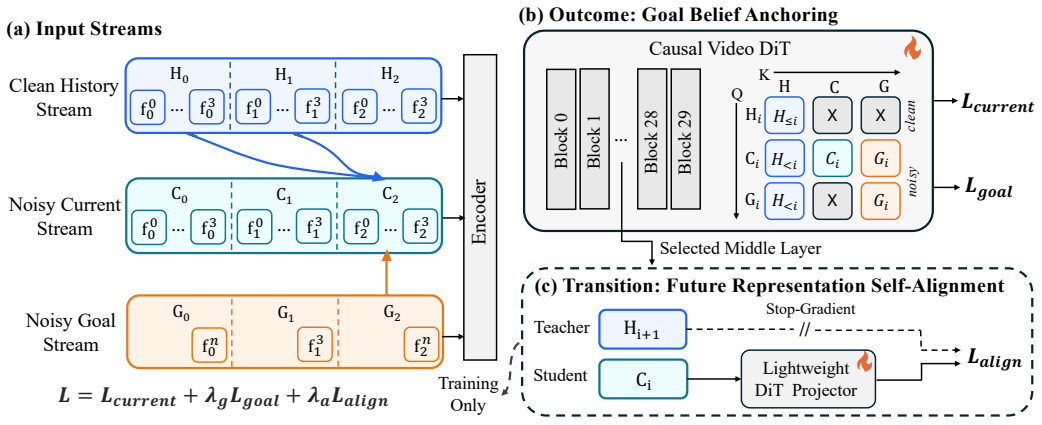


Figure 2: Overview of ProAR. (a) Input: For a given video, the model processes the history, current, and goal streams within a forward pass. (b) Outcome: The causal DiT jointly denoises the current chunk and goal frame, where an attention mask enables the clean history and evolving goal belief to guide current generation. (c) Transition: A training-only projector aligns the middle-layer hidden states of the noisy current chunk (student) with the clean next chunk (teacher).

Each conditional distribution is modeled with flow matching (Lipman et al., 2022). Given a clean target chunk $\mathbf{z}_0^i$, we construct its noisy state as $\mathbf{z}_t^i = (1-t)\mathbf{z}_0^i + t\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $t \in [0, 1]$. The velocity network $\mathbf{v}_\theta$ is trained with

$$\mathcal{L}_{\text{AR}} = \mathbb{E}_{i,t,\boldsymbol{\epsilon}} \left[\left\| \mathbf{v}_\theta \left(\mathbf{z}_t^i, t \mid \mathbf{z}_0^{<i}, \mathbf{c} \right) - \left(\boldsymbol{\epsilon} - \mathbf{z}_0^i \right) \right\|_2^2 \right]. \tag{2}$$

We adopt teacher-forcing during training. All target chunks are denoised in parallel within a single forward pass, while a causal attention mask ensures each noisy chunk is conditioned only on its clean ground-truth history. During inference, the model generates chunks sequentially and appends each to the history for subsequent prediction. Thus, training provides chunk-level denoising supervision, whereas inference proceeds as an irreversible rollout over model-generated history.

From Next-Chunk Prediction to Prospective Reasoning. Video-native reasoning requires each generated video to complete a specified task. This imposes two requirements beyond visual plausibility: (1) *goal completion*, where the final state must satisfy the task objective, and (2) *progression validity*, where intermediate transitions must follow task logic or physical constraints.

However, the local objective $\mathcal{L}_{\text{AR}}$ (Eq. 2) supervises only the current chunk given its preceding history. It does not directly supervise whether the prediction supports the final objective or enables valid subsequent transitions. During inference, every generated chunk is appended to the history and cannot be revised, allowing early errors to persist and compound throughout the autoregressive rollout. As illustrated in Fig. 1(a), a locally plausible rollout may therefore culminate in task failure.

This gap motivates *prospective reasoning* in Fig. 1(b): anticipating state evolution and future outcomes from available observations, then using these expectations to guide current generation. It requires guidance at two temporal scales. Long-range guidance maintains an expectation of the final outcome, while short-range guidance anticipates the next stage of progression. These requirements motivate the outcome and transition mechanisms introduced below.

3.2 Outcome: Goal Belief for Long-Horizon Planning

Outcome guidance addresses *goal completion* by answering ***What should the generation achieve?*** It introduces a *goal belief* that explicitly represents the anticipated final state and evolves with the generated history. By jointly predicting the current chunk and goal belief at every AR step, the model can guide local generation toward the expected long-range outcome.

Let $\mathbf{g}^0$ denote the clean ground-truth goal frame (final frame). At autoregressive step i, the model jointly denoises the noisy current chunk $\mathbf{z}_i^t$ and the noisy goal input $\mathbf{g}_i^t$, conditioned on the clean history $\mathbf{z}_{<i}^0$ and generation condition $\mathbf{c}$. Both inputs are constructed at the same flow-matching

timestep t:

$$\mathbf{z}_i^t = (1-t)\mathbf{z}_i^0 + t\boldsymbol{\epsilon}_{z,i}, \qquad \mathbf{g}_i^t = (1-t)\mathbf{g}^0 + t\boldsymbol{\epsilon}_{g,i}, \tag{3}$$

where $\boldsymbol{\epsilon}_{z,i}, \boldsymbol{\epsilon}_{g,i} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ are sampled independently. A shared backbone predicts the velocity fields for both streams in a single forward pass. The current stream is supervised by the AR objective $\mathcal{L}_{\mathrm{AR}}$ in Eq. 2, while the goal stream is supervised by

$$\mathcal{L}_{\mathrm{goal}} = \mathbb{E}_{i,t,\boldsymbol{\epsilon}_{g,i}} \left[\left\| \mathbf{v}_\theta^g \left(\mathbf{g}_i^t, t \mid \mathbf{z}_{<i}^0, \mathbf{c} \right) - \left(\boldsymbol{\epsilon}_{g,i} - \mathbf{g}^0 \right) \right\|_2^2 \right], \tag{4}$$

where $\mathbf{v}_\theta^g$ denotes the velocity predicted for the goal stream.

Joint prediction alone cannot guarantee the desired information flow. We therefore introduce an asymmetric attention mask (Fig. 2b). Specifically, goal tokens attend only to the clean history and themselves, shielded from the noisy current chunk. Conversely, current tokens additionally attend to these goal tokens. This target-first dependency prevents the uncertain current state from corrupting the goal belief, while allowing the anticipated goal to guide ongoing generation, modulated by the loss weight λ_{goal} in Eq. 7.

During training, the goal branch at every AR step is supervised by the same ground-truth final frame $\mathbf{g}^0$. During inference, the model jointly denoises the current chunk and goal input, but only the generated current chunk is appended to the history. The goal belief is re-estimated at the next step from the expanded history, allowing it to be progressively refined as the video unfolds.

3.3 Transition: Representation Alignment for Progressive Reasoning

Outcome guidance specifies where the generation should end, but leaves how the current state should evolve unconstrained. Transition guidance addresses *progression validity* by answering ***What should happen next?*** Our key intuition is that the current representation should anticipate the next stage of the task. We realize this intuition through *future representation self-alignment*.

Teacher and Student. Recent studies show that intermediate hidden states of diffusion models encode rich semantic and temporal abstractions, and that stronger hidden states can guide weaker ones to improve generation (Jiang et al., 2025; Haghighi et al., 2026). Building on this insight, we construct a self-supervised Teacher–Student mechanism within the backbone feature space.

Our core idea is to use the clean representation of the next chunk to guide the noisy representation of the current chunk, as illustrated in Fig. 2(c). The clean next-chunk hidden states contain uncontaminated information about the next stage and therefore serve as a strong, prospective *teacher*. In contrast, the noisy current hidden states contain incomplete and noise-corrupted information, serving as the weaker *student*. Thus, the teacher provides not only a cleaner representation but also a short-range temporal direction for the current state.

Formally, at a selected backbone layer ℓ, let $\mathbf{H}_{i+1}^\ell$ and $\mathbf{C}_i^{t,\ell}$ denote the hidden states of the clean next chunk $\mathbf{z}_{i+1}^0$ and the noisy current chunk $\mathbf{z}_i^t$, respectively (simplified as $\mathbf{H}_{i+1}$ and $\mathbf{C}_i$ in Fig. 2c). For each valid current chunk ($i < N$), a lightweight predictor $\mathcal{P}_\phi$ maps the student toward the fixed teacher representation:

$$\overline{\mathbf{H}}_{i+1}^\ell = \mathrm{sg}\left(\mathbf{H}_{i+1}^\ell\right), \qquad \widehat{\mathbf{H}}_{i+1}^{t,\ell} = \mathcal{P}_\phi\left(\mathbf{C}_i^{t,\ell}\right), \tag{5}$$

where sg denotes the stop-gradient operator. The predictor (default 3 transformer blocks) learns to transform the student features to anticipate future dynamics, while preserving the backbone's token grid and dimension. This ensures direct spatial alignment without dimensional compression.

