Counterfactual World Models via Digital Twin-conditioned Video Diffusion

Yiqing Shen, Aiza Maksutova, Chenjia Li, Mathias Unberath
Johns Hopkins University
{yshen92, unberath}@jhu.edu

Abstract

World models learn to predict the temporal evolution of visual observations given a control signal, potentially enabling agents to reason about environments through forward simulation. Because of the focus on forward simulation, current world models generate predictions based on factual observations. For many emerging applications, such as comprehensive evaluations of physical AI behavior under varying conditions, the ability of world models to answer counterfactual queries – such as "what would happen if this object was removed?" – is of increasing importance. We formalize counterfactual world models that additionally take interventions as explicit inputs, predicting temporal sequences under hypothetical modifications to observed scene properties. Traditional world models operate directly on entangled pixel-space representations where object properties and relationships cannot be selectively modified. This modeling choice prevents targeted interventions on specific scene properties. We introduce CWMDT, a framework to overcome those limitations, turning standard video diffusion models into effective counterfactual world models. First, CWMDT constructs digital twins of observed scenes to explicitly encode objects and their relationships, represented as structured text. Second, CWMDT applies large language models to reason over these representations and predict how a counterfactual intervention propagates through time to alter the observed scene. Third, CWMDT conditions a video diffusion model with the modified representation to generate counterfactual visual sequences. Evaluations on two benchmarks show that the CWMDT approach achieves state-of-the-art performance, suggesting that alternative representations of videos, such as the digital twins considered here, offer powerful control signals for video forward simulation-based world models.

1. Introduction

World models learn to predict the temporal evolution of visual observations, generating future states from current observation [18]. Recent work demonstrates their effective-ness in reinforcement learning [49], robotic control [20, 59], and game playing [28], where agents learn policies through predicted interactions rather than direct environmental exploration. Yet, current world models generate only factual predictions following a given scene [14], lacking the capability to reason about alternative outcomes under hypothetical modifications. Consider an autonomous vehicle encountering an obstacle: beyond predicting the default trajectory, it needs to evaluate how counterfactual scenarios evolve over time, such as "*what sequence of events would unfold if the obstacle moved?*" or "*how would the scene dynamics change if road conditions were different?*" [29]. Therefore, we propose *counterfactual world models* that extend traditional formulations by incorporating interventions as explicit inputs, predicting temporal sequences that capture both immediate intervention effects and their propagation through subsequent time steps.

However, existing world models suffer from two constraints that prevent counterfactual reasoning. First, traditional world models learn direct mappings from observations to future states without explicit factorization of scene components, preventing targeted interventions on specific objects or relationships [18]. Video diffusion models like OpenAI's SORA, LTX-video and Wan2.2 [4, 6, 19, 24, 54, 56], while capable of temporal generation, lack the intervention capabilities required for counterfactual world models [39, 43]. They learn entangled pixel-space representations where object properties, spatial relationships, and temporal dynamics are encoded within the latent distribution [12, 22]. When attempting to implement interventions directly in this entangled space, modifying one object's properties cannot be isolated from other scene elements, preventing controlled propagation of intervention effects through time. Furthermore, the existing world models, particularly those video diffusion models, lack the explicit reasoning capability to determine how interventions should propagate [42].

We introduce CWMDT (Counterfactual World Model with Digital Twin Representation Conditioned Diffusion Model), a framework that can transform video diffusion models into counterfactual world models. Rather than oper-

ating directly on entangled pixel space, we first extract digital twin representations *i.e.*, a structured intermediate representation that explicitly encode objects and relationships in text. The digital twin representation enables large language models (LLMs) to simulate counterfactual dynamics, predicting how interventions affect object states and relationships over time rather than merely generating modified pixels [50, 53]. Afterwards, the modified digital twin representations from LLM condition a video diffusion model to synthesize corresponding visual frames, translating the LLM-predicted temporal evolution into pixel-space video sequences. In other words, the digital twin representation enables a decoupling between reasoning and synthesis, separating the logical determination of how interventions affect scene dynamics from the pixel-level generation process that existing world models cannot achieve.

The major contributions are three-fold. First, we formalize counterfactual world models as an extension of traditional world models that incorporate interventions to generate alternative trajectories. Second, we present CWMDT, a novel framework to turn video diffusion model into counterfactual world model by decomposition of counterfactual generation into perception, intervention, and synthesis through digital twin representations. It demonstrates how video diffusion models can be augmented with explicit reasoning capabilities for LLMs. Third, we validate our approach through extensive experiments on reasoning-intensive benchmarks, where CWMDT achieves superior performance.

2. Related Work

World Models. World models learn latent representations of environment dynamics to generate future states from current observations [18]. For example, early work [18] introduced variational autoencoders combined with recurrent networks to compress visual observations into compact latent representation, allowing agents to train policies through simulated rather than direct interaction. Recent work has explored transformer-based architectures for world modeling [9, 40, 65], showing improved sample efficiency and long-range dependency modeling. Diffusion-based world models have also emerged [1, 13, 45, 55], integrating transformer backbones into diffusion processes for scalable video generation. Beyond architectural improvements, learning paradigms have evolved to optimize for decision-making rather than reconstruction. MuZero [49] learns value-equivalent models that preserve decision-relevant information while discarding reconstruction fidelity. DreamerV3 [21] trains policies by back propagating through predicted trajectories in learned latent space, extending world models to continuous control domains. These approaches have found applications in autonomous driving [15, 26, 38, 57] and embodied AI [35], where simulated interac-

tions enable policy learning and scenario forecasting. Video diffusion models like SORA [44] have been characterized as world simulators due to their emergent object permanence and temporal coherence [62], though recent studies reveal limitations in complex physical reasoning and out-of-distribution generalization [36, 42]. Despite these advances, existing world models generate predictions conditioned solely on observed states and selected actions. Our work formalizes counterfactual world models that accept interventions as explicit inputs, generating multiple plausible trajectories under modified scene conditions.

Video Diffusion Models. Video diffusion models such as OpenAI's SORA [44] show simulation capabilities through large-scale training on diverse visual data, with approaches spanning latent space diffusion [4], text-conditioned generation [54], and real-time synthesis [19, 56]. Recent efforts have adapted video diffusion models toward action-conditioned world models, with Genie [6] learning latent action spaces from unlabeled videos and AVID [48] introducing learned adapters that modify intermediate diffusion outputs based on action inputs. Motion control approaches such as Pandora [27] and Go-with-the-Flow [7] enable trajectory manipulation through structured noise and optical flow guidance [3, 8]. However, video diffusion models generate frames through entangled latent distributions where object properties, spatial relationships, and temporal dynamics are implicitly encoded [12, 22]. Some approaches like NewtonGen [64] attempt to inject physical constraints into generation but remain limited to implicit physical priors embedded in data distributions without structured reasoning about intervention effects. We address this limitation by introducing digital twin representations that decouple reasoning from synthesis, enabling explicit intervention determination before video generation.

Digital Twin Representations. Previous work [50] argues that foundation models (such as the world model) require digital twin representations to capture fine-grained spatial-temporal dynamics and perform causal reasoning. The argument rests on the observation that learned representations in foundation models encode scene properties in entangled latent spaces, making it difficult to isolate and manipulate individual factors such as object positions or physical relationships. Previous work such as just-in-time digital twin framework [53] demonstrates that LLM can dynamically construct digital twin representations from video using vision models, decoupling perception from reasoning to allow multi-step spatial-temporal inference without model fine-tuning. These representations encode object attributes, spatial relationships, and dynamic states in natural language, creating an interface for LLM to apply world knowledge during reasoning [50, 53]. Unlike video dif-

fusion models that learn implicit scene dynamics through entangled latent distributions, digital twin representations make scene factors explicit and separable, allowing controlled modifications to individual objects or relationships.

