ST-GEN4D: EMBEDDING 4D SPATIOTEMPORAL COGNITION INTO WORLD MODEL FOR 4D GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative models have achieved success in producing apparently coherent 2D videos, but remain challenging in the physical world due to lack of 4D spatiotemporal scale. Typically, existing 4D generative models directly embed macro scale constraints to enhance overall spatiotemporal consistency. However, these methods only ensure global appearance coherence and fail to reveal the local dynamics of the physical world. Our insight is that global appearance structure and local dynamic topology empower 4D spatiotemporal cognition, thereby enabling 4D generation with spatiotemporal regularities. In this work, we propose ST-Gen4D, a 4D generation framework with 4D spatiotemporal cognition-based world model. Our model is guided by four key designs: 1) Spatiotemporal representation. We encode various modalities into multiple representations as a feature basis. 2) Spatiotemporal cognition. We sculpture these representations into global appearance graph and local dynamic graph, and fuse them via semantic-bridged spatiotemporal fusion to obtain a 4D cognition graph. 3) Spatiotemporal reasoning. We utilize a world model to derive future state based on the 4D cognition. 4) Spatiotemporal generation. We leverage the derived cognition as condition to guide latent diffusion for 4D Gaussian generation. By deeply integrating 4D intrinsic cognition with generative priors, our model guarantees the structural rationality and topological consistency of 4D generation. Moreover, we propose ST-4D datasets by aggregating public 4D datasets and self-built subset. Extensive experiments demonstrate the superiority of our ST-Gen4D across 3D and 4D generation tasks. Our code will be made publicly available once the paper is accepted.

1 INTRODUCTION

Generative models have achieved remarkable progress in 2D video generation by learning the distribution mapping between textual prompts and spatiotemporal visual signals (Ho et al., 2022; Singer et al., 2022; Blattmann et al., 2023; Hong et al., 2022; Villegas et al., 2022; Brooks et al., 2024). However, lifting such generative capability to the 4D physical world remains challenging, since 2D generative priors mainly capture frame-level appearance evolution but lack explicit modeling of 4D spatiotemporal scale. In the physical world, dynamic scenes are continuous 4D processes governed by spatial-temporal regularities rather than independent visual frames. Thus, coherent 4D generation across space and time requires deeper spatiotemporal understanding.

Existing 4D generation methods typically introduce macro-scale constraints, such as 4D-aware diffusion priors (Jiang et al., 2023; Zhao et al., 2023; Zhu et al., 2025) or smoothness regularizations (Ling et al., 2024a; Bahmani et al., 2024; Lyu et al., 2026), to enhance overall spatiotemporal consistency. Although these constraints improve global coherence, they are usually imposed externally and mainly regularize generated results at a coarse level. Thus, they remain insufficient to reveal the intrinsic organization of dynamic 4D scenes, especially how overall appearance structures are preserved and how local regions move, deform, and interact over time. This motivates us to move beyond macro-scale constraints and introduce an intrinsic spatiotemporal cognition for 4D generation.

In this work, we define 4D spatiotemporal cognition as a structured state that encodes global appearance structure and local dynamic topology. The former describes spatial organization and cross-view consistency, while the latter characterizes temporal evolution and cross-time coherence. As shown in Fig. 1, this design is driven by two key insights: 1) Spatiotemporal encoders capture

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.

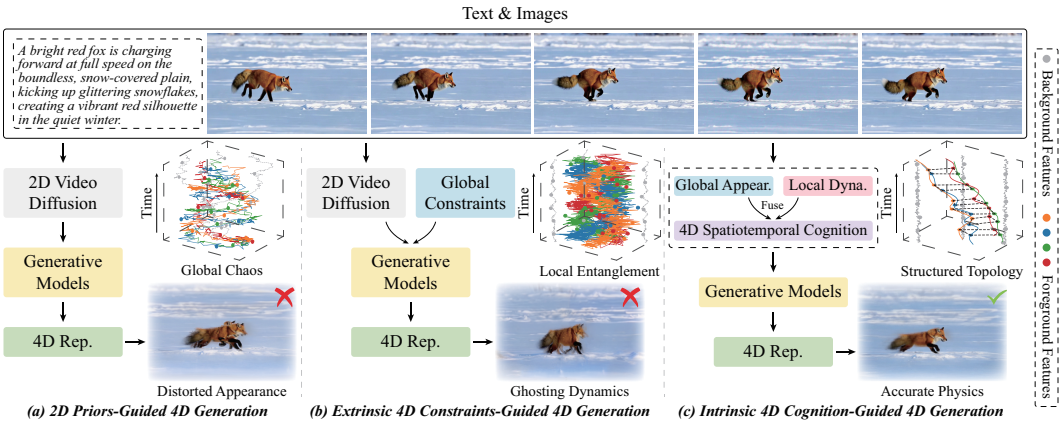


Figure 1: **Paradigm shift of 4D generation.** 2D Priors-guided generation mainly relies on 2D video diffusion, leading to collapsed appearance. Extrinsic constraints-guided generation introduces macro scale constraints but suffers from motion artifacts. Our intrinsic 4D cognition-guided paradigm leverages a spatiotemporal cognition to ensure plausible appearance and coherent dynamics.