Single-Pass Self-Supervision. A key advantage of our design is that both Teacher and Student hidden states are obtained from the same backbone forward pass. During teacher-forcing training, all noisy current chunks are denoised in parallel, each attending only to its clean ground-truth history (Fig. 2(a)). Consequently, the clean next chunk $\mathbf{z}_{i+1}^0$ is inherently encoded in the history stream to condition later denoising targets, while the noisy current chunk $\mathbf{z}_i^t$ is processed simultaneously in the current stream. We thus obtain their hidden states $\mathbf{H}_{i+1}^\ell$ and $\mathbf{C}_i^{t,\ell}$ entirely for free, eliminating the overhead of an external representation encoder or an additional backbone pass.

Motivated by evidence that reasoning capabilities primarily localize in intermediate DiT blocks (Wang et al., 2026b), we extract these representations from a shared middle layer. We empirically validate this structural choice through layer-wise ablations in Appendix A.2.



Figure 3: Qualitative Results. Across tasks from VBVR, VideoRLVR, and WorldArena, ProAR consistently improves reasoning performance by promoting goal completion and valid progression.

Representation Alignment. We align the student's prediction $\widehat{\mathbf{H}}_{i+1}^{t,\ell}$ with the teacher target $\overline{\mathbf{H}}_{i+1}^{\ell}$ using a token-wise cosine distance loss. Given M tokens per chunk, the alignment objective is:

$$\mathcal{L}_{\text{align}} = \mathbb{E}_{i<N,t}\left[\frac{1}{M}\sum_{m=1}^{M}\left(1-\cos\left(\widehat{\mathbf{H}}_{i+1,m}^{t,\ell}, \overline{\mathbf{H}}_{i+1,m}^{\ell}\right)\right)\right]. \tag{6}$$

The stop-gradient operation keeps the teacher fixed, while gradients propagate through the predictor $\mathcal{P}_\phi$ and the student branch of the backbone. This directional optimization encourages the current hidden states to anticipate upcoming temporal dynamics.

Overall Objective and Inference. The complete training objective combines current generation, goal prediction (Sec.3.2), and future representation alignment (Sec.3.3):

$$\mathcal{L} = \mathcal{L}_{\text{current}} + \lambda_{\text{goal}}\mathcal{L}_{\text{goal}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}}. \tag{7}$$

At inference, the alignment predictor $\mathcal{P}_\phi$ is removed entirely, so Transition guidance adds no inference cost. Outcome guidance only requires jointly denoising 1 auxiliary goal frame with each current chunk in the same forward pass. Detailed computational costs are provided in Appendix A.1.

4 Experiments

4.1 Experimental Setup

Benchmarks. We evaluate ProAR on three complementary video-native reasoning benchmarks.

(1) VBVR. VBVR (Wang et al., 2026a) contains 100 tasks, each with 10K training examples. We select 10 tasks that primarily assess perceptual and spatial reasoning: Concentric Ring (G194), Outline Match (G25), Stable Sort (G3), Star Match (G5), Domino Chain (O23), Ball Eater (O31), Grid Shift (O36), Pipe Puzzle (O44), Slide Puzzle (O47), and Balancing Vessels (O75). Since the official benchmark provides only five test examples per task, we use its official data generator to expand the test set to 50 examples per task, yielding 500 test examples in total.

Table 1: Performance comparison across 10 selected tasks in VBVR. The best result for each metric is shown in **bold**, and the second-best result is underlined.

Model	Concentric Ring	Outline Match	Stable Sort	Star Match	Domino Chain	Ball Eater	Grid Shift	Pipe Puzzle	Slide Puzzle	Balancing Vessels	Mean
Closed-Source Models											
Seedance 2.0	0.653	0.658	0.693	0.598	0.547	0.696	0.482	0.530	0.799	0.616	0.627
Bidirectional Models											
Wan2.2 5B	0.241	0.480	0.400	0.312	0.419	0.349	0.000	0.256	0.034	0.362	0.285
Wan2.2 5B SFT	0.798	0.702	**0.905**	0.827	0.521	0.779	0.585	0.514	0.860	0.872	0.736
Autoregressive Models											
Standard AR	0.794	0.753	0.565	0.689	**0.658**	0.754	0.720	0.424	0.388	0.880	0.663
w/ Outcome	0.940	0.755	0.890	0.858	0.639	**0.784**	0.726	0.486	0.908	0.869	0.786
w/ Transition	0.860	0.758	0.699	0.659	0.646	0.765	0.651	0.348	0.558	**0.887**	0.683
ProAR (ours)	**0.949**	**0.777**	0.895	**0.871**	0.628	0.781	**0.770**	**0.546**	**0.931**	0.863	**0.801**

Table 2: Ablation study on the Transition training start point. (* denotes default.)

Start Point	Concentric Ring	Outline Match	Stable Sort	Star Match	Domino Chain	Ball Eater	Grid Shift	Pipe Puzzle	Slide Puzzle	Balancing Vessels	Mean
Step 0	0.909	0.771	0.909	0.850	0.566	**0.784**	**0.782**	**0.591**	**0.978**	0.857	0.799
Step 5000	0.913	0.755	**0.923**	0.870	0.610	0.766	0.741	0.423	0.953	**0.864**	0.782
Step 7500*	**0.949**	**0.777**	0.895	**0.871**	**0.628**	0.781	0.770	0.546	0.931	0.863	**0.801**

(2) VideoRLVR. VideoRLVR (Zhu et al., 2026b) evaluates abstract reasoning through 3 games: Maze, FlowFree, and Sokoban. Each task contains 10K training and 1K test samples. Together, VBVR and VideoRLVR cover 13 tasks spanning perceptual, spatial, and abstract reasoning.

(3) WorldArena. As an extension to embodied reasoning, we evaluate ProAR on WorldArena (Shang et al., 2026), following its official protocol on 50 bimanual manipulation tasks from RoboTwin 2.0 (Chen et al., 2025). Each task provides 50 training and 20 separately generated test samples. This setting tests whether ProAR can generate actionable sequences as an embodied video simulator.

Baselines. For each benchmark, we evaluate both bidirectional and autoregressive video generation.

(1) Bidirectional Models: On the VBVR, we evaluate both the pretrained Wan2.2-TI2V-5B model and its SFT variant trained following the official protocol (Wang et al., 2026a). On VideoRLVR, we use the reported Wan2.2-TI2V-5B results after supervised fine-tuning and reinforcement learning.

(2) Autoregressive Models: We adapt the open-source Wan2.2-TI2V-5B (Wan et al., 2025) backbone to autoregressive generation using causal attention and teacher-forcing training. This Standard AR baseline is trained and evaluated under the same settings as ProAR.

Implementation Details. We optimize all models using AdamW with a learning rate of 5×10^{-6} and a global batch size of 16. The chunk size is 4 frames for VBVR and VideoRLVR, and 8 frames for WorldArena. We conduct every training on 2-4 NVIDIA B200 GPUs.

For *Outcome Guidance*, the final ground-truth frame serves as the goal target. For *Transition Guidance*, we extract both student (noisy current chunk) and teacher (clean next chunk) hidden states from backbone block 15 in a single forward pass. A lightweight 3-block DiT predictor maps the Student to the Teacher representation and is fully removed at inference.

We set $\lambda_{\text{goal}} = 0.25$ and $\lambda_{\text{align}} = 0.01$ by default. Since the raw cosine alignment loss is much larger than the generation losses, the weighted alignment term contributes approximately 3%–10% of the total loss. Transition guidance is activated in the later training stages: at step 7,500/10,000 for VBVR and step 3,000/5,000 for VideoRLVR. See more details in Appendix A.1.

4.2 Main Results

Quantitative Results on VBVR. Table 1 summarizes performance across the selected 10 perceptual and spatial reasoning tasks. Consistent with the short-horizon limitation discussed in Sec. 3.1, the teacher-forced Standard AR baseline achieves a mean score of 0.663, below the 0.736 obtained by bidirectional Wan2.2 SFT. With prospective guidance, ProAR raises the mean score to 0.801,

Table 3: Performance comparison across 3 tasks in VideoRLVR. † are results reported from VideoRLVR (Zhu et al., 2026b).