3. Methods

Formulation of Counterfactual World Model. The world model can be defined as a predictor for future visual observations, formulated as $f : \mathcal{V}_t \times \mathcal{C} \to \mathcal{P}(\mathcal{V}_{t+1:t+k})$. Here, $\mathcal{V}_t$ denotes the space of visual observations in time t, representing a single video frame, while $\mathcal{V}_{t+1:t+k}$ denotes a sequence of future video frames k that span time $t + 1$ to $t + k$. The space $\mathcal{C}$ represents all possible text prompts as conditions, and $\mathcal{P}(\mathcal{V}_{t+1:t+k})$ denotes the probability distribution over these future visual observations. We extend this definition to counterfactual world models by introducing an intervention space $\mathcal{I} \subseteq \mathcal{C}$, which represents conditions that specify counterfactual modifications to scene such as "*what would the scene look like if condition X were different?*" The counterfactual world model can therefore be formulated as $f_{\text{cf}} : \mathcal{V}_t \times \mathcal{I} \to \mathcal{P}(\tilde{\mathcal{V}}_{t:t+k})$, where $\tilde{\mathcal{V}}_{t:t+k}$ represents the space of counterfactual video sequences from time t to $t + k$. Formally, given an initial visual observation v_t and an intervention $i \in \mathcal{I}$, the counterfactual world model generates $\tilde{v}_{t:t+k} \sim f_{\text{cf}}(v_t, i)$ that incorporates both the immediate effects of the intervention and its propagation through subsequent time steps. Sampling from this distribution yields multiple possible outputs $\{\tilde{v}_{t:t+K}^{(1)}, \tilde{v}_{t:t+K}^{(2)}, \ldots, \tilde{v}_{t:t+K}^{(N)}\}$ from a single intervention.

Method Overview. We introduce CWMDT (Counterfactual World Model with Digital Twin Representation Conditioned Diffusion Model), an end-to-end implementation of the counterfactual world model using digital twin representations as an intermediate layer, as shown in Fig. 1. Formally, we decompose the counterfactual world model into three consecutive mappings, depicted as

$$f_{\text{cf}} = f_{\text{synth}} \circ f_{\text{interv}} \circ (f_{\text{percept}}, \text{id}), \tag{1}$$

where id denotes the identity function, ensuring that $(f_{\text{percept}}, \text{id})$ transforms the input pair (v_t, i) into (s_t, i). Perception mapping $f_{\text{percept}} : \mathcal{V}_t \to \mathcal{S}_t$ converts video frames to digital twin representations through vision models [53], where $\mathcal{S}_t$ denotes the space of digital twin representations. The intervention mapping $f_{\text{interv}} : \mathcal{S}_t \times \mathcal{I} \to \mathcal{P}(\tilde{\mathcal{S}}_{t:t+k})$ generates modified digital twin representations under interventions $i \in \mathcal{I}$ through an LLM. Here, $\tilde{\mathcal{S}}_{t:t+k}$ denotes the space of counterfactual digital twin representation sequences spanning from time t to $t + k$ that reflect both the intervention and their predicted temporal evolution. Innovatively, rather than operating on video frames directly, interventions are applied to the digital twin representation s_t to

enable explicit reasoning over scene factors with embedded world knowledge in LLM Finally, the synthesis mapping $f_{\text{synth}} : \tilde{\mathcal{S}}_{t:t+k} \to \mathcal{P}(\tilde{\mathcal{V}}_{t:t+k})$ generates video frames conditioned on the modified digital twin representation through a video diffusion model, where $\tilde{\mathcal{V}}_{t:t+k}$ represents the space of counterfactual video sequences.

Digital Twin Representation Construction. To enable counterfactual reasoning over the scene, we first transform each given video frame $v_t \in \mathcal{V}_t$ into a digital twin representation $s_t \in \mathcal{S}_t$, depicted as:

$$s_t = \{(j, c_j^{(t)}, a_j^{(t)}, p_j^{(t)}, m_j^{(t)})\}_{j=1}^{N_t}, \tag{2}$$

where N_t denotes the number of object instances in the frame v_t. Each instance tuple contains an identifier j that maintains correspondence across frames, a semantic category $c_j^{(t)}$ describing the object class, attribute descriptions $a_j^{(t)}$ capturing visual properties such as color and texture, spatial properties $p_j^{(t)} = (x, y, z, w, h)$ encoding centroid coordinates, depth, width, and height, and a segmentation mask $m_j^{(t)}$ defining the precise object locations. We construct s_t through various vision foundation models operating on individual frames. Object segmentation and cross-frame tracking are performed through SAM-2 [30, 47], which generates instance-level masks and maintains object identity across video sequences. Depth estimation network DepthAnything [61] computes per-pixel depth maps that we sample at object centroids to obtain spatial positioning. Semantic categorization assigns each detected instance to conceptual classes through object detection model, *i.e*, OWLv2 [41]. QWen2.5-VL [2] generates natural language descriptions of object attributes by analyzing localized image regions corresponding to each segmentation mask. We serialize the resulting digital twin representation s_t in structured text format of JSON, which transforms the counterfactual world model problem from reasoning over visual observations to reasoning over explicit textual scene descriptions.

Counterfactual Reasoning over Digital Twin Representation. Given a digital twin representation s_t and an intervention $i \in \mathcal{I}$, we then implement the intervention mapping $f_{\text{interv}} : \mathcal{S}_t \times \mathcal{I} \to \mathcal{P}(\tilde{\mathcal{S}}_{t:t+k})$ to generate a sequence of modified digital twin representations with LLM. First, the LLM analyzes the given intervention to identify which objects and relationships within s_t are directly affected. Then, LLM predicts the temporal evolution of these changes in the subsequent video frames. For instance, given an intervention such as "*remove the obstacle from the path*," the LLM identifies the relevant object instance in s_t, determines which spatial relationships change as a result, and predicts how other objects might respond to the newly avail-
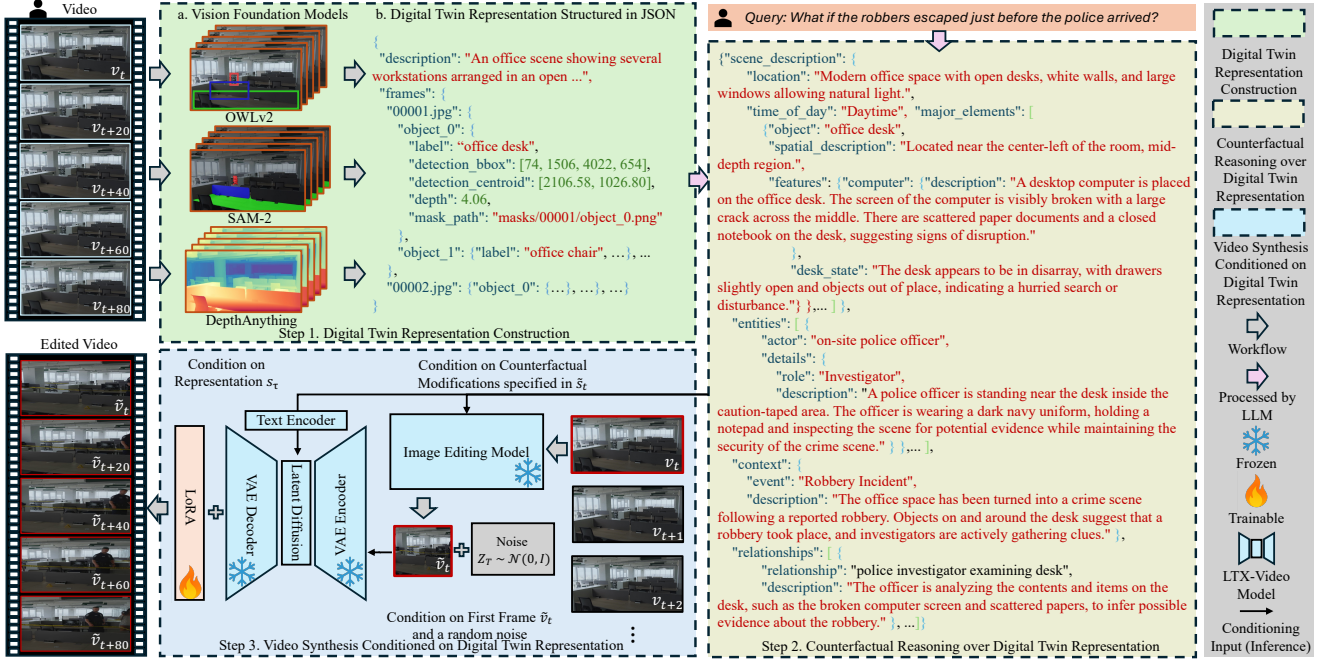
Figure 1. Method overview for CWMDT. Our approach consists of three stages. (1) Digital twin representation construction: Vision models extract structured scene representations s_t from video frames v_t. (2) Counterfactual reasoning: An LLM processes intervention queries to predict temporal evolution, generating modified digital twin representations $\tilde{s}_{t:t+k}$. (3) Video synthesis: A fine-tuned diffusion model generates counterfactual videos $\tilde{v}_{t:t+k}$ conditioned on the edited first frame $\tilde{v}_t$ and the modified digital twin representation $\tilde{s}_{t:t+k}$.