deep visual correlations to support robust spatial alignment and cross-time continuity. 2) Cognition graphs impose structural and topological constraints to preserve semantic boundaries and mitigate feature entanglement during complex object interactions. More importantly, unlike static conditions or external regularization terms, our cognition can be evolved by a world model to infer future spatiotemporal states, which are further used to condition the subsequent 4D generation process.

Based on this idea, we propose **ST-Gen4D**, a 4D generation framework with a spatiotemporal cognition-based world model. Our framework consists of four stages: 1) Spatiotemporal foundational representation. We introduce the 4D-VGGT (Wang et al., 2025a) encoder to extract robust features for cross-view consistency and temporal continuity. 2) Spatiotemporal structural cognition. We construct a global appearance graph and a local dynamic graph, and fuse them to obtain spatiotemporal cognition. 3) Spatiotemporal predictive reasoning. We utilize the world model to derive the evolution of dynamic scenes based on cognition. 4) Spatiotemporal consistent generation. We leverage the predicted cognition as the condition of a latent diffusion for 4D Gaussians generation, ensuring the appearance rationality and dynamics consistency. By integrating intrinsic 4D cognition with generative priors, ST-Gen4D achieves spatially stable and temporally coherent 4D generation.

Our main contributions are summarized as:

- We propose ST-Gen4D, a novel 4D generation framework guided by a 4D spatiotemporal cognition-based world model for superior spatial stability and temporal coherence in the physical world.
- Observing that spatiotemporal encoders ensure robust spatiotemporal consistency while cognition graphs provide structural constraints, we design a world model reasoning paradigm based on 4D cognition to accurately predict dynamic scene evolution.
- We develop a latent diffusion generation process guided by the future 4D spatiotemporal cognition inferred from the world model. Conditioning 4D Gaussian generation on these predicted states enforces appearance structure and dynamic topologic consistency.
- We aggregate a series of 4D datasets and build a validation subset, collectively named the spatiotemporal 4D dataset (**ST-4D**). Extensive experiments have demonstrated our significant effectiveness.

2 Related Works

4D Generation. Driven by the success of 2D video generation, recent efforts in 4D generation predominantly utilize 2D video diffusion models to optimize or reconstruct 4D scenes (Jiang et al., 2023; Zhao et al., 2023; Zhu et al., 2025; Bahmani et al., 2024; Lyu et al., 2026; Bahmani et al., 2025; Zheng et al., 2024; Liu et al., 2025). For instance, Consistent4D (Jiang et al., 2023) optimizes 4D scenes via SDS loss (Poole et al., 2022), whereas Lyra (Bahmani et al., 2025) directly reconstructs them from generated videos. Despite visual appeal, their reliance on 2D priors lacks 4D spatiotempo-

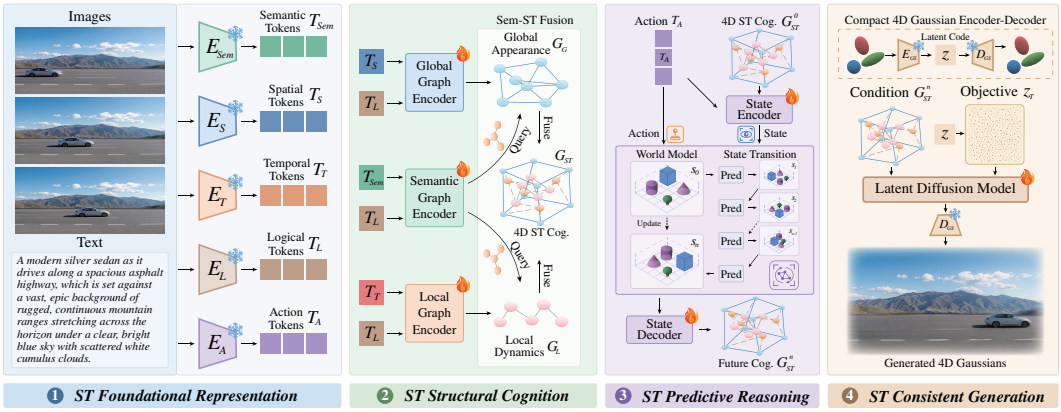


Figure 2: **Overview of ST-Gen4D.** The model first obtains multiple cognition tokens, which are then encoded into structured spatiotemporal cognition graph. We then utilize an autoregressive world model to predict state transition based on the cognition and action tokens. Finally, the deduced state severs as the condition of latent diffusion to generate 4D Gaussians.

ral scale, resulting in spatial structural distortions and temporal incoherence. We aim to introduce intrinsic 4D spatiotemporal modeling to significantly enhance spatiotemporal representations.