Model	Maze				FlowFree				Sokoban			
	Prec.	Rec.	F1	SR	Prec.	Rec.	F1	SR	Prec.	Rec.	F1	SR
Bidirectional Models												
Wan2.2 SFT†	80.40	85.10	82.70	69.00	43.10	42.50	42.80	2.50	32.80	11.60	17.10	2.70
VideoRLVR†	82.10	86.90	84.40	72.20	44.30	43.80	44.00	7.90	34.00	12.50	29.40	6.10
Autoregressive Models												
Standard AR	89.28	91.86	90.16	79.50	55.22	54.91	55.05	29.40	40.18	34.74	34.62	44.00
w/ Outcome	93.01	**96.54**	**94.65**	82.30	57.83	55.16	56.10	23.20	43.63	32.73	35.59	44.30
w/ Transition	89.29	91.99	90.15	80.90	55.80	55.43	55.59	26.70	45.41	**34.93**	36.65	**49.60**
ProAR (ours)	**93.33**	95.88	94.48	**82.70**	**60.57**	**58.86**	**59.52**	**30.20**	**47.62**	32.56	**36.87**	46.00

a 20.8% relative improvement over Standard AR. It also exceeds the closed-source Seedance 2.0 and Wan2.2 SFT baselines by 0.174 and 0.065, respectively. At the task level, ProAR outperforms Standard AR on 8 of the 10 tasks, with particularly large absolute gains on Stable Sort (+0.330), Pipe Puzzle (+0.122), and Slide Puzzle (+0.543). The gains across tasks with different objectives indicate that the improvement is not driven by a single reasoning pattern. Overall, these results show that prospective guidance effectively reduces the mismatch between local chunk prediction and reasoning task completion.

Quantitative Results on VideoRLVR. Table 3 reports results on the 3 abstract reasoning games. Surprisingly, Standard AR already outperforms both bidirectional Wan2.2 SFT and its reinforcement-learning variant, suggesting that causal generation is well suited to structured tasks involving sequential state transitions. Building on this strong baseline, ProAR improves the average success rate from 50.97 to 52.97 and the average precision from 61.56 to 67.17. Precision increases on all 3 games, including a substantial gain from 40.18 to 47.62 on Sokoban. ProAR also improves the success rate over Standard AR on every game, reaching 82.70 on Maze, 30.20 on FlowFree, and 46.00 on Sokoban. These consistent gains demonstrate the effectiveness of prospective guidance across different abstract reasoning environments.

Qualitative Results. As illustrated in Fig. 3, ProAR overcomes the short-sighted failures of standard AR by consistently ensuring valid progression and successful goal completion. On VBVR, AR produces disorganized movements and incomplete shapes, whereas ProAR correctly aligns objects to form the intended patterns.. On VideoRLVR, standard AR generates fragmented or invalid paths, while ProAR produces continuous, rule-consistent trajectories that reach the target.

4.3 Training Efficiency

We next examine whether prospective guidance improves training efficiency in addition to final performance. Fig. 4 compares ProAR with the standard AR baseline at matched training checkpoints on VBVR 10-task subset. ProAR converges substantially faster. After only 2,500 steps, it reaches a mean score of 0.7079, already exceeding the 0.6635 obtained by the standard AR baseline after 10,000 steps. After Transition guidance is activated at step 7,500, the score further increases from 0.7666 to 0.8010 by step 10,000. Overall, these results show that prospective guidance improves both convergence speed and final reasoning performance without requiring additional optimization steps.

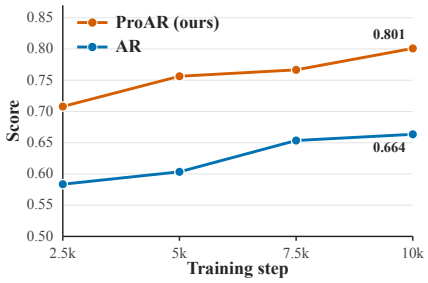


Figure 4: Performance curve on VBVR 10 tasks subsets.

4.4 Ablations and Analysis

Complementarity of Outcome and Transition. To isolate the contribution of each component, we evaluate ProAR with only outcome guidance (*w/ Outcome*) or transition guidance (*w/ Transition*) in Tables 1 and 3. On VBVR, outcome and transition individually improve the mean score of

Table 4: Extension study on WorldArena. Within the last two rows, the higher result for each metric is highlighted in orange.

Method	Image Qual.	JEPA Sim.	Flow Score	Back. Cons.	Photo. Cons.	Interact. Qual.	Traj. Acc.	Depth Acc.	Instr. Follow.	EWMScore
Bidirectional Models										
CogVideoX	36.23	94.84	21.77	88.38	33.79	66.96	34.79	91.09	75.58	58.79
Cosmos-Predict 2.5 (text)	**65.49**	46.69	17.55	73.98	35.39	65.14	11.60	70.71	64.70	53.06
Cosmos-Predict 2.5 (action)	49.48	90.65	9.95	72.72	45.21	60.46	27.49	90.35	60.54	54.29
Vidar	41.75	61.10	17.03	78.24	20.57	60.84	21.26	79.77	61.54	51.92
FlowWAM	48.87	**97.14**	20.47	89.97	29.78	73.82	**64.26**	**98.97**	**82.76**	**63.71**
Autoregressive Models										
IRASim	41.85	93.72	11.66	78.69	**45.70**	62.76	35.92	93.50	65.40	56.15
Genie Envisioner	28.18	33.81	19.91	87.54	16.04	22.74	2.63	86.83	20.36	41.37
Ctrl-World	42.44	92.77	**33.57**	**90.30**	12.88	62.62	48.20	93.25	67.68	59.98
Standard AR	50.01	95.30	19.73	87.10	23.17	73.12	41.74	88.42	80.88	60.43
ProAR (ours)	50.23	95.34	20.71	88.11	23.12	**73.86**	42.34	91.11	81.96	60.95

Standard AR from 0.663 to 0.786 and 0.683, respectively, while their combination further increases it to 0.801. A similar trend appears on VideoRLVR: Outcome and Transition raise the average precision from 61.56 to 64.82 and 63.50, whereas the complete ProAR reaches 67.17. Overall, ProAR achieves the best aggregate performance on both benchmarks, leading among the AR variants on 7 of the 10 VBVR tasks and 8 of the 12 VideoRLVR metrics. These results show that long-range outcome awareness and short-range transition anticipation provide distinct and complementary guidance.

When to Activate Transition Alignment. We investigate when to activate the transition module during training in Table 2. Applying transition alignment throughout training from scratch (Step 0) yields comparable or slightly inferior performance to activating it at a later stage (Step 7500). We hypothesize that, early in training, the backbone has not yet developed stable representations for the target reasoning domain, making its internal teacher targets less informative. A task-specific fine-tuning warm-up allows more reliable, task-relevant structure to emerge before self-alignment begins. We therefore adopt the late-start strategy as the default configuration in all main experiments.

Ablations on the attention mask, transition layer, and predictor depth are provided in Appendix A.2.

4.5 Extension on Embodied Reasoning

We further evaluate whether ProAR extends beyond visual reasoning to embodied manipulation. Table 4 reports selected WorldArena results, with the complete evaluation and metrics illustration provided in Appendix A.5. ProAR achieves an EWMScore of 60.95, the second-highest result among the methods shown, remaining highly competitive with computationally heavier bidirectional baselines and specialized embodied world models.

Under the controlled comparison with Standard AR, ProAR improves 8 of the 9 component metrics and raises EWMScore from 60.43 to 60.95. The gains are particularly evident in task-relevant dimensions, including Depth Accuracy (88.42 to 91.11), Instruction Following (80.88 to 81.96), Interaction Quality (73.12 to 73.86), and Trajectory Accuracy (41.74 to 42.34). ProAR also improves motion and background consistency while maintaining comparable visual quality. These results suggest that prospective guidance transfers effectively to embodied settings, helping AR video models generate physically coherent and task-directed manipulation trajectories. As shown in Fig. 3, on WorldArena, standard AR rollouts gradually drift and lead to failed grasps, whereas ProAR preserves stable physical interactions and completes the manipulation successfully.

5 Conclusion

In this paper, we present ProAR, a novel framework that equips autoregressive video generation with prospective reasoning capability. To address the short-sightedness of next-chunk prediction, ProAR combines outcome guidance for goal completion with transition guidance for valid intermediate progression. Experiments across perceptual, spatial, abstract, and embodied reasoning benchmarks demonstrate consistent improvements with modest computational overhead. These results highlight prospective modeling as a promising direction for advancing video-native reasoning, opening new possibilities for physical and embodied intelligence.

AI USE STATEMENT

Generative AI tools supported manuscript review, experimental planning, code development, and language editing. The authors conducted the experiments, verified the code, results, text, and references, and take full responsibility for the final manuscript.

ETHICS STATEMENT

This work develops prospective reasoning methods for video generation, evaluated on established benchmarks and simulated manipulation tasks. The research aims to improve goal-directed prediction and understanding of video models. We identify no specific ethical concerns arising from the proposed methodology or the reported experimental setting.

REPRODUCIBILITY STATEMENT

Section 3 describes the autoregressive formulation, goal conditioning, and transition alignment objectives. Section 4.1 and Appendix A.1 specify datasets, training and inference settings, and computational costs. Section 4.4 and Appendix A.2 provide ablation studies. Appendices A.4 and A.5 detail the reasoning-task evaluation protocols and WorldArena metrics.