able space across subsequent time steps. The output consists of a sequence of modified digital twin representations $\tilde{s}_{t:t+k} = \{\tilde{s}_t, \tilde{s}_{t+1}, \ldots, \tilde{s}_{t+k}\}$, where $\tilde{s}_t$ encodes the immediate effects of the intervention, and subsequent representations capture the predicted temporal propagation. Each $\tilde{s}_\tau$ maintains the same structural format as s_t, preserving the object-level decomposition. Finally, by sampling the distribution $\mathcal{P}(\tilde{\mathcal{S}}_{t:t+k})$, we may obtain multiple plausible counterfactual trajectories that reflect uncertainty in how interventions might propagate.

Video Synthesis Conditioned on Digital Twin Representation. The synthesis mapping $f_{\text{synth}} : \tilde{\mathcal{S}}_{t:t+k} \to \mathcal{P}(\tilde{\mathcal{V}}_{t:t+k})$ generates counterfactual video sequences from the modified digital twin representations via a video diffusion model. Formally, we adopt a pre-trained video diffusion model as the backbone and fine-tune it on paired data of digital twin representations and corresponding video frames. During fine-tuning, the backbone video diffusion model learns to condition the denoising process on both the digital twin representations s_τ at each frame τ as textual input and the corresponding first frame v_t as visual input to generate subsequent frames. Through this fine-tuning process, the video diffusion model therefore learns to predict subsequent frame dynamics from the initial visual state while respecting the temporal evolution specified

by the digital twin representations. During inference, we first apply an image editing method to modify the original frame v_t according to the counterfactual modifications specified in $\tilde{s}_t$, producing an edited frame $\tilde{v}_t$ that visually reflects the intervention effects. This editing step ensures consistency between the visual starting point and the textual scene description, as directly conditioning on the unmodified frame v_t would create a mismatch with the counterfactual digital twin sequence. Given the counterfactual digital twin sequence $\tilde{s}_{t:t+k}$ and the edited initial frame $\tilde{v}_t$, the fine-tuned video diffusion model then generates the corresponding counterfactual video frames. Eventually, by sampling multiple digital twin sequences from $\mathcal{P}(\tilde{\mathcal{S}}_{t:t+k})$, we can therefore generate various counterfactual videos through repeated inference runs on this video diffusion model.

4. Experiments

Implementation Details. We implement all experiments using PyTorch 2.8.0 on one NVIDIA GeForce RTX 4090 GPU with 48 GB memory. The intervention mapping uses Qwen3-VL-8B-Instruct [2] as the LLM backbone to perform counterfactual reasoning on digital twin representations. For video synthesis, we adopt LTX-Video [19] as the pre-trained video diffusion model backbone and fine-tune it on 95 paired samples of digital twin representations and
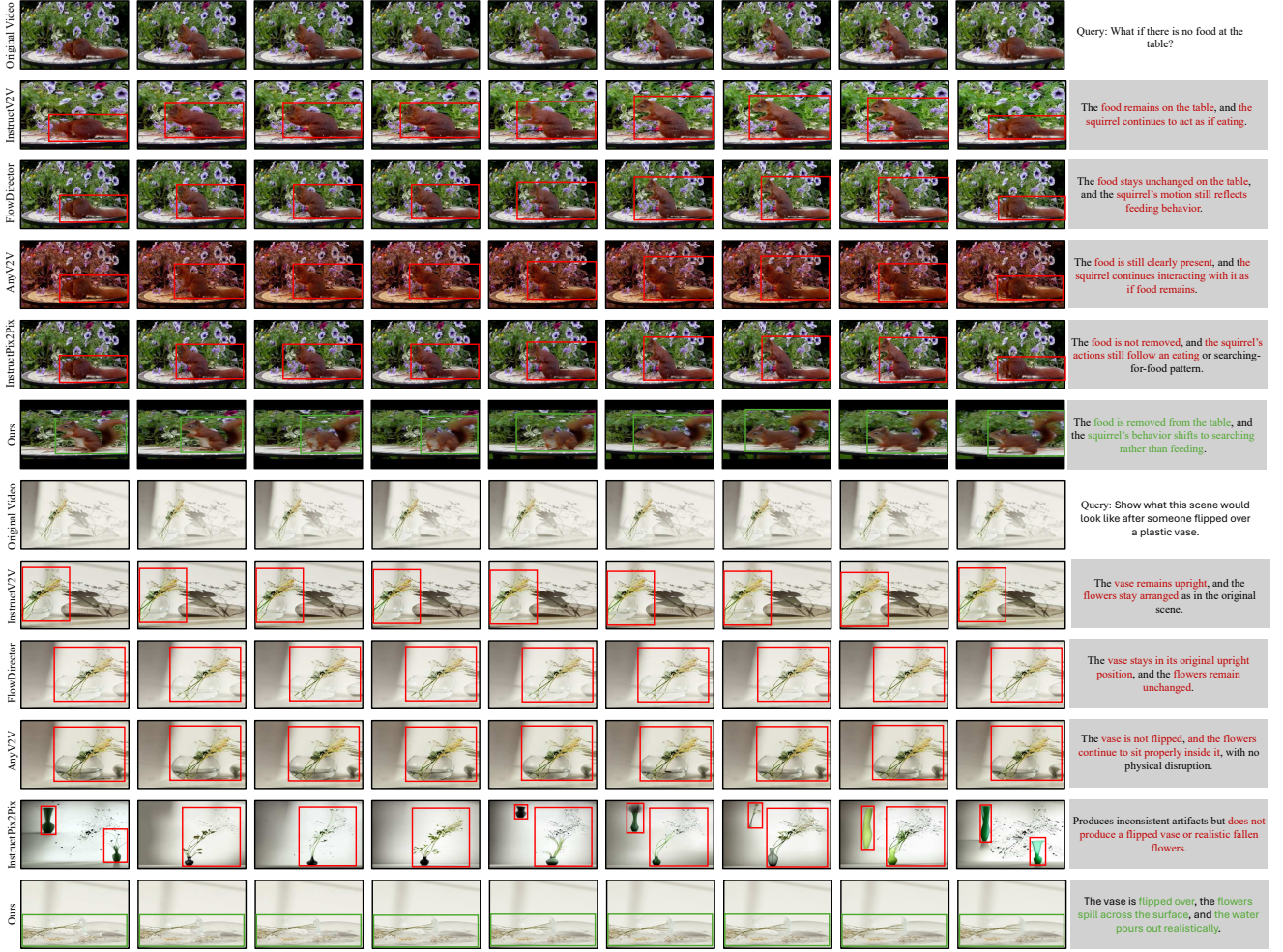
Figure 2. Qualitative comparison of counterfactual world model capabilities across different methods. Two intervention scenarios test whether models can predict alternative temporal sequences. CWMDT correctly generates counterfactual trajectories. Compared methods fail to execute these interventions. Red boxes indicate regions where intervention effects should appear.