Spatiotemporal-based 4D Generation. Existing 4D generation methods typically introduce macro-scale constraints like 4D-aware diffusion (Jiang et al., 2023; Zhao et al., 2023; Zhu et al., 2025) or smoothness regularizations (Ling et al., 2024a; Bahmani et al., 2024; Lyu et al., 2026) to enhance overall spatiotemporal consistency. Although these approaches improve macroscopic appearance consistency, they struggle to accurately model local dynamics in the physical world. This is because applying uniform global regularizations causes spatial-temporal over-smoothing, thereby ignoring complex dynamics. Thus, we introduce a structured spatiotemporal cognition graph to simultaneously comprehend global appearance structures and local dynamic topologies.

World Model. Achieving 4D generation with plausible appearance and coherent dynamics relies on the accurate spatiotemporal deduction of the physical world. While world models (LeCun et al., 2022; Bardes et al., 2024; Maes et al., 2026) excel at predicting these transitions, existing frameworks restrict them to inferring superficial pixel-level appearances across views and times (Lu et al., 2025; Chen et al., 2025b; Yang et al., 2026; Team et al., 2026). This neglects their intrinsic capacity to predict future structural relationships. Consquently, we aim to leverage the world model to simulate the transition process of 4D spatiotemporal cognition, thus accurately predicting the appearance structure and dynamic topology evolution of physical world.

3 Our ST-Gen4D

Overview. As shown in Fig. 2, our ST-Gen4D achieves structurally-rational and topologically-consistent 4D generation through the following four stages:

1) **Spatiotemporal Foundational Representation.** We extract multiple tokens $(T_{Sem}, T_S, T_T, T_L, T_A)$ from input image sequences I and texts T as a feature basis:

$$T_{Sem}, T_S, T_T, T_L, T_A = E_{Found}(I, T). \tag{1}$$

2) **Spatiotemporal Structural Cognition.** Semantic-bridged spatiotemporal fusion (Sem-STF) fuses independent topological graphs encoded from tokens into a spatiotemporal cognition graph G_{ST}:

$$G_{Sem}, G_S, G_T = \text{GE}(T_{Sem,S,T}, T_L), G_{ST} = \text{Sem-STF}(G_{Sem}, G_S, G_T). \tag{2}$$

3) **Spatiotemporal Predictive Reasoning.** An autoregressive world model uses the initial cognition graph G^t_{ST} and action tokens T_A to infer a continuous sequence of future cognitive states $\hat{G}^{t+1:T}_{ST}$ with strict kinematic priors:

$$\hat{G}^{t+1:T}_{stc} = WM(G^t_{ST}, T_A). \tag{3}$$

4) **Spatiotemporal Consistent Generation.** The predicted spatiotemporal cognition steers a latent diffusion process, which is then decoded into 4D Gaussians $\hat{\mathcal{G}}_{4D}$:

$$Z_{4D}^{(0)} = \text{CG-DiT}(Z_{4D}^{(T)} \mid \hat{G}_{ST}^{t+1:T}), \quad \hat{\mathcal{G}}_{4D} = D_{GS}(Z_{4D}^{(0)}). \tag{4}$$

Remarks. The output $\hat{\mathcal{G}}_{4D}$ denotes the generated 4D Gaussians representing the dynamic scene. Spatiotemporal cognition graphs provide structural conditions for structurally-rational and topologically-consistent 4D generation in latent diffusion models.

3.1 Spatiotemporal Foundational Representation

We use semantic, spatial, temporal, logical, and action encoders to extract distinct representations for cognition construction and world model reasoning.

Spatial and Temporal Encoders. For robust 4D structural cognition, we extract visual semantics, global appearance, and local dynamics. Wan2.1 3D Causal VAE (Wan et al., 2025) extracts semantic tokens. In Fig. 3, 4D-VGGT (Wang et al., 2025a) overcomes the feature drift of ResNet50 (He et al., 2016), providing strict cross-view and cross-time consistency. We thus adopt 4D-VGGT as core encoders. Sharing a DINOv2 backbone, the spatial encoder integrates multi-view appearance via pair-wise cross-attention, while the temporal encoder captures local dynamics via sliding window attention (Wang et al., 2025a).

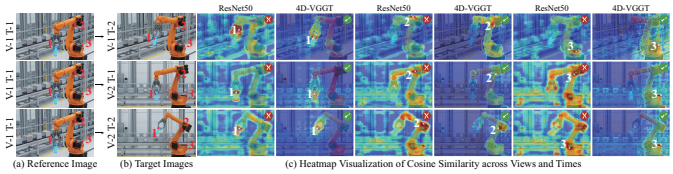


Figure 3: Cross-view and cross-time consistency analysis.