REFERENCES

Arslan Ali, Junjie Bai, Maciej Bala, Yogesh Balaji, Aaron Blakeman, Tiffany Cai, Jiaxin Cao, Tianshi Cao, Elizabeth Cha, Yu-Wei Chao, et al. World simulation with video foundation models for physical ai. *arXiv preprint arXiv:2511.00062*, 2025.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15619–15629. IEEE, 2023.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first international conference on machine learning*, 2024.

Hila Chefer, Patrick Esser, Dominik Lorenz, Dustin Podell, Vikash Raja, Vinh Tong, Antonio Torralba, and Robin Rombach. Self-supervised flow matching for scalable multi-modal synthesis. *arXiv preprint arXiv:2603.06507*, 2026.

Tianxing Chen, Zanxin Chen, Baijun Chen, Zijian Cai, Yibin Liu, Zixuan Li, Qiwei Liang, Xianliang Lin, Yiheng Ge, Zhenyu Gu, et al. Robotwin 2.0: A scalable data generator and benchmark with strong domain randomization for robust bimanual robotic manipulation. *arXiv preprint arXiv:2506.18088*, 2025.

Xinyan Chen, Ziyu Guo, Renrui Zhang, Dongzhi Jiang, and Hongsheng Li. Opencof: Learning to reason through video generation. *arXiv preprint arXiv:2607.08763*, 2026a.

Yixiang Chen, Peiyan Li, Yuan Xu, Qisen Ma, Jiabing Yang, Kai Wang, Jianhua Yang, Dong An, He Guan, Gaoteng Liu, et al. Flowwam: Optical flow as a unified action representation for world action models. *arXiv preprint arXiv:2607.13017*, 2026b.

Yukang Chen, Luozhou Wang, Wei Huang, Shuai Yang, Bohan Zhang, Yicheng Xiao, Ruihang Chu, Weian Mao, Qixin Hu, Shaoteng Liu, et al. Longlive-2.0: An nvfp4 parallel infrastructure for long video generation. *arXiv preprint arXiv:2605.18739*, 2026c.

Junhao Cheng, Liang Hou, Tianxiong Zhong, Xin Tao, Pengfei Wan, Kun Gai, and Jing Liao. Vlms are good teachers for video reasoning via adaptive test-time optimization. *arXiv preprint arXiv:2606.02564*, 2026.

Jiaxing 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. In *International Conference on Learning Representations*, volume 2026, pp. 85802–85822, 2026.

Yao Feng, Hengkai Tan, Xinyi Mao, Chendong Xiang, Guodong Liu, Shuhe Huang, Hang Su, and Jun Zhu. Vidar: Embodied video diffusion model for generalist manipulation. *arXiv preprint arXiv:2507.12898*, 2025.

Robert Geirhos, Yuxuan Li, Thaddäus Wiedemer, Neha Kalibhat, Zi Wang, Mani Malek, Oyvind Tafjord, Kevin Swersky, Been Kim, and Priyank Jaini. Visual prompt engineering for video models. *arXiv preprint arXiv:2607.25537*, 2026.

Yanjiang Guo, Lucy Shi, Jianyu Chen, and Chelsea Finn. Ctrl-world: A controllable generative world model for robot manipulation. In *International Conference on Learning Representations*, volume 2026, pp. 6121–6138, 2026.

Yasaman Haghighi, Bastien Van Delft, Mariam Hassan, and Alexandre Alahi. Layersync: Self-aligning intermediate layers. In *International Conference on Learning Representations*, volume 2026, pp. 29918–29950, 2026.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, et al. Matrix-game 2.0: An open-source real-time and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2025.

Xuan He, Cong Wei, Yuhao Cheng, Linrui Ma, Yuxuan Zhang, Zuojun Li, Yuhao Wen, Zeyi Liu, Yuren Hao, Songcheng Cai, et al. Vgi-bench: Probing visual intelligence in video generation models. *arXiv preprint arXiv:2608.19583*, 2026.

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

Ziqi Huang, Ning Yu, Gordon Chen, Haonan Qiu, Paul Debevec, and Ziwei Liu. Vchain: Chain-of-visual-thought for reasoning in video generation. *arXiv preprint arXiv:2510.05094*, 2025.

Dengyang Jiang, Mengmeng Wang, Liuzhuozheng Li, Lei Zhang, Haoyu Wang, Wei Wei, Guang Dai, Yanning Zhang, and Jingdong Wang. No other representation component is needed: Diffusion transformers can provide representation guidance by themselves. *arXiv preprint arXiv:2505.02831*, 2025.

Sejoon Jun, Zheng Ding, Huangyuan Su, Weirui Ye, and Yilun Du. Temporal backtracking search for test-time generative video reasoning. *arXiv preprint arXiv:2606.13861*, 2026.

Hirokatsu Kataoka, Yoshihiro Fukuhara, Yonglong Tian, Shangzhe Wu, Oishi Deb, Ryousuke Yamada, Christian Rupprecht, Jianyuan Wang, Kohsuke Ide, Koichi Namekata, et al. Visual general intelligence: A white paper. *arXiv preprint arXiv:2608.25924*, 2026.

Xingjian Leng, Jaskirat Singh, Yunzhong Hou, Zhenchang Xing, Saining Xie, and Liang Zheng. Repa-e: Unlocking vae for end-to-end tuning with latent diffusion transformers. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 18262–18272. IEEE, 2025.

Lin Li, Qihang Zhang, Yiming Luo, Shuai Yang, Ruilin Wang, Fei Han, Mingrui Yu, Zelin Gao, Nan Xue, Xing Zhu, et al. Causal world modeling for robot control. *arXiv preprint arXiv:2601.21998*, 2026.

Yue Liao, Pengfei Zhou, Siyuan Huang, Donglin Yang, Shengcong Chen, Yuxin Jiang, Yue Hu, Si Liu, Jianlan Luo, Liliang Chen, et al. Genie envisioner: A unified world foundation platform for robotic manipulation. In *International Conference on Learning Representations*, volume 2026, pp. 88446–88463, 2026.

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

Kunhao Liu, Wenbo Hu, Jiale Xu, Ying Shan, and Shijian Lu. Rolling forcing: Autoregressive long video diffusion in real time. In *International Conference on Learning Representations*, volume 2026, pp. 91177–91196, 2026.

Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. *Advances in Neural Information Processing Systems*, 36:47500–47510, 2023.

Jiayi Luo, Qiyan Liu, Tengyang Wang, JunHao Liu, Jiayu Chen, Cong Wang, Hanxin Zhu, Chen Gao, Xiaobin Hu, Qingyun Sun, et al. Future forcing: Future-aware training-free kv cache policy for autoregressive video generation. *arXiv preprint arXiv:2605.30083*, 2026.

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

Yu Qi, Xinyi Xu, Ziyu Guo, Siyuan Ma, Renrui Zhang, Xinyan Chen, Ruichuan An, Ruofan Xing, Jiayi Zhang, Haojie Huang, et al. Mme-cof-pro: Evaluating reasoning coherence in video generative models with text and visual hints. *arXiv preprint arXiv:2603.20194*, 2026.

Zezhong Qian, Xiaowei Chi, Chak-Wing Mak, Tianze Zhou, Ruibin Yuan, Yuhan Rui, Hengzhe Sun, Zhuoqun Wu, Yuming Li, Siyuan Qian, et al. Hierarchical denoising for multi-step visual reasoning. *arXiv preprint arXiv:2607.15278*, 2026.

Yu Shang, Zhuohang Li, Yiding Ma, Weikang Su, Xin Jin, Ziyou Wang, Lei Jin, Xin Zhang, Yinzhou Tang, Haisheng Su, et al. Worldarena: A unified benchmark for evaluating perception and functional utility of embodied world models. *AI Open*, 7:208–226, 2026.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. *Advances in neural information processing systems*, 36: 1363–1389, 2023.

Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, WQ Zhang, Weifeng Luo, et al. Magi-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025.

Yuchuan Tian, Hanting Chen, Mengyu Zheng, Yuchen Liang, Chao Xu, and Yunhe Wang. U-repa: Aligning diffusion u-nets to vits. *Advances in Neural Information Processing Systems*, 38:11003–11024, 2026.

Jingqi Tong, Yurong Mou, Hangcheng Li, Mingzhe Li, Yongzhuo Yang, Ming Zhang, Qiguang Chen, Tianyi Liang, Xiaomeng Hu, Yining Zheng, et al. Thinking with video: Video generation as a promising multimodal reasoning paradigm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 41121–41129, 2026.

Team Wan, 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, Mengyang Feng, 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. *arXiv preprint arXiv:2503.20314*, 2025.

Maijunxian Wang, Ruisi Wang, Juyi Lin, Ran Ji, Thaddäus Wiedemer, Qingying Gao, Dezhi Luo, Yaoyao Qian, Lianyu Huang, Zelong Hong, et al. A very big video reasoning suite. *arXiv preprint arXiv:2602.20159*, 2026a.