Table 1. Quantitative evaluation on RVEBench. Each metric is assessed across three levels of reasoning complexity (L1, L2, L3) in percentage (%). Higher scores indicate better performance for all metrics.

Method	CLIP-Text (↑)			CLIP-F (↑)			GroundingDINO (↑)			LLM-as-a-Judge (↑)		
	L1	L2	L3	L1	L2	L3	L1	L2	L3	L1	L2	L3
InstructDiff [17]	19.57	18.23	17.21	86.80	86.40	86.17	14.73	10.08	9.38	44.33	39.89	36.79
InstructV2V [46]	22.22	21.43	19.78	91.80	91.44	90.59	11.48	7.50	5.13	38.86	35.95	34.70
FlowDirector [33]	19.56	18.70	17.15	93.90	94.06	94.10	8.73	8.11	5.56	35.52	30.75	29.96
AnyV2V [31]	17.05	16.29	15.54	93.71	93.85	93.61	13.72	12.38	9.94	20.01	18.47	17.68
InstructPix2Pix [5]	18.49	17.45	16.73	92.18	92.73	93.06	6.19	6.32	9.68	34.53	30.76	29.49
CWMDT (Ours)	**26.18**	**25.42**	**26.39**	**97.87**	**98.45**	**98.48**	**29.16**	**28.57**	**33.33**	**62.47**	**64.06**	**58.81**

corresponding video frames from RVTBench [51]. During fine-tuning of the video diffusion models, we perform LoRA [25] fine-tuning with rank 32 for 100 epochs with a batch size of 2 using the AdamW optimizer and Cosine scheduler with a learning rate of 1e-4. The diffusion model generates videos at 24 fps with a resolution of 768×768 pixels over 65 frames. For image editing to produce the modified initial frame $\tilde{v}_t$, we use Qwen-Image-Edit-2509 [58].
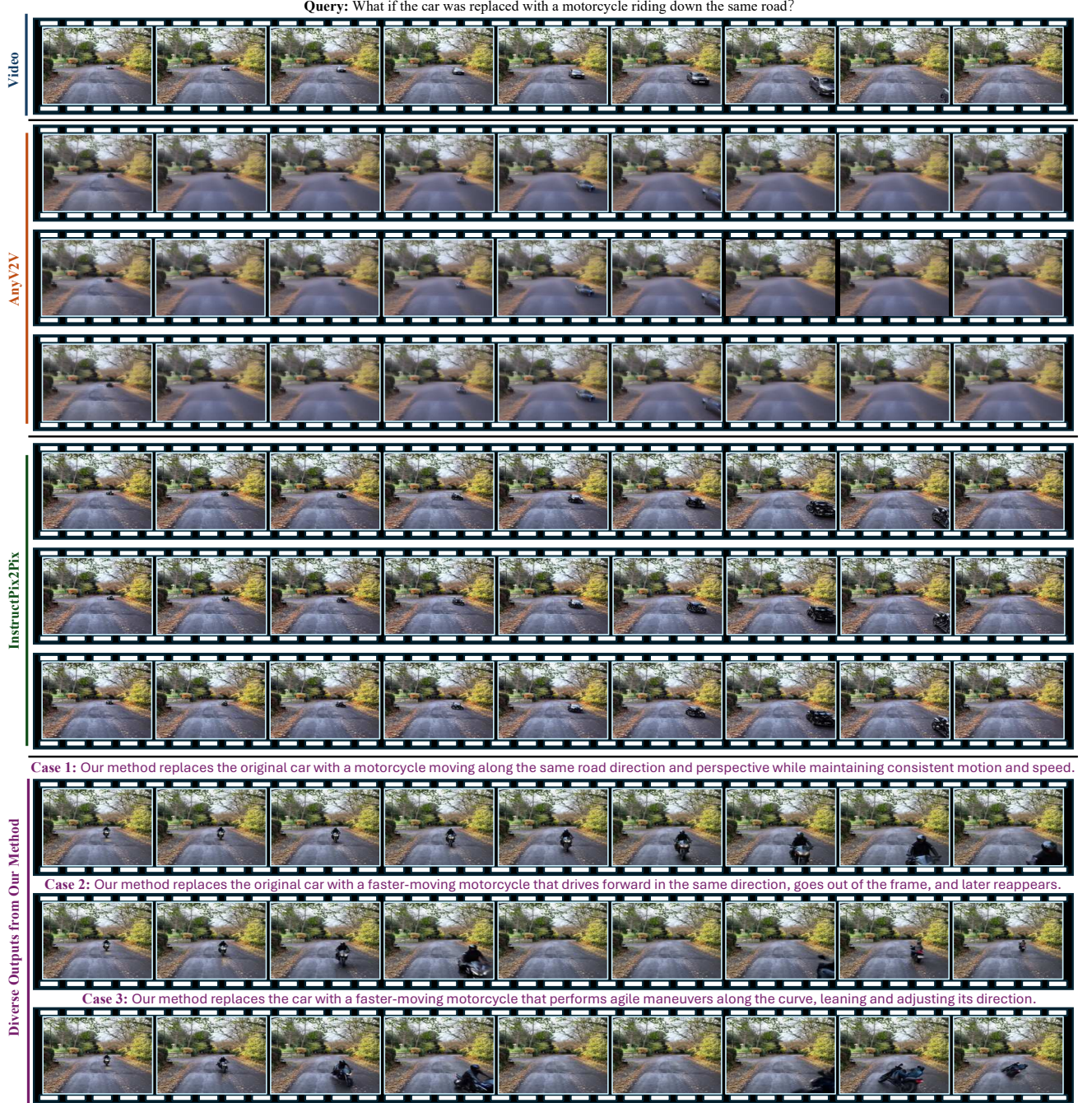
Figure 3. Demonstration of diverse counterfactual trajectory generation from a single intervention. Given the query to replace a car with a motorcycle, CWMDT produces three distinct plausible scenarios: maintaining the original motion pattern (Case 1), accelerating beyond the frame boundary and reentering (Case 2), and executing agile cornering maneuvers (Case 3). Each trajectory respects physical constraints while exploring different behavioral possibilities that could arise from the same initial intervention. Baseline methods either fail to execute the vehicle replacement or produce visually inconsistent results, lacking the ability to reason about multiple plausible outcomes.

During inference, we sample 3 counterfactual digital twin sequences from $\mathcal{P}(\tilde{\mathcal{S}}_{t:t+k})$ for each intervention to generate multiple plausible counterfactual trajectories.

Benchmark Datasets and Metrics. We evaluate CWMDT on two benchmarks that test different aspects of counterfactual world model capabilities. First, RVEBench [52] provides 100 videos with 519 queries for

Table 2. Ablation study evaluating the contribution of each component in CWMDT on the FiVE benchmark. Checkmarks (✓) indicate component presence, while crosses (✗) indicate removal. Results show that digital twin representations and LLM-based intervention reasoning are important for accurate counterfactual world modeling, while the edited initial frame ensures visual-textual consistency.