Logical and Action Encoders. Logical and action encoders bridge spatiotemporal cognition and provide latent dynamic cues for reasoning. The logic encoder extracts image and text features via CLIP ViT and Text encoders (Radford et al., 2021), then maps them through an MLP into unified logical tokens. The action encoder uses InternVideo2 (Wang et al., 2024) and multimodal alignment to distill inputs into tokens enriched with action priors. Logical representations solidify cognition, while action priors facilitate spatiotemporal inference.

3.2 Spatiotemporal Structural Cognition

High-fidelity 4D generation requires guidance equipped with global appearance and local dynamics. In Fig. 4, simple semantic features often fail to distinguish entities during complex interactions, whereas structured spatiotemporal features maintain high discriminability. This motivates our structured spatiotemporal cognition in Fig. 5 for robust inference.

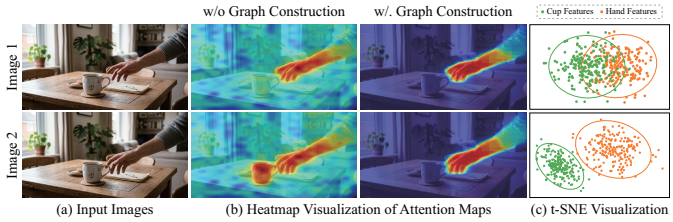


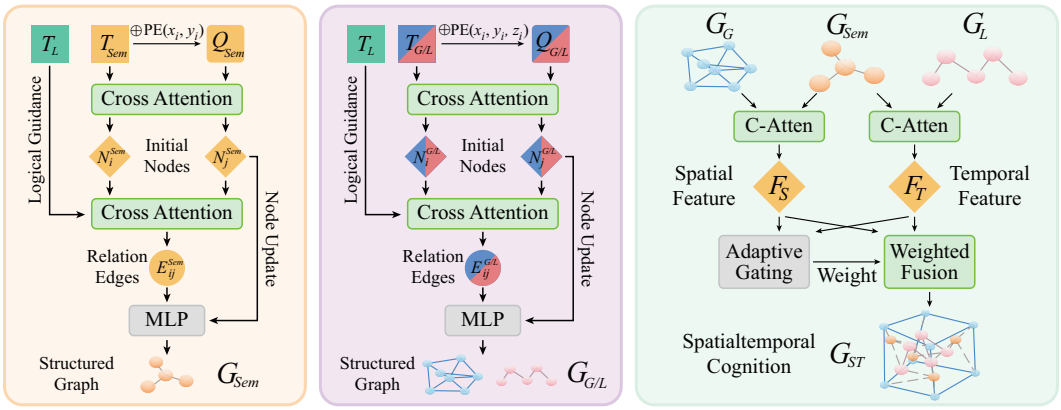
Figure 4: Feature robustness and discriminability analysis.

Semantic-Global-Local Graph Encoder. Three graphs encode semantics, global appearance, and local dynamics through a unified paradigm to capture comprehensive visual cues.

Semantic Graph Construction. Semantic query Q_{Sem} concatenates T_{Sem} with a 2D positional embedding $\text{PE}(x_p, y_p)$. Cross-attention between Q_{Sem} and T_{Sem} yields n initial nodes N_i^{Sem}. Logical tokens T_L guide semantic edges E_{ij}^{Sem} between nodes N_i^{Sem} and N_j^{Sem}. An MLP aggregates messages m_{ij}^{Sem} to update nodes into semantic graph G_{Sem}:

$$m_{ij}^{Sem} = \text{MLP}(N_i^{Sem}, N_j^{Sem}, E_{ij}^{Sem}), \quad G_{Sem} = N_i^{Sem} + \sum_j m_{ij}. \tag{5}$$

Global Appearance and Local Dynamic Graph Construction. We construct global appearance graph G_G and local dynamics graph G_L similarly, replacing semantic tokens with spatial and temporal tokens, respectively. Replacing the 2D positional embedding with a 3D geometry embedding $\text{GE}(x_p, y_p, z_p)$ incorporates 3D geometric awareness.



a) Semantic Graph Cons. b) Global-Local Graph Cons. c) Semantic-Guided Spatiotemporal Fusion

Figure 5: **Architectural details of the spatiotemporal structured cognition.** (a, b) The construction pipelines for semantic, global, and local graphs from foundational representations. (c) Details of the fusion progress between global and local graphs guided by common semantic graph.