Ruisi Wang, Zhongang Cai, Fanyi Pu, Junxiang Xu, Wanqi Yin, Maijunxian Wang, Ran Ji, Chenyang Gu, Bo Li, Ziqi Huang, et al. Demystifying video reasoning. *arXiv preprint arXiv:2603.16870*, 2026b.

Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, et al. Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory. *arXiv preprint arXiv:2604.08995*, 2026c.

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

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

Junxiang Xu, Ruisi Wang, Fanyi Pu, Maijunxian Wang, Ran Ji, Tongxi Zhou, Chenyang Gu, Jing Zuo, Hongcan Xiao, Yimeng Geng, et al. Vbvr-pro: A scalable and verifiable suite for native visual reasoning. *arXiv preprint arXiv:2608.26105*, 2026b.

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

Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *International Conference on Learning Representations*, volume 2025, pp. 83048–83077, 2025b.

Runmao Yao, Kairui Hu, Yukang Cao, Ruisi Wang, Shulin Tian, Ziang Cao, Weichen Fan, Ziqi Huang, Yuhao Dong, Hao Li, Zhaoxi Chen, Zhongang Cai, Lei Yang, and Ziwei Liu. Apple-π: Benchmarking thinking with video towards law-grounded physical intelligence, 2026. URL https://arxiv.org/abs/2607.16401.

Seonghyeon Ye, Yunhao Ge, Kaiyuan Zheng, Shenyuan Gao, Sihyun Yu, George Kurian, Suneel Indupuru, You Liang Tan, Chuning Zhu, Jiannan Xiang, et al. World action models are zero-shot policies. *arXiv preprint arXiv:2602.15922*, 2026.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22963–22974. IEEE, 2025.

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. *arXiv preprint arXiv:2410.06940*, 2024.

Yonghao Yu, Lang Huang, Runyi Li, Zerun Wang, and Toshihiko Yamasaki. Video-mirai: Autoregressive video diffusion models need foresight. *arXiv preprint arXiv:2606.03971*, 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. *Advances in Neural Information Processing Systems*, 38:122647–122676, 2026.

Fangqi Zhu, Hongtao Wu, Song Guo, Yuxiao Liu, Chilam Cheang, and Tao Kong. Irasim: A fine-grained world model for robot manipulation. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9834–9844. IEEE, 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. *arXiv preprint arXiv:2602.02214*, 2026a.

Tinghui Zhu, Sheng Zhang, James Y Huang, Selena Song, Xiaofei Wen, Yuankai Li, Hoifung Poon, and Muhao Chen. Video models can reason with verifiable rewards. *arXiv preprint arXiv:2605.15458*, 2026b.

Yixuan Zhu, Jiaqi Feng, Wenzhao Zheng, Yuan Gao, Xin Tao, Pengfei Wan, Jiwen Lu, and Jie Zhou. Astra: General interactive world model with autoregressive denoising. In *International Conference on Learning Representations*, volume 2026, pp. 79167–79184, 2026c.

A APPENDIX

In this supplement, we provide implementation and computational details (Sec.A.1), additional ablations of the Outcome and Transition designs (Sec.A.2), visual analyses of AR failure modes and future representation alignment (Sec.A.3), task specifications and evaluation protocols (Sec.A.4), complete WorldArena results (Sec.A.5), a discussion of limitation (Sec.A.6), and further qualitative examples (Sec.A.7).

A.1 IMPLEMENTATION DETAILS

Video Configuration. In addition to the chunk settings detailed in the main paper, VBVR videos are generated at 512×512 resolution with task-specific sequence lengths. VideoRLVR videos consist of 81 frames at 832×480 resolution. During inference, we employ 20 denoising steps per chunk. Further ablations on the configuration settings are detailed in Sec. A.2.

WorldArena Settings. WorldArena (Shang et al., 2026) evaluates video world models as embodied simulators across 50 RoboTwin 2.0 (Chen et al., 2025) tasks (Chen et al., 2025). Since the original trajectory lengths vary substantially, we follow the FlowWAM (Chen et al., 2026b) protocol to standardize the training data: each video is uniformly sampled to 121 frames and resized to 640×480 resolution. For evaluation, WorldArena introduces 1,000 newly generated test trajectories (20 per task) beyond the official RoboTwin dataset. Following WorldArena's default configuration, we generate these test videos in the identical 121-frame and 640×480 format. This ensures structural consistency between training and inference while preserving the complete physical manipulation process. Given these longer videos, we use a chunk size of 8 and set $\lambda_{\text{goal}} = 0.25$ and $\lambda_{\text{align}} = 0.05$. We train for 10000 steps and start the transition module at step 7000.

Computational Cost. ProAR introduces only modest computational overhead over Standard AR during both training and inference.

Training: Outcome guidance appends one goal frame to each 4-frame chunk, increasing the token count by 8%–11% for videos of 3–8 chunks. Transition guidance is highly efficient: it reuses internal representations from the same forward pass without requiring an external teacher. Its lightweight 3-block predictor (236M parameters, just 4.7% of the 5B backbone) adds a mere 2%–3% to the training computation. Overall, ProAR increases the total training cost by only 12%–14%.

Inference: The Transition predictor is discarded entirely, yielding zero overhead. Meanwhile, the Outcome module jointly denoises the goal frame alongside the current chunk without any additional forward passes or sampling steps. Measured on an NVIDIA B200 GPU, ProAR increases the overall inference time by approximately 12% for a chunk size of 4 compared to standard AR generation. Since only a single goal frame is added per step, this relative overhead diminishes further with larger chunk sizes. For example, generating 8-frame chunks in WorldArena incurs a marginal time increase of only $\sim$8%.

A.2 ADDITIONAL ABLATION STUDIES

In this section, we provide additional ablations on the attention mask, alignment layer, and predictor depth, complementing the studies in Sec. 4.4.

Outcome: Asymmetric Attention Mask. We first validate the asymmetric attention design introduced in Sec. 3.2. As shown in Table 5, removing the mask allows unrestricted interaction between the noisy current chunk and the noisy goal tokens. Our asymmetric design achieves a slightly higher mean score, improving from 0.799 to 0.801, and performs better on 6 of the 10 tasks. Although the overall difference is modest, the result supports the proposed information flow: the goal belief is protected from the uncertain current state while remaining available to guide current generation. Moreover, the asymmetric attention pattern is sparser, making it more computationally efficient. We therefore use the asymmetric mask as the default Outcome configuration.

Transition: Alignment Layer. We next study where future representation self-alignment should be applied within the backbone. Table 6 compares Layers 10, 15, and 20, which span the middle

Table 5: Ablation on asymmetric attention mask.

Model	Concentric Ring	Outline Match	Stable Sort	Star Match	Domino Chain	Ball Eater	Grid Shift	Pipe Puzzle	Slide Puzzle	Balancing Vessels	Mean
w/o Asym. Mask	0.911	0.750	**0.913**	0.865	0.620	**0.797**	0.742	**0.588**	**0.950**	0.861	0.799
ProAR	**0.949**	**0.777**	0.895	**0.871**	**0.628**	0.781	**0.770**	0.546	0.931	**0.863**	**0.801**

portion of the 30-block DiT. All three settings outperform both the standard AR baseline (0.663; Causal-Forcing in Table 1) and Outcome guidance alone (0.786; w/ Outcome in Table 1), showing that Transition guidance is effective across a range of intermediate layers. Layer 15 achieves the highest mean score of 0.801, compared with 0.788 and 0.791 for Layers 10 and 20, respectively. We therefore align the Student and Teacher hidden states at Layer 15 by default.

Table 6: Ablations on the choice of alignment layer. (* denotes default.)

Alignment Layer	Concentric Ring	Outline Match	Stable Sort	Star Match	Domino Chain	Ball Eater	Grid Shift	Pipe Puzzle	Slide Puzzle	Balancing Vessels	Mean
Layer 10	0.948	0.763	0.868	0.866	**0.647**	**0.785**	0.646	**0.563**	0.928	**0.867**	0.788
Layer 15*	**0.949**	**0.777**	0.895	**0.871**	0.628	0.781	0.770	0.546	**0.931**	0.863	**0.801**
Layer 20	0.944	0.767	**0.907**	0.867	0.643	0.764	**0.773**	0.483	0.902	0.862	0.791

Transition: Predictor Depth. Finally, we ablate the depth of the lightweight DiT projector in Table 7. Increasing its depth from 2 to 3 blocks improves the mean score from 0.792 to 0.801. This suggests that the additional block provides greater capacity to map the noisy current hidden states to the clean next-chunk representation. Since the predictor is used only during training, the 3-block design improves alignment without increasing inference cost. We adopt it as our default configuration.

Table 7: Ablation study on the number of projector DiTs. (* denotes default.)