Digital Twin Representation	LLM Intervention Reasoning	LLM Scale	Modified Initial Frame $\tilde{v}_t$	CLIP-Text (↑)	CLIP-F (↑)	MUSIQ (↑)	SSIM (↑)	PSNR (↑)	GroundingDINO (↑)	LLM-as-a-Judge (↑)
✗	✓	8B	✓	$26.35_{\pm1.47}$	$97.21_{\pm0.14}$	$45.74_{\pm2.32}$	$45.09_{\pm0.55}$	$13.17_{\pm0.60}$	$17.65_{\pm12.42}$	$43.62_{\pm9.82}$
✓	✗	8B	✓	$27.10_{\pm1.30}$	$97.15_{\pm0.20}$	$43.50_{\pm2.10}$	$46.00_{\pm0.60}$	$14.00_{\pm0.65}$	$19.50_{\pm11.00}$	$46.99_{\pm8.75}$
✓	✓	1.5B	✓	$28.90_{\pm1.10}$	$98.00_{\pm0.28}$	$47.20_{\pm1.85}$	$50.10_{\pm0.50}$	$16.50_{\pm0.55}$	$24.00_{\pm10.50}$	$51.26_{\pm6.13}$
✓	✓	8B	✗	$27.98_{\pm0.99}$	$97.09_{\pm0.36}$	$36.53_{\pm1.66}$	$45.38_{\pm0.53}$	$13.66_{\pm0.54}$	$17.34_{\pm11.58}$	$48.31_{\pm7.51}$
✓	✓	8B	✓	$\mathbf{30.59}_{\pm1.83}$	$\mathbf{98.85}_{\pm0.25}$	$\mathbf{50.19}_{\pm1.95}$	$\mathbf{53.47}_{\pm0.52}$	$\mathbf{18.32}_{\pm0.57}$	$\mathbf{30.18}_{\pm6.25}$	$\mathbf{63.02}_{\pm5.01}$

Figure 4. Diverse counterfactual scenarios generated from a single original video sequence using the proposed CWMDT.

reasoning video editing, which tests whether the model can reason about counterfactual scenarios that require multi-hop reasoning. It is organized into three levels of complexity in reasoning (L1, L2, L3), where each level requires progressively more reasoning steps to identify the intervention targets from implicit queries. FiVE benchmark [34] contains 100 videos with 420 object-level query pairs across six fine-grained editing types, testing the model's ability to execute precise interventions while maintaining temporal consistency. We employ four metrics: CLIP-Text [23] measures the semantic alignment between the generated counterfactual video and the intervention description; CLIP-F [23, 60, 63] evaluates the temporal coherence between frames in the counterfactual sequence; GroundingDINO [37] assesses whether the intervention targets are correctly localized in the generated video; and LLM-as-a-Judge [66] assesses whether the counterfactual outcome aligns with the intervention intent. We report all metrics as percentage.

Compared Methods. We compare CWMDT with five video generative models that represent different approaches to instruction-driven visual manipulation. Three methods operate directly on video: InstructV2V [46] performs end-to-end instruction-based editing through diffusion models, FlowDirector [33] applies optical flow for localized modi-

fications, and AnyV2V [31] converts image editing models into video editors through temporal feature injection. Two image editing methods are also included by frame-by-frame processing in the videos: InstructDiff [17] interprets natural language instructions for image manipulation, while InstructPix2Pix [5] learns to follow editing instructions through conditional diffusion training. These baselines reveal the limitations of existing approaches when confronted with counterfactual reasoning in world models as they operate directly on pixel representations without explicit scene understanding. Our comparison evaluates whether decomposing counterfactual world modeling into perception, reasoning, and synthesis through digital twin representations offers advantages over direct pixel-space editing.

Evaluations on RVEBench. Table 1 presents quantitative results in RVEBench, where CWMDT outperforms all compared methods in all metrics and complexity levels. For GroundingDINO, CWMDT achieves 29.16%, 28.57%, and 33.33% at L1, L2, and L3 respectively, compared to the next best scores of 14.73%, 12.38%, and 9.94%. This improvement demonstrates that digital twin representations enable precise spatial grounding during counterfactual reasoning. Similarly, LLM-as-a-Judge scores show CWMDT achieving 62.47%, 64.06%, and 58.81%, substantially higher than others' 20.01%-44.33%, 18.47%-39.89%, and 17.68%-36.79% across the three levels. These results validate that separating reasoning from synthesis through digital twin representations produces counterfactual trajectories that align better with intervention semantics. The compared methods show declining performance as complexity increases from L1 to L3, reflecting their limitation in propagating intervention effects through time without explicit scene understanding. On the contrary, CWMDT maintains consistent performance and even improves at L3 for GroundingDINO. The high CLIP-F scores across all methods (above 86%) confirm temporal consistency in video generation, yet CWMDT achieves the highest scores (97.87%-98.48%), demonstrating that conditioning on digital twin representations preserves coherent temporal dynamics while executing interventions.

Fig. 2 presents qualitative comparisons. For example,

when asked to remove food from the table, CWMDT generates video sequences where the squirrel's behavior adapts from feeding to searching, while the compared methods fail to execute the intervention and continue showing the squirrel interacting with the still-present food. Fig. 3 demonstrates CWMDT's ability to generate multiple plausible counterfactual trajectories from a single intervention. Fig. 4 illustrates qualitative examples in which CWMDT generates realistic counterfactual scenarios for the same given image with different counterfactual queries. These findings confirm that by introducing interventions as explicit inputs and reasoning over compositional scene structure, CWMDT generates alternative trajectories that accurately reflect hypothetical modifications to scene properties.

Table 3. Quantitative evaluation of video editing methods on FiVE dataset. Each metric assesses editing quality from different perspectives. Higher scores indicate better performance for all metrics.

Method	CLIP-Text	CLIP-F	GroundingDINO	LLM-as-a-Judge
InstructDiff [17]	$24.81_{\pm 0.28}$	$88.03_{\pm 0.36}$	$17.67_{\pm 4.07}$	$54.83_{\pm 7.11}$
InstructV2V [46]	$25.31_{\pm 0.25}$	$91.84_{\pm 0.28}$	$14.67_{\pm 3.96}$	$59.85_{\pm 10.50}$
FlowDirector [33]	$20.50_{\pm 0.44}$	$96.91_{\pm 0.11}$	$19.59_{\pm 7.50}$	$37.20_{\pm 8.16}$
AnyV2V [31]	$24.73_{\pm 0.36}$	$96.98_{\pm 0.12}$	$20.00_{\pm 5.26}$	$54.85_{\pm 8.76}$
InstructPix2Pix [5]	$23.22_{\pm 0.27}$	$92.26_{\pm 0.25}$	$22.81_{\pm 4.45}$	$41.85_{\pm 12.34}$
CWMDT (Ours)	$\mathbf{30.59}_{\pm 1.83}$	$\mathbf{98.85}_{\pm 0.25}$	$\mathbf{30.18}_{\pm 6.25}$	$\mathbf{63.02}_{\pm 5.01}$

Evaluations on FiVE Benchmark. Tab. 3 shows the evaluation results on the FiVE becnhmark. CWMDT achieves 30.59% CLIP-Text score, 98.85% CLIP-F score, 30.18% GroundingDINO score, and 63.02% LLM-as-a-Judge score, outperforming all compared methods. The improvements over baselines are particularly observable in CLIP-Text (20.8% relative gain over the 25.31% achieved by InstructV2V) and GroundingDINO (32.3% relative gain over the 22.81% achieved by InstructPix2Pix), demonstrating that digital twin representations provide advantages beyond complex reasoning scenarios. Unlike RVEBench, where the compared methods showed a consistent decline in complexity levels, the FiVE results reveal that pixel-space approaches achieve high temporal consistency (CLIP-F scores greater than 88%) but struggle with semantic alignment and spatial grounding. This pattern suggests that existing video editing methods can maintain frame-to-frame coherence but fail to execute interventions that require understanding and modifying specific scene components. The standard deviation analysis provides additional evidence. CWMDT shows comparable or lower variance than compared methods on CLIP-F (0.25) and LLM-as-a-Judge (5.01), indicating stable performance despite the added complexity of three-stage decomposition. Baseline methods exhibit higher variance on GroundingDINO (ranging from 3.96 to 7.50), reflecting inconsistent spatial

grounding. These results show that CWMDT's advantages extend beyond reasoning-intensive benchmarks to general video editing tasks.