Semantic-Bridged Spatiotemporal Fusion. To aggregate these isolated graphs, we design a semantic-bridged spatiotemporal fusion module. Treating the semantic graph G_{Sem} as a query, we apply dual-stream cross-attention to the global appearance graph G_G and local dynamics graph G_L to extract aligned features. An adaptive gating mechanism then generates dynamic weights, selectively aggregating these spatial and temporal streams into the semantic anchor via a feed-forward network. The resulting spatiotemporal cognition graph G_{ST} provides a robust state for subsequent inference.

3.3 Spatiotemporal Predictive Reasoning

Our cognition-based world model extrapolates continuous future dynamics from the current spatiotemporal cognition graph and extracted action tokens.

World Action-State Formulation. We distill the expansive cognition graph G_{ST} into a compact state s_t via cross-attention with learnable queries. Concurrently, action tokens T_A are processed through temporal attention pooling and an MLP to form a dynamic condition vector a_t. Using adaptive layer normalization (adaLN), a_t generates affine parameters to modulate s_t into an action-conditioned state $s_t^{'}$. This integration enforces kinematic constraints during autoregressive modeling.

Cognition-based World Model Reasoning. A causal Transformer predictor inspired by LeWorld-Model (Maes et al., 2026) processes the historical sequence $[s_{t-T}^{'}, \ldots, s_t^{'}]$ to infer the next latent state $\hat{s}_{t+1}$. A cross-attention state resampling decoder projects it into cognition space, using original graph G_{ST}^t as queries and $\hat{s}_{t+1}$ as keys and values to iteratively construct future cognition $\hat{G}_{ST}^{t+1}$. Future states preserve rigorous spatiotemporal dynamics to guide physically plausible 4D generation.

3.4 Spatiotemporal Consistent Generation

Cognition-guided latent diffusion and a latent 4D Gaussian encoder-decoder generate high-fidelity, physically consistent 4D scenes conditioned on predicted spatiotemporal cognition graphs.

Cognition-Guided Latent Diffusion. A cognition-guided diffusion Transformer (CG-DiT) trained via flow-matching generates 4D scene latents. Stacked DiT blocks process an initial noise sequence $z_{4D}^{(T)}$, modulated by timestep τ via adaLN. Structured cross-attention conditions intermediate features on the predicted graph $\hat{G}_{ST}$ as keys and values. A projection head progressively denoises the sequence to the clean latent $z_{4D}^{(0)}$, ensuring plausible spatial appearances and temporal coherence.

Latent 4D Gaussian Encoder-Decoder. A pre-trained KL-regularized autoencoder compresses massive 4D Gaussians into tractable latents. Its sparse 4D convolutional encoder E_{GS} voxelizes Gaussian points into a compact latent sequence z_{4D}. A lightweight decoder D_{GS} uses self-attention and parallel MLPs to reconstruct spatial attributes and polynomial deformation fields, reducing the overhead of generating millions of Gaussian splats and bridging explicit 4D rendering and diffusion.

3.5 OPTIMIZATION

We train ST-Gen4D across the progressive stages using three loss functions:

World Model Loss ($\mathcal{L}_{WM}$). To learn accurate state transitions in a robust latent space, we combine an MSE prediction loss with a SIGReg term to prevent representation collapse(Maes et al., 2026):

$$\mathcal{L}_{WM} = \mathbb{E}_t \left[|\hat{s}_{t+1} - s_{t+1}|_2^2 \right] + \alpha \mathcal{L}_{SIGReg}(s), \tag{6}$$

where z_{t+1} and $\hat{z}_{t+1}$ are ground-truth and predicted future latents, and α weights regularization.

SDS Loss ($\mathcal{L}_{SDS}$). Score Distillation Sampling (SDS) from a pre-trained multi-view or video diffusion model supervises appearance using generative priors:

$$\mathcal{L}_{SDS}(\phi, x) = \mathbb{E}_{t,\epsilon,c} \left[w(t) \|\hat{\epsilon}_\phi(x_t; t, c) - \epsilon\|_2^2 \right], \tag{7}$$

where ϵ is our predicted noise and $\hat{\epsilon}_\phi$ denotes the frozen pre-trained noise predictor. x_t is the noisy latent at diffusion timestep t, and c represents the images and texts conditioning prompts.

Spatiotemporal Loss ($\mathcal{L}_{ST}$): To equip the model with spatiotemporal perception, we introduce a loss combining Chamfer Distance (CD) for spatial supervision, temporal smoothness for deformation consistency, and DDIM-based reconstruction for rendering fidelity:

$$\mathcal{L}_{ST} = \mathcal{L}_{CD}(\hat{\mathcal{G}}, \mathcal{G}) + \beta \sum_t \|\Delta\hat{\mathcal{G}}_t - \Delta\hat{\mathcal{G}}_{t-1}\|_2^2 + \gamma \mathcal{L}_{DDIM}(I(\hat{\mathcal{G}}), I(\mathcal{G})), \tag{8}$$

where $\hat{\mathcal{G}}$ and $\mathcal{G}$ are the generated and ground-truth 4D Gaussians, $\Delta\hat{\mathcal{G}}_t$ denotes vertex displacement between adjacent Gaussians, $I(\cdot)$ represents rendered images, and β, γ are weighting factors.