DiT #	Concentric Ring	Outline Match	Stable Sort	Star Match	Domino Chain	Ball Eater	Grid Shift	Pipe Puzzle	Slide Puzzle	Balancing Vessels	Mean
2 DiTs	0.932	0.770	0.860	0.865	0.609	**0.785**	**0.791**	0.510	**0.933**	**0.867**	0.792
3 DiTs*	**0.949**	**0.777**	**0.895**	**0.871**	**0.628**	0.781	0.770	**0.546**	0.931	0.863	**0.801**

A.3 Analysis

We provide visual analyses that complement the motivation in Sec. 3.1 and examine the Transition mechanism introduced in Sec. 3.3. Specifically, we illustrate the characteristic failure modes of Standard AR and visualize how future representation alignment shifts current hidden states toward the next video stage.

Failure Modes of Standard AR. As motivated in Sec. 3.1, video-native reasoning requires both goal completion and valid progression. Fig. 5 illustrates how Standard AR can fail either requirement: a rollout may remain locally plausible but end without completing the task (*Goal Failure*), or produce invalid intermediate interactions that violate task rules (*Rule Violation*). In contrast, a successful trajectory preserves valid transitions and reaches the intended goal. Red boxes highlight erroneous states, while green arrows indicate the successful path.

Visualization of Future Representation Alignment. As described in Sec. 3.3, transition guidance injects prospective information into current hidden states. To examine this effect, we visualize the resulting representations. During teacher-forcing training, all video chunks are processed in parallel within a single forward pass. For each transition from chunk i to chunk $i+1$, we extract the noisy current hidden states from Layer 15 at a moderate noise level ($\sigma = 0.5$). We then pass them through the trained lightweight DiT predictor to obtain the student prediction. The clean hidden

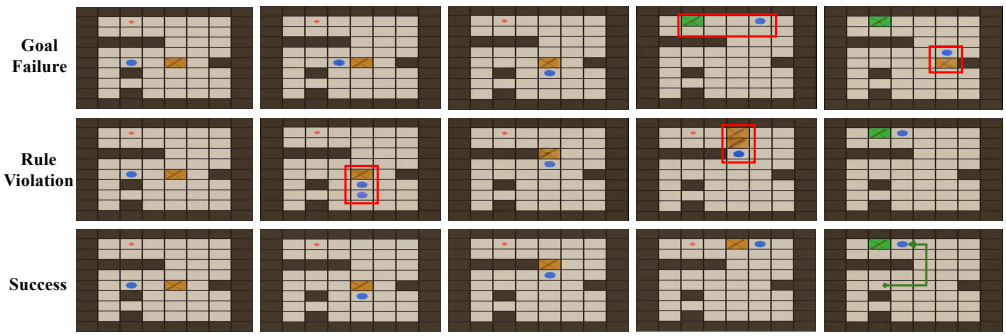


Figure 5: Failure Modes of Standard AR.

states of chunk $i+1$, extracted from the clean video stream at the same layer, serve as the stop-gradient teacher target. We project the current, predicted, and target representations into RGB using a shared PCA basis and normalization range, allowing their spatial structures to be compared.

As shown in Fig. 6, the noisy current representations primarily reflect the spatial state of the current chunk. After prediction, the Student representations shift toward the corresponding clean Teacher representations, recovering the spatial layout and object positions of the next chunk. This pattern is observed for both transitions, from i to $i+1$ and from $i+1$ to $i+2$. The visualization suggests that the predictor does not merely reconstruct the current state. Instead, it learns to map current hidden states toward the upcoming video state, encouraging the backbone to encode prospective features that support progressive reasoning.

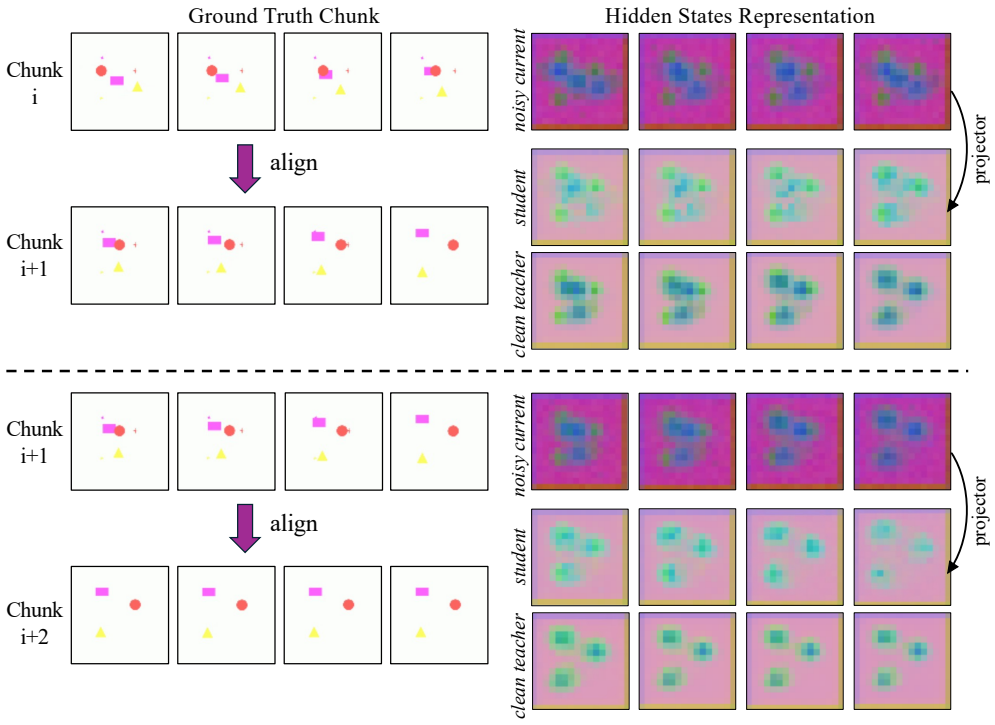


Figure 6: PCA Visualization of Future Representation Alignment.

A.4 Reasoning Tasks and Evaluation Protocols

This section details the reasoning tasks and evaluation protocols used for VBVR and VideoRLVR.

Table 8: Specifications of the 10 selected VBVR tasks (Wang et al., 2026a).

Task	Description
Concentric Ring (G194)	Move two circles to the same center while preserving their sizes and colors.
Outline Match (G25)	Move each colored shape into its matching outline.
Stable Sort (G3)	Group shapes by type and sort each group from smallest to largest.
Star Match (G5)	Move each colored object to the star marker of the same color.
Domino Chain (O23)	Trigger the first domino and propagate the falling sequence through the complete chain.
Ball Eater (O31)	Guide the black ball to consume smaller balls, growing after each step.
Grid Shift (O36)	Shift all blocks by the specified distance and direction.
Pipe Puzzle (O44)	Rotate the pipe tiles until they form a continuous connected path.
Slide Puzzle (O47)	Slide tiles into the empty space to sort them.
Balancing Vessels (O75)	Simulate liquid flow until all communicating vessels reach the same level.

Table 9: Specifications of the 3 VideoRLVR tasks (Zhu et al., 2026b).

Task	Description
Maze	Move the agent from the green start to the red goal through open cells without crossing any walls.
FlowFree	Connect each pair of same-colored endpoints with a continuous path, without overlapping other paths, until the entire grid is filled.
Sokoban	Move the agent through free cells and push the box onto the green target without crossing walls or pulling the box.

Task Specifications. Tables 8 and 9 summarize the reasoning tasks used in our experiments.

Evaluation Metrics. We follow the official evaluation protocols of VBVR and VideoRLVR.

For VBVR (Wang et al., 2026a), task-specific rule-based evaluators assess the generated video against its reference solution and task objective. Depending on the task, they measure final-state correctness, object and spatial consistency, transition validity, and rule compliance. Each video receives a continuous score in $[0, 1]$, where a higher value indicates greater task completion. This score should not be interpreted as a binary success rate.

For VideoRLVR (Zhu et al., 2026b), we report Precision, Recall, F1, and Success Rate (SR), macro-averaged over the test samples. F1 is the harmonic mean of the per-sample Precision and Recall. The task-specific metrics are defined as follows. For *Maze*, Precision and Recall measure the overlap between the generated and reference path regions, while SR requires a connected path from start to goal without corrupting the maze walls. For *FlowFree*, Precision and Recall measure the color correctness and coverage of filled non-endpoint cells. SR additionally requires valid non-overlapping paths, preserved endpoints, correct color connectivity, and sufficient grid coverage. For *Sokoban*, Precision and Recall compare the generated and reference action sequences at corresponding steps. SR requires all boxes to reach their targets through sufficiently valid state transitions.

Precision, Recall, and F1 measure local agreement with the reference solution, whereas SR evaluates strict task completion. Consequently, a high F1 score does not necessarily imply a high success rate.

A.5 Additional WorldArena Evaluation

This section supplements Sec. 4.5 with detailed metric definitions, descriptions of the evaluated baselines, and complete quantitative results.

Metrics. We follow the official WorldArena evaluation protocol, which includes 16 metrics across 6 dimensions. All metrics are higher-is-better.