Ablation Study. We perform ablation on the FiVE benchmark to evaluate the contribution of each component in CWMDT, as shown in Tab. 2. Removing digital twin representations and instead directly conditioning the diffusion model on input text prompts together with the edited first frame results in decreased GroundingDINO scores (17.65% versus 30.18%) and LLM-as-a-Judge scores (43.62% versus 63.02%). It demonstrates that structured digital twin representations enable more accurate spatial localization and intervention execution compared to entangled text embeddings. Removing LLM intervention reasoning by switching Qwen3 to non-reasoning mode reduces LLM-as-a-Judge scores from 63.02% to 46.99%, indicating that explicit multi-hop reasoning over digital twin representations produces counterfactual trajectories that better align with intervention semantics. Scaling down the LLM from Qwen3-8B to Qwen3-1.5B decreases performance across all metrics, with GroundingDINO dropping from 30.18% to 24.00% and LLM-as-a-Judge declining from 63.02% to 51.26%, confirming that larger LLM provide stronger reasoning capabilities for determining how interventions should propagate over time. Removing the modified initial frame $\tilde{v}_t$ and instead using the original frame v_t leads to degraded MUSIQ scores (36.53% versus 50.19%) and lower LLM-as-a-Judge scores (48.31% versus 63.02%). It reveals that visual-textual consistency between the starting frame and the counterfactual digital twin sequence is necessary for the diffusion model to generate alternative trajectories. These results confirm that all components contribute to counterfactual world modeling, with digital twin representations and LLM-based reasoning being the most important for producing accurate interventions and their temporal propagation.

5. Conclusion

World models enable forward simulation of environment dynamics, yet existing methods generate only factual predictions from observed states. We formalize counterfactual world models that accept interventions as explicit inputs alongside visual observations, extending forward simulation to hypothetical scenarios. This extension serves physical AI evaluation, where agents must reason about alternative outcomes before committing to actions. CWMDT demonstrates that video diffusion models can be transformed into counterfactual world models through a three-stage decomposition: perception constructs digital twin representations that make scene structure explicit, intervention reasoning through LLMs determines how modifications propagate across time, and synthesis generates cor-

responding visual sequences. Digital twin representations function as alternative control signals for video forward simulation, exposing compositional scene factors that enable selective modifications to specific objects and relationships rather than operating on entangled pixel distributions. This decomposition separates logical intervention determination from visual generation, allowing world models to leverage embedded world knowledge in LLMs for reasoning about counterfactual dynamics. Future work may expand digital twin representations to capture finer-grained physical properties and explore how counterfactual world models can guide decision-making in autonomous systems where evaluating hypothetical scenarios is necessary for safe operation.

A. Additional Experiments

To validate CWMDT beyond the primary benchmarks, we select the CausalVQA [16] debug dataset split. This choice aligns naturally with our counterfactual world model formulation for three reasons. First, CausalVQA explicitly tests counterfactual reasoning through questions that probe alternative outcomes under hypothetical modifications to observed scenes, directly matching our model's design objective of predicting temporal sequences under interventions. Second, the benchmark grounds its questions in real-world physical scenarios captured through egocentric videos, providing the complex visual dynamics and object interactions that our digital twin representations are designed to capture. Third, unlike synthetic simulation benchmarks that simplify physical scenes, CausalVQA presents the authentic complexity of real environments while maintaining focus on physically grounded causal reasoning rather than purely descriptive visual understanding.

Experimental Settings. We conduct our evaluation on the CausalVQA debug dataset split, which contains 20 samples categorized as "easy" difficulty, with each sample paired with two question variants to test robustness to language perturbations. We select this split because it provides ground-truth target values for every case, enabling detailed analysis of model behavior. A similarly fine-grained examination on the full test split remains infeasible as those target values are withheld for leaderboard purposes.

For each question in this debug split, we first apply CWMDT to generate the corresponding counterfactual video sequence based on the intervention specified in the question. We then construct the input to each VLM by concatenating the original video with the generated counterfactual video, followed by the question text. This allows the VLM to compare the factual trajectory against the counterfactual trajectory predicted by CWMDT when answering. For baseline comparisons, we feed only the original video and the question to the models, following the standard CausalVQA evaluation protocol. This comparison reveals whether CWMDT's counterfactual predictions contain information that improves model performance on counterfactual reasoning tasks, or alternatively, whether the generated videos introduce artifacts that degrade answer quality.

Compared Methods. We evaluated the same set of models used in the original CausalVQA paper [16] under identical inference configurations. Open-source VLMs include LLaVA-OneVision [32], Qwen2.5-VL [2], PerceptionLM [11], and InternVL-2.5 [10]. Commercial closed VLMs consist of GPT-4o and Gemini 2.5 Flash. To establish a human baseline, we recruit five independent annotators with no prior exposure to the dataset to answer the benchmark questions.

Results and Analysis. Table 4 presents the results on the CausalVQA debug dataset across five question categories. CWMDT augmentation (Qwen2.5VL+CWMDT) achieves the best model performance on anticipation questions with 62.50% accuracy, exceeding the strongest baseline by 7.50%. For counterfactual questions, our method matches the top-performing closed model Gemini 2.5 Flash at 70.00%, while outperforming the base Qwen2.5VL model by 17.50%. On hypothetical questions, CWMDT reaches 72.50%, tying with GPT-4o for the highest score. These results suggest that explicit counterfactual video generation provides the most value for question types that require reasoning about alternative temporal trajectories (anticipation, counterfactual, hypothetical), while offering smaller improvements for questions that primarily test factual understanding (descriptive) or goal-oriented reasoning (planning).

B. Details of Digital Twin Representations

In this section, we describe the structure of our digital twin representation. Specifically, it includes a global scene summary, a spatial trajectory description, per-object frame-level captions, and numerical traces including area curves, depth estimates, and centroid movements. Formally, as shown in Fig. 5, each digital twin representation is represented as a JSON object containing the following components: (1) a `summary` describing the overall scene, (2) a `spatial_summary` explaining object motion and spatial behavior across the sequence, (3) a `major_elements` with per-frame annotations and numerical attributes, and (4) a `frame_range` denoting the temporal span.

However, such digital twin representation is expressive and large. Therefore, for the diffusion model fine-tuning and inference, we introduce a condensed version of digital twin representation that retains only most relevant elements. The condensed form preserves the global summary, the spatial description, and a compressed set of object attributes, while removing redundant frames, long numerical traces, and other high-granularity metadata.

C. Details for Editing the Digital Twin Representations

In this section, we present example for the edited digital twin representation. Given an initial digital twin representation as input, the LLM does more than rewrite the global `summary` and `spatial_summary`: it also generates a physically coherent update to the underlying motion cues, object trajectories, and depth evolution. Moreover, LLM adjusts frame-level descriptions, modifies object deformation patterns, and refines the numerical signals that govern

Original Digital Twin Representation

{"summary": "A pink lotus flower sways gently in a pond.",
 "spatial_summary":"Object 0 (pink lotus flower) stays in the center foreground, drifting only slightly right as it opens and closes. Its size stays constant, depth shifts are mild, and the centroid barely moves. The whole pattern matches a lotus gently swaying on pond water.",
 "major_elements": [
    {"object_0": "pink lotus flower",
     "frames": {"00000": {"description": "A pink lotus flower blooms gracefully on a green stem against a blurred natural background."},
                "00040": {"description": "A pink lotus flower blooms gracefully on a green stem against a blurred green leaves."},
     "area": [...] ,
     "depth": [...] ,
     "center": [...] }],
  "frame_range": {
           "start": 0,
           "end": 80}
}

Condensed Digital Twin Representation

"This JSON describes a video as a digital twin representation. SCENE_INFO gives total frames and 3x3 grid cut-lines forming regions TL,T,TR,L,C,R,BL,B,BR. Each object has: REGIONS (grid cell spans), PATH (8-way compass with counts), DEPTH (N=near, M=mid, F=far) spans, SCALE (^ grow, = steady, v shrink) spans. summary and spatial_sum provide scene context and relations. Frame indices in spans are actual frame numbers, computed from the source sampling rate."