The total objective is:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{WM} + \lambda_2 \mathcal{L}_{SDS} + \lambda_3 \mathcal{L}_{ST}, \tag{9}$$

where $\lambda 1$, λ_2, and λ_3 weight state prediction, appearance alignment, and spatiotemporal refinement.

4 TRAINING DATASETS AND PIPELINE

4.1 OUR ST-4D DATASET

As shown in Fig. 6, we propose **ST-4D**, a comprehensive spatiotemporal 4D dataset that integrates public sources and self-built data generated from Genie-3 (Bruce et al., 2024), which is designed to support cognition-based 4D generation in a unified formulation. It aggregates diverse 3D spatial datasets (Baruch et al., 2021; Dai et al., 2017; Yeshwanth et al., 2023; Deitke et al., 2023; Zhou et al., 2018; Li & Snavely, 2018; Ling et al., 2024b; Fu et al., 2021) and 4D spatiotemporal datasets (Liu et al., 2022; Karaev et al., 2023; Sun et al., 2020; Cabon et al., 2020; Wang et al., 2020; Black et al., 2023; Zheng et al., 2023; Wang et al., 2025b), covering global appearance, local dynamics, and long-term dynamic evolution, which provide essential priors for our whole framework. We standardize heterogeneous data in three steps: *Image prompt rendering* projects 3D/4D assets into temporally ordered 2D sequences as visual prompts. *Text prompt generation* employs Gemini3 to produce semantic descriptions for text-conditioned generation. *Unified Gaussian conversion* transforms diverse formats into consistent Gaussian representations for efficient differentiable rendering and diffusion training. ST-4D supports unified training across all four stages of our framework: spatiotemporal representation learning, cognition graph construction, world model reasoning, and cognition-conditioned 4D generation.

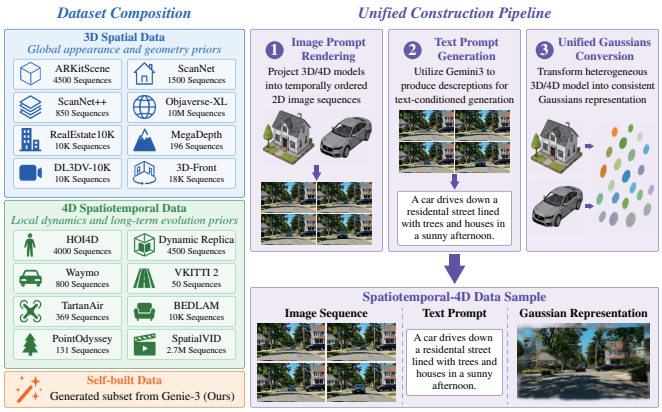


Figure 6: Composition and construction pipeline of ST-4D.

Table 1: Text-to-3D comparison on T3Bench.

	Method	T3Bench (He et al., 2023)			
		Single Obj.↑	Single w/ Surr.↑	Multi Obj.↑	Average↑
3D Generation	DREAMFUSION (Poole et al., 2022)	24.4	19.8	11.7	18.7
	Magic3D (Lin et al., 2023)	37	35.4	25.7	32.7
	SJC (Wang et al., 2023a)	24.7	19.8	11.7	18.7
	ProlificDreamer (Wang et al., 2023b)	49.4	44.8	35.8	43.3
	GaussianDreamer (Yi et al., 2024)	54	48.6	34.5	45.7
	Cog2Gen3D (Wang et al., 2026)	58.3	57.9	53.6	56.6
4D Generation	4Dfy (Bahmani et al., 2024)	52.6	45.1	36.1	44.6
	Dream-in-4D (Zheng et al., 2024)	55.2	50.3	42.4	49.3
	Ours	**58.9**	**58.7**	**55.2**	**57.6**

Table 2: Image-to-3D comparison on 3D-Front.

	Method	3D-Front (Fu et al., 2021)		
		Chamfer Distance ↓	F-Score ↑	IoU-B ↑
3D Generation	LucidDreamer (Chung et al., 2023)	0.138	39.99	0.299
	Gen3DSR (Ardelean et al., 2025)	0.123	40.07	0.363
	EchoScene (Zhai et al., 2024)	0.105	45.62	0.458
	Layout2Scene (Chen et al., 2025a)	0.094	48.36	0.492
	Cog2Gen3D (Wang et al., 2026)	0.063	58.43	0.682
4D Generation	Free4D (Liu et al., 2025)	0.114	42.85	0.412
	NeoVerse (Yang et al., 2026)	0.099	47.14	0.475
	WorldFM (Team et al., 2026)	0.078	53.46	0.587
	Ours	**0.051**	**62.15**	**0.713**



(a) Input Text (b) Dreamfusion (c) GaussianDreamer (d) Cog2Gen3D (e) Dream-in-4D (f) 4D-fy (g) Ours

Figure 7: Visualization of Text-to-3D generation on the T3Bench dataset.