- *Visual Quality.* Image Quality (IQ) measures frame-level perceptual quality using MUSIQ, while Aesthetic Quality (AQ) evaluates overall visual appeal. JEPA Similarity (JEPA) com-

Table 10: Performance comparison on WorldArena. Within the last two rows, the higher result for each metric is highlighted in orange.

Method	Visual Quality			Motion Quality			Content Consistency			Physics		3D Accuracy		Controllability			EWMScore
	IQ	AQ	JEPA	DD	FS	MS	SC	BC	PC	IntQ	TA	DA	Per	InsF	SA	AcF	
Bidirectional Models																	
CogVideoX	36.23	36.30	94.84	31.47	21.77	73.30	80.97	88.38	33.79	66.96	34.79	91.09	85.46	75.58	89.70	0.00	58.79
Cosmos-Predict 2.5 (text)	**65.49**	**44.33**	46.69	21.62	17.55	70.93	66.19	73.98	35.39	65.14	11.60	70.71	**98.22**	64.70	86.34	**10.01**	53.06
Cosmos-Predict 2.5 (action)	49.48	34.14	90.65	18.96	9.95	65.66	67.24	72.72	45.21	60.46	27.49	90.35	84.68	60.54	89.86	1.31	54.29
Vidar	41.75	39.87	61.10	21.05	17.03	**79.11**	70.37	78.24	20.57	60.84	21.26	79.77	81.88	61.54	88.46	7.86	51.92
FlowWAM	48.87	40.83	**97.14**	30.41	20.47	71.65	82.46	89.97	29.78	73.82	**64.26**	**98.97**	94.56	**82.76**	**89.93**	3.50	**63.71**
Autoregressive Models																	
IRASim	41.85	32.86	93.72	21.60	11.66	64.31	72.82	78.69	**45.70**	62.76	35.92	93.50	84.16	65.40	89.43	3.94	56.15
Genie Envisioner	28.18	26.39	33.81	32.41	19.91	69.17	74.53	87.54	16.04	22.74	2.63	86.83	54.32	20.36	85.98	1.07	41.37
Ctrl-World	42.44	37.05	92.77	**41.82**	**33.57**	77.34	**83.56**	**90.30**	12.88	62.62	48.20	93.25	83.66	67.68	88.68	3.90	59.98
Causal-Forcing	50.01	39.92	95.30	31.31	19.73	69.74	79.90	87.10	23.17	73.12	41.74	88.42	94.80	80.88	89.65	2.07	60.43
ProAR (ours)	50.23	39.65	95.34	31.38	20.71	69.95	80.86	88.11	23.12	**73.86**	42.34	91.11	95.04	81.96	89.71	1.88	60.95

pares high-level spatiotemporal features between generated and reference videos using V-JEPA.

- *Motion Quality.* Dynamic Degree (DD) measures salient motion in active regions, and Flow Score (FS) captures the overall magnitude of optical flow. Motion Smoothness (MS) evaluates whether motion remains continuous and free from abrupt changes or temporal flickering.
- *Content Consistency.* Subject Consistency (SC) measures whether robots and objects preserve their appearance across frames. Background Consistency (BC) evaluates scene stability, while Photometric Consistency (PC) measures local appearance consistency after optical-flow alignment.
- *Physics.* Interaction Quality (IntQ) uses a vision-language model to assess whether robot–object interactions are physically plausible, including contact, grasping, and object motion. Trajectory Accuracy (TA) compares the generated and reference end-effector trajectories after detecting and aligning the robot gripper.
- *3D Accuracy.* Depth Accuracy (DA) compares the depth structure of generated and reference videos. Perspectivity (Per) evaluates whether scale, occlusion, spatial relationships, and perspective remain geometrically plausible.
- *Controllability.* Instruction Following (InsF) evaluates whether the generated video correctly executes the given instruction. Semantic Alignment (SA) measures agreement with the reference video at the task-semantic level. Action Following (AcF) evaluates whether different action conditions produce distinguishable behaviors, rather than directly measuring task success.
- *EWMScore.* WorldArena summarizes overall performance using EWMScore, computed as the arithmetic mean of the 16 metrics after normalizing each to $[0, 100]$:

$$\text{EWMScore} = \frac{1}{16}\sum_{k=1}^{16} s_k. \tag{8}$$

It provides a unified measure of visual fidelity, motion, content consistency, physical plausibility, 3D structure, and controllability.

Baselines. We compare ProAR with general video generators, embodied world models, and a controlled autoregressive baseline. We use the results reported from FlowWAM (Chen et al., 2026b).

General Video Generators.

- *CogVideoX.* CogVideoX (Yang et al., 2025b) is a latent video diffusion model built on an expert transformer and a temporally compressed 3D VAE.
- *Cosmos-Predict 2.5.* Cosmos-Predict 2.5 (Ali et al., 2025) is a flow-based world foundation model that unifies text-, image-, and video-conditioned generation. We evaluate both its text-conditioned and action-conditioned variants.

Embodied World Models.

- *IRASim.* IRASim (Zhu et al., 2025) is an action-conditioned latent diffusion transformer that predicts future robot observations from historical frames and action trajectories.
- *Vidar.* Vidar (Feng et al., 2025) combines large-scale multi-view video diffusion pretraining with masked inverse dynamics to infer robot actions from predicted visual trajectories.
- *Genie Envisioner.* Genie Envisioner (Liao et al., 2026) uses an instruction-conditioned video diffusion model for autoregressive robotic simulation and policy learning.
- *Ctrl-World.* Ctrl-World (Guo et al., 2026) combines joint multi-view prediction, frame-level action conditioning, and pose-conditioned memory retrieval to support controllable and temporally consistent rollouts.
- *FlowWAM.* FlowWAM (Chen et al., 2026b) represents actions as dense optical-flow videos and jointly models flow and RGB observations using a dual-stream diffusion framework.

Controlled Autoregressive Baseline.

- *Standard AR.* We adapt the Wan2.2-TI2V-5B (Wan et al., 2025) backbone to autoregressive generation using causal attention and teacher-forcing training. This serves as our Standard AR baseline and follows the same training and evaluation protocol as ProAR.

Results. As shown in Table 10, ProAR achieves an EWMScore of 60.95, ranking second among the autoregressive models and fourth across all evaluated systems. More importantly, under the controlled comparison with Causal-Forcing, ProAR improves 13 of the 16 base metrics and raises EWMScore from 60.43 to 60.95. The gains span several task-relevant dimensions, including Depth Accuracy (88.42 to 91.11), Instruction Following (80.88 to 81.96), Background Consistency (87.10 to 88.11), Interaction Quality (73.12 to 73.86), and Trajectory Accuracy (41.74 to 42.34). These results indicate that prospective guidance improves not only visual consistency but also spatial accuracy, physical interaction, and task alignment, supporting the extension of ProAR from visual reasoning to embodied video simulation.

A.6 Limitations and Discussions

While ProAR establishes a powerful paradigm for video-native reasoning, our current implementation of Outcome guidance relies on a simple heuristic: utilizing the final video frame to represent the target outcome. This design choice is highly effective for most goal-oriented reasoning tasks, where the final frame naturally encapsulates the goal completion state.

However, this rigid assumption introduces a limitation. In scenarios where the final frame is semantically uninformative, such as tasks with post-completion idle states or purely process-driven dynamics, predicting this frame may provide weak supervision. Consequently, the model may struggle to extract meaningful long-range guidance.

A promising direction for future work is to transition from a fixed final-frame target to adaptive milestone guidance. Rather than strictly anchoring to the last frame, future frameworks could dynamically detect the most salient keyframes or critical bottleneck states within the visual sequence, utilizing them as flexible goals or intermediate subgoals. This would further generalize prospective reasoning to more open-ended and complex physical environments.