{"summary": "A pink lotus flower sways gently in a pond."
 "spatial_summary":"Object 0 (pink lotus flower) stays in the center foreground, drifting only slightly right as it opens and closes. Its size stays constant, depth shifts are mild, and the centroid barely moves. The whole pattern matches a lotus gently swaying on pond water.",
 "major_elements": [
    {"object_0": "pink lotus flower",
     "frames": {"00000": {"description": "A pink lotus flower blooms gracefully on a green stem against a blurred natural background."},
                "00040": {"description": "A pink lotus flower blooms gracefully on a green stem against a blurred green leaves."},
     "regions": "L@0-0 > R@4-4 > BR@8-8 > L@12-12 > T@16-16 > L@20-20",
     "path": "Ex1 SEx1 Wx1 NEx1 SWx1" ,
     "depth": "F@0-0 > N@4-8 > M@12-16 > F@20-20"}],
  "frame_range": {
           "start": 0,
           "end": 80}
}

Edited Condensed Twin Representation

"This JSON describes a video as a digital twin representation.SCENE_INFO gives total frames and 3x3 grid cut-lines forming regions TL,T,TR,L,C,R,BL,B,BR. Each object has: REGIONS (grid cell spans), PATH (8-way compass with counts), DEPTH (N=near, M=mid, F=far) spans, SCALE (^ grow, = steady, v shrink) spans. summary and spatial_sum provide scene context and relations. Frame indices in spans are actual frame numbers, computed from the source sampling rate."

{"summary": "An hour after the snowfall, the pink lotus flower remains damp and slightly wilted, surrounded by cold mist rising from the pond."
 "spatial_summary":"Object 0 (the pink lotus flower) stays near the center foreground but shows a heavier, slower sway. Its area decreases slightly as petals curl inward, retaining moisture from melting snow. Depth values rise marginally, suggesting the flower leans closer to the water surface. The pond reflects pale light under a dim sky, with thin patches of ice and lingering droplets on nearby leaves, implying the temperature remains just above freezing.",
 "major_elements": [
    {"object_0": "pink lotus flower",
     "frames": {"00000": {"description": "A pink lotus flower stays near the center and sways slowly, getting a bit smaller as its petals curl in and it leans closer to the water. "},
                "00040": {"description": "A pink lotus flower remains damp and surrounded by a frozen pond. The pond shows dim light, a few thin ice patches, and droplets on the leaves, meaning the temperature is just above freezing."}},
     "regions": "L@0-0 > C@4-4 > C@8-8 > R@12-12 > C@16-16 > R@20-20",
     "path": "Ex1 Ex1 SEx1 Ex1 Wx1" ,
     "depth": "M@0-0 > M@4-8 > M@12-16 > M@20-20"}],
  "frame_range": {
           "start": 0,
           "end": 80}
}

Figure 5. Evolution of digital twin representations through the CWMDT. **Left**: Original digital twin representation extracted from video, containing per-frame descriptions, numerical traces for area, depth, and centroid coordinates. **Middle**: Condensed representation retaining scene summaries and compressed spatial attributes through compact notation for regions, motion paths, and depth spans. **Right**: LLM-edited representation reflecting the counterfactual intervention. The LLM modifies not only textual descriptions but also spatial trajectories, depth evolution, and motion patterns to maintain physical coherence under the hypothetical condition.

Table 4. Performance comparison on CausalVQA debug dataset. All values represent accuracy (%). Each category contains 40 question pairs. The best model performance in each category is shown in **bold**, with human performance in *italics*.

| | | Large/Closed | | Open (7-8B) | | | | | |
| | | GPT-4o | Gemini 2.5 Flash | InternVL2.5 | LLaVA-OneVision | PerceptionLM | Qwen2.5VL | **Qwen2.5VL+CWMDT** | *Human* |
Category	Difficulty								
Anticipation	Easy	50.00	55.00	52.50	27.50	47.50	47.50	**62.50**	*86.00*
Counterfactual	Easy	65.00	**70.00**	50.00	57.50	60.00	52.50	**70.00**	*93.12*
Descriptive	Easy	70.00	**80.00**	50.00	60.00	55.00	65.00	67.50	*90.50*
Hypothetical	Easy	**72.50**	67.50	67.50	60.00	67.50	40.00	**72.50**	*88.75*
Planning	Easy	65.00	**70.00**	**70.00**	52.50	70.00	57.50	67.50	*93.00*

area traces, centroid drift, and depth scaling. As a result, the edited digital twin representation is not merely a textual reinterpretation of the original scene, but a fully revised spatiotemporal representation that reflects the hypothetical or counterfactual conditions requested by the user. This edited form serves as a self-consistent, semantically aligned counterpart to the input twin, enabling downstream models to reason about alternative scene states with accurate and coherent structural detail.

References

[1] Eloi Alonso, Eloi Valevski, Vincent Micheli, and François Fleuret. Statespacediffuser: Bringing long context to diffusion world models. *arXiv preprint arXiv:2410.22849*, 2024. 2

[2] Shuai Bai et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3, 4, 10

[3] Zhipeng Bao, Anurag Bagchi, Yu-Xiong Wang, Pavel Tokmakov, and Martial Hebert. Video diffusion models learn the structure of the dynamic world. 2024. 2

[4] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22563–22575, 2023. 1, 2

[5] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18392–18402, 2023. 5, 7, 8

[6] Jake Bruce et al. Genie: Generative interactive environments. *arXiv preprint arXiv:2402.15391*, 2024. 1, 2

[7] Francesca Burgert, Vincent Sitzmann, and Jacob Steinhardt. Go-with-the-flow: Motion-controllable video diffusion models. *CVPR*, 2025. 2

[8] Weifeng Chen et al. Control-a-video: Controllable text-to-video diffusion models with motion prior and reward feedback learning. *arXiv preprint arXiv:2305.13840*, 2023. 2

[9] Weipu Chen et al. Storm: Efficient stochastic transformer based world models for reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2023. 2

[10] Zhe Chen et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 10

[11] Jang Hyun Cho et al. Perceptionlm: Open-access data and models for detailed visual understanding. *arXiv preprint arXiv:2504.13180*, 2025. 10

[12] Oriol Comas, Alex Pareja-Zubieta, Marc Gil-Ortin, Miquel Gonz'alez-Duque, Antonio Agudo, and Francesc Moreno-Noguer. Learning object-centric dynamic modes from video and emerging properties. In *Conference on Lifelong Learning Agents (CoLLAs)*, 2023. 1, 2

[13] Fei Deng, Ingook Tsang, and Ran Chen. Facing off world model backbones: Rnns, transformers, and s4. *Advances in Neural Information Processing Systems*, 37, 2024. 2

[14] Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, Jie Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, et al. Understanding world or predicting future? a comprehensive survey of world models. *ACM Computing Surveys*, 58(3):1–38, 2025. 1

[15] Tuo Feng, Wenguan Wang, and Yi Yang. A survey of world models for autonomous driving. *arXiv preprint arXiv:2501.11260*, 2025. 2

[16] Aaron Foss, Chloe Evans, Sasha Mitts, Koustuv Sinha, Ammar Rizvi, and Justine T Kao. Causalvqa: A physically grounded causal reasoning benchmark for video models. *arXiv preprint arXiv:2506.09943*, 2025. 10