(a) Input Image (b) LucidDreamer (c) Cog2Gen3D (d) WorldFM (e) Ours

Figure 8: Visualization of image-to-3D generation on the 3D-Front dataset.

4.2 Training Pipeline

Three stages progressively develop foundational understanding and consistent 4D generation:

Stage 1: Foundational Cognition-Evolution Learning. We fix the action encoder and use Wan2.1 (Wan et al., 2025) as the state encoder to optimize the predictor. We then freeze the predictor and train the state encoder with the spatiotemporal cognition graph as input, providing structurally consistent representations and supporting long-term state evolution.

Stage 2: Pixel Appearance Alignment. To learn high-fidelity visual perception and generative priors, we pre-train on large-scale videos using Score Distillation Sampling (SDS). This warm-up with multi-view diffusion models enables visually plausible scene generation.

Stage 3: Spatiotemporal Perception Refinement. Global fine-tuning using 3D/4D point cloud data refines spatiotemporal perception, ensuring the generated 4D scenes maintain plausible global appearance structure and coherent local dynamics.

5 Experiments

5.1 Implementation Details

We extract $n = 256$ nodes for each semantic, global, and local graph, balancing representational capacity and efficiency. Latent diffusion uses 1000 noise timesteps for flow-matching and 50 sampling steps. We use AdamW with an initial learning rate of 1×10^{-4}, a 0.01 weight decay, linear warmup, and cosine annealing. Internal loss weights in Eq. (6) and Eq. (8) are $\alpha = 0.1$, $\beta = 0.5$, and $\gamma = 1.0$. Final joint fine-tuning weights in Eq. (9) are $\lambda_1 = 1.0$, $\lambda_2 = 0.1$, and $\lambda_3 = 0.5$. We perform all model training and evaluation on a total of 8 NVIDIA A800 GPUs.

5.2 Comparison Experiments

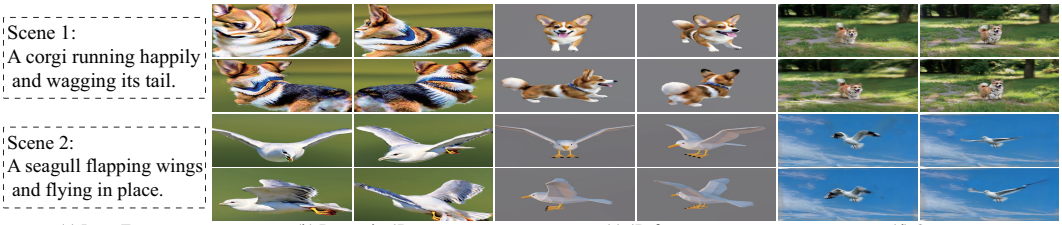
Text-to-3D Generation. We evaluate text-to-3D generation on T3Bench (He et al., 2023) across three levels against SOTA methods (Poole et al., 2022; Lin et al., 2023; Wang et al., 2023a;b; Yi et al., 2024; Wang et al., 2026; Bahmani et al., 2024; Zheng et al., 2024). ST-Gen4D achieves superior scores across all metrics, especially for challenging multi-object tasks (Table 1), and consistently

Table 3: Text-to-4D evaluation on VBench and User Study.

Method	VBench (Huang et al., 2024)				User Study			
	Text Align.↑	Consist.↑	Dyna.↑	Aesth.↑	Align.↑	Qual.↑	Consist.↑	Smooth.↑
AYG (Ling et al., 2024a)	22.35	89.41	36.45	43.27	6.91	7.45	7.82	7.55
4Dfy (Bahmani et al., 2024)	25.12	91.75	52.45	52.63	7.54	8.36	8.21	8.24
Dream-in-4D (Zheng et al., 2024)	24.64	90.23	53.72	53.14	7.78	8.15	8.52	8.63
4Real (Yu et al., 2024)	25.87	92.14	47.13	55.69	7.95	8.42	8.36	8.45
Free4D (Liu et al., 2025)	25.71	93.56	54.27	59.42	8.12	8.71	9.04	8.55
Ours	**36.92**	**97.43**	**67.75**	**69.84**	**8.79**	**9.56**	**9.67**	**9.28**

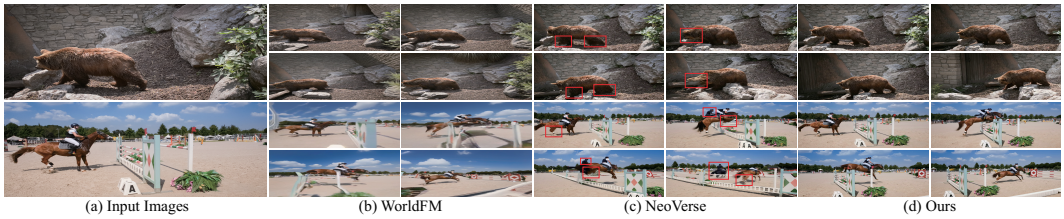
Table 4: Image-to-4D evaluation.