A.7 Qualitative Results

We provide additional visualization results on VBVR (Fig. 7, 8), VideoRLVR (Fig. 9), and WorldArena (Fig. 10)

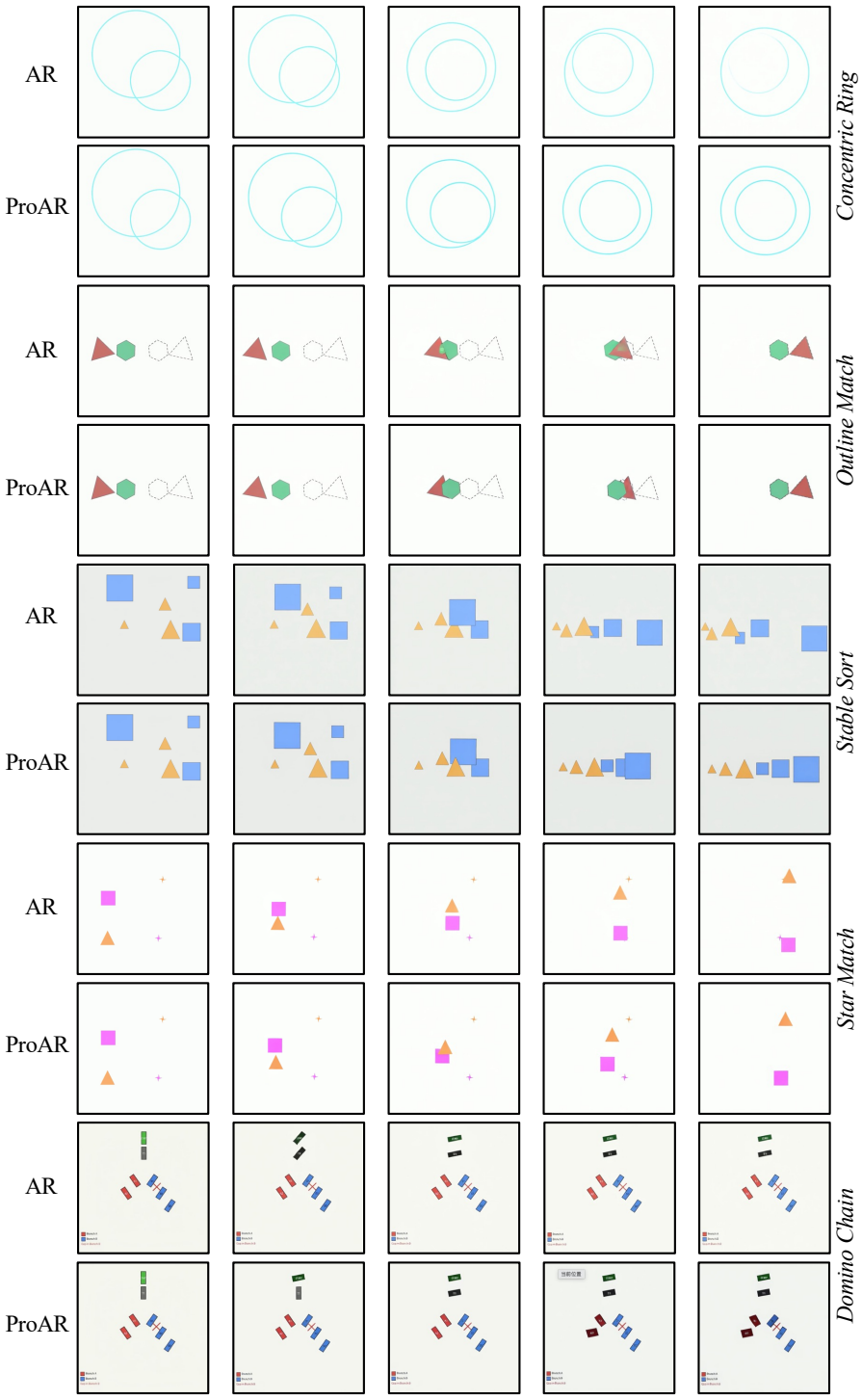


Figure 7: Qualitative Results on VBVR.

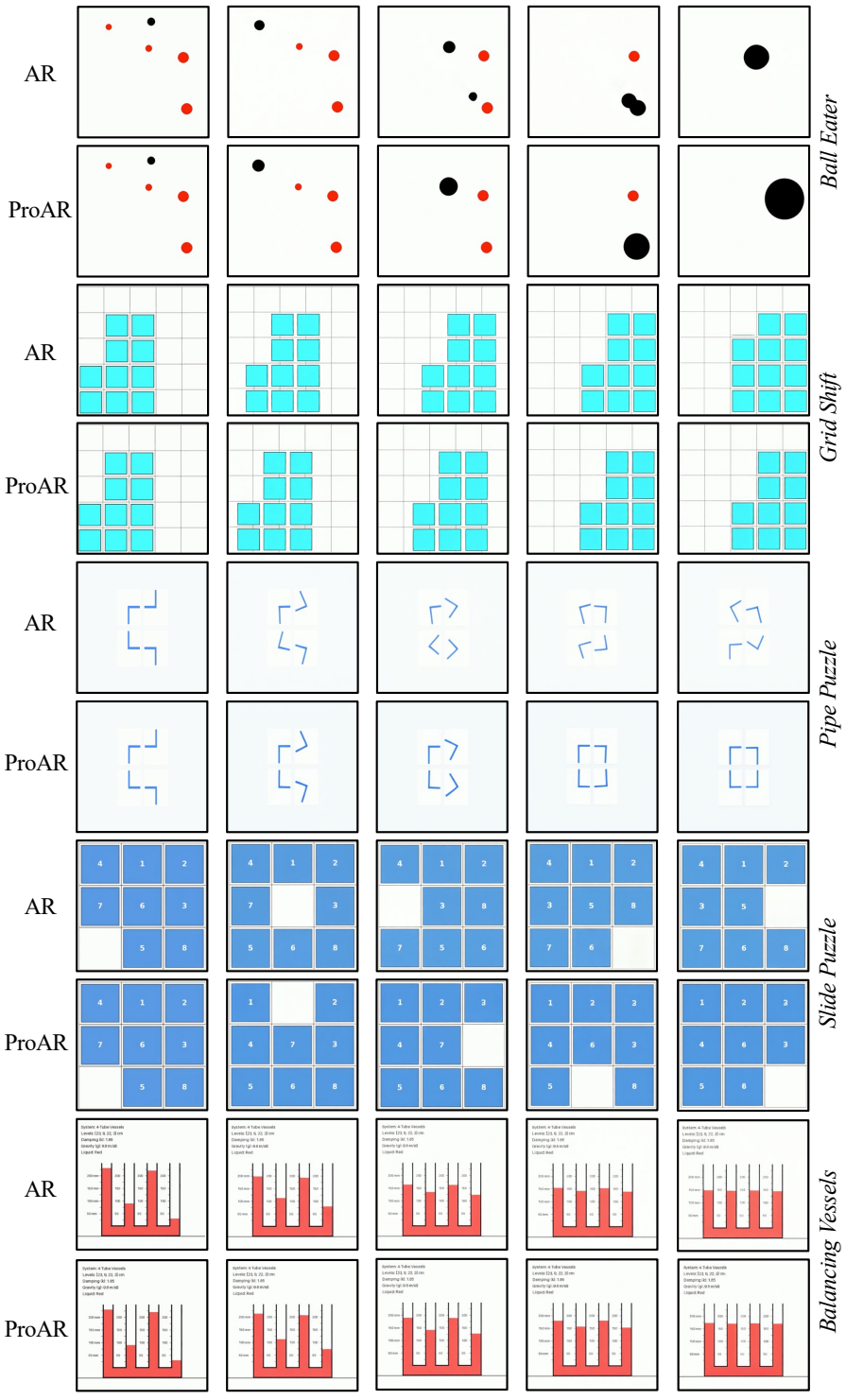


Figure 8: Qualitative Results on VBVR.

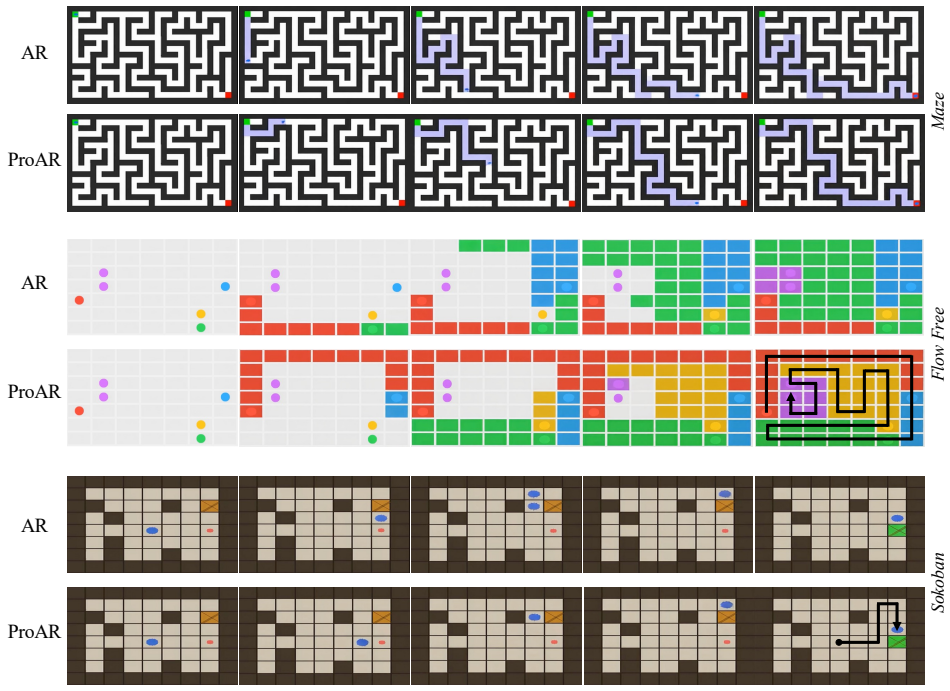


Figure 9: Qualitative Results on VideoRLVR.



Figure 10: Qualitative Results on WorldArena.