[17] Zigang Geng, Binxin Yang, Tiankai Hang, Chen Li, Shuyang Gu, Ting Zhang, Jianmin Bao, Zheng Zhang, Houqiang Li, Han Hu, et al. Instructdiffusion: A generalist modeling interface for vision tasks. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 12709–12720, 2024. 5, 7, 8

[18] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3), 2018. 1, 2

[19] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, et al. Ltx-video: Realtime video latent diffusion. *arXiv preprint arXiv:2501.00103*, 2024. 1, 2, 4

[20] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019. 1

[21] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023. 2

[22] Paul Henderson and Christoph H Lampert. Unsupervised object-centric video generation and decomposition in 3d. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3106–3117, 2020. 1, 2

[23] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7612–7624, 2021. 7

[24] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. In *Advances in Neural Information Processing Systems*, 2022. 1

[25] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 5

[26] Xiaofeng Hu et al. Drivedreamer: Towards real-world-driven world models for autonomous driving. *European Conference on Computer Vision*, pages 646–662, 2024. 2

[27] Jianxiong Huang et al. Pandora: Towards general world model with natural language actions and video states. *arXiv preprint arXiv:2406.09455*, 2024. 2

[28] Lukasz Kaiser, Mohammad Babaeizadeh, Piotr Milos, Blazej Osinski, Roy H Campbell, Konrad Czechowski, Dumitru Erhan, Chelsea Finn, Piotr Kozakowski, Sergey Levine, et al. Model-based reinforcement learning for atari. *arXiv preprint arXiv:1903.00374*, 2019. 1

[29] Lara Kirfel, Robert J MacCoun, Thomas Icard, and Tobias Gerstenberg. Anticipating the risks and benefits of counterfactual world simulation models. In *AI Meets Moral Philosophy and Moral Psychology Workshop (NeurIPS 2023)*, 2023. 1

[30] Alexander Kirillov et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 3

[31] Max Ku et al. Anyv2v: A tuning-free framework for any video-to-video editing tasks. *arXiv preprint arXiv:2403.14468*, 2024. 5, 7, 8

[32] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 10

[33] Guangzhao Li, Yanming Yang, Chenxi Song, and Chi Zhang. Flowdirector: Training-free flow steering for precise text-to-video editing. *arXiv preprint arXiv:2506.05046*, 2025. 5, 7, 8

[34] Minghan Li et al. Five-bench: A fine-grained video editing benchmark for evaluating emerging diffusion and rectified flow models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16672–16681, 2025. 7

[35] Xinqing Li, Xin He, Le Zhang, and Yun Liu. A comprehensive survey on world models for embodied ai. *arXiv preprint arXiv:2510.16732*, 2025. 2

[36] Bingzhi Liu, Dongchen Yang, Tianyi Zhang, Mingzhe Chen, Hao Wang, Yi Zhou, et al. How far is video generation from world model: A physical law perspective. *arXiv preprint arXiv:2411.02385*, 2024. 2

[37] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 426–443. Springer, 2023. 7

[38] Shenyuan Liu, Jiaxin Fan, Yihang Zhang, Dongxiang Zhang, et al. Vista: A generalizable driving world model with high fidelity and versatile controllability. *Advances in Neural Information Processing Systems*, 38, 2024. 2

[39] Yu Lu, Linchao Zhu, Hehe Fan, and Yi Yang. Flowzero: Zero-shot text-to-video synthesis with llm-driven dynamic scene syntax. In *arXiv preprint arXiv:2311.15813*, 2023. 1

[40] Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world models. *International Conference on Learning Representations (ICLR)*, 2023. 2

[41] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[42] Saman Motamed et al. Do generative video models understand physical principles? In *arXiv preprint arXiv:2501.09038*, 2025. 1, 2

[43] Haomiao Ni, Changhao Shi, Kai Li, Sharon X. Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7951–7961, 2023. 1

[44] OpenAI. Video generation models as world simulators. *OpenAI Technical Report*, 2024. 2

[45] William Peebles and Saining Xie. Scalable diffusion models with transformers. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 2

[46] Bosheng Qin, Juncheng Li, Siliang Tang, Tat-Seng Chua, and Yueting Zhuang. Instructvid2vid: Controllable video editing with natural language instructions. *arXiv preprint arXiv:2305.12328*, 2024. 5, 7, 8

[47] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3

[48] Marc Rigter, Tarun Gupta, Agrin Hilmkil, and Chao Ma. Avid: Adapting video diffusion models to world models. *Reinforcement Learning Conference (RLC)*, 2025. 2

[49] Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020. 1, 2

[50] Yiqing Shen, Hao Ding, Lalithkumar Seenivasan, Tianmin Shu, and Mathias Unberath. Position: Foundation models need digital twin representations. *arXiv preprint arXiv:2505.03798*, 2025. 2

[51] Yiqing Shen, Chenjia Li, Chenxiao Fan, and Mathias Unberath. Rvtbench: A benchmark for visual reasoning tasks. *arXiv preprint arXiv:2505.11838*, 2025. 5

[52] Yiqing Shen, Chenjia Li, and Mathias Unberath. Text-driven reasoning video editing via reinforcement learning on digital twin representations. *arXiv preprint arXiv:2511.14100*, 2025. 6

[53] Yiqing Shen, Bohan Liu, Chenjia Li, Lalithkumar Seenivasan, and Mathias Unberath. Online reasoning video segmentation with just-in-time digital twins. *arXiv preprint arXiv:2503.21056*, 2025. 2, 3

[54] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *arXiv preprint arXiv:2209.14792*, 2022. 1, 2

[55] Eloi Valevski, Eloi Alonso, and François Fleuret. Diffusion world models. *arXiv preprint arXiv:2402.03570*, 2024. 2

[56] Wan-AI. Wan2.2: Open and advanced large-scale video generative models, 2025. 1, 2

[57] Chen Min Wang et al. Driveworld: 4d pre-trained scene understanding via world models for autonomous driving. *arXiv preprint arXiv:2405.04390*, 2024. 2

[58] Chenfei Wu et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025. 5

[59] Philipp Wu, Alejandro Escontrela, Danijar Hafner, Pieter Abbeel, and Ken Goldberg. Daydreamer: World models for physical robot learning. In *Conference on robot learning*, pages 2226–2240. PMLR, 2023. 1

[60] Zhangkai Wu, Xuhui Fan, Zhongyuan Xie, Kaize Shi, Zhidong Li, and Longbing Cao. Fame: Fairness-aware attention-modulated video editing. *arXiv preprint arXiv:2510.22960*, 2025. 7

[61] Lihe Yang, Bingyi Kang, Zilong Huang, et al. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 3

[62] Shen Yang, Yuchen Zhang, Ziwei Liu, Chen Change Loy, and Bo Dai. Sora as an agi world model? a complete survey on text-to-video generation. *arXiv preprint arXiv:2403.05131*, 2024. 2

[63] Shoubin Yu et al. Veggie: Instructional editing and reasoning of video concepts with grounded generation. *arXiv preprint arXiv:2503.14350*, 2025. 7

[64] Yu Yuan, Xijun Wang, Tharindu Wickremasinghe, Zeeshan Nadir, Bole Ma, and Stanley H Chan. Newtongen: Physics-consistent and controllable text-to-video generation via neural newtonian dynamics. *arXiv preprint arXiv:2509.21309*, 2025. 2

[65] Chang Zhang, Rui Chen, Mengqi Xu, and Jian Shi. Transdreamer: Reinforcement learning with transformer world models. *arXiv preprint arXiv:2202.09481*, 2022. 2

[66] Lianmin Zheng, Wei-Lin Chiang, Yonghao Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, pages 46595–46623, 2023. 7