Method	VBench (Huang et al., 2024)			User Study			
	Consis.↑	Dyna.↑	Aesth.↑	Align.↑	Qual.↑	Consist.↑	Smooth.↑
GenXD (Zhao et al., 2024)	82.52	86.34	47.58	7.42	7.76	7.88	7.53
DimensionX (Sun et al., 2024)	89.14	89.73	52.39	8.21	8.18	8.57	8.49
Free4D (Liu et al., 2025)	92.35	94.38	56.87	8.52	8.83	9.05	8.68
4DNeX (Chen et al., 2025c)	85.43	90.53	49.27	7.95	8.26	8.34	8.11
One4D (Mi et al., 2025)	89.62	82.74	48.56	7.84	8.45	8.12	7.95
NeoVerse (Yang et al., 2026)	91.47	92.16	51.56	8.35	8.64	8.71	8.52
WorldFM (Team et al., 2026)	93.62	95.42	54.14	8.84	9.15	8.98	8.76
Ours	**97.83**	**99.74**	**59.37**	**9.46**	**9.72**	**9.85**	**9.12**



(a) Input Text (b) Dream-in-4D (c) 4D-fy (d) Ours

Figure 9: Visualization of text-to-4D generation on the ST-4D dataset.



(a) Input Images (b) WorldFM (c) NeoVerse (d) Ours

Figure 10: Visualization of image-to-4D generation on the DAVIS (Pont-Tuset et al., 2017) dataset.

generates detailed 3D scenes that avoid the blurred appearances and collapsed structures observed in the 3D scenes generated by competing methods (Fig. 7).

Image-to-3D Generation. We evaluate image-to-3D generation on 3D-Front (Fu et al., 2021) against baselines (Liu et al., 2025; Yang et al., 2026; Team et al., 2026; Wang et al., 2026; Chung et al., 2023; Ardelean et al., 2025; Zhai et al., 2024; Chen et al., 2025a) using Chamfer Distance, F-Score, and IoU-B. Our framework significantly outperforms competitors across all metrics (Table 2) and synthesizes high-fidelity 3D assets with coherent topologies, overcoming the distorted geometries frequently produced by the competing baseline methods (Fig. 8).

Text-to-4D Generation. We evaluate text-to-4D generation against SOTA methods (Ling et al., 2024a; Bahmani et al., 2024; Zheng et al., 2024; Liu et al., 2025; Yu et al., 2024) using VBench (Huang et al., 2024) and a user study. Our framework achieves the highest automated and human evaluation scores (Table 3) and synthesizes 4D scenes with seamless state transitions and natural movements, overcoming competitors' severe motion artifacts and temporal flickering (Fig. 9).

Image-to-4D Generation. We evaluate image-to-4D generation against competitive baselines (Liu et al., 2025; Yang et al., 2026; Team et al., 2026; Zhao et al., 2024; Sun et al., 2024; Chen et al., 2025c; Mi et al., 2025) using VBench (Huang et al., 2024) and a user study. Our framework achieves superior performance across all automated metrics and human preferences (Table 4) and generates highly realistic 4D representations, overcoming the severe appearance degradation and dynamics drift typically observed in the alternative models (Fig. 10).

5.3 Ablation Study

Effectiveness of Multiple Spatiotemporal Representations. We ablate semantic, spatial, and temporal tokens, evaluating three metrics through a user study. Omitting any token severely degrades its primary metric and compromises other dimensions (Table 5), confirming that their synergy establishes a robust cognitive foundation for 4D scene synthesis.

Superiority of Graph Construction. We replace structured graph construction with direct concatenation of all extracted tokens. Table 6 shows a substantial decline across all VBench metrics, confirming that structured graph construction is crucial for robust 4D generation.

Impact of Separate Global and Local Graphs. We independently ablate the global appearance and local dynamic graphs. Fig. 11 shows that omitting either significantly degrades performance in its

Table 5: Ablation study of different tokens on user study metrics.

Settings			User Study		
Semantic Tokens	Spatial Tokens	Temporal Tokens	Semantic Fidelity ↑	Appearance Plausibility ↑	Dynamics Coherence ↑
✓			8.2	4.3	3.5
	✓		4.1	8.5	3.8
		✓	4.5	4.8	8.0
✓	✓		8.9	9.2	4.6
✓		✓	9.0	5.2	8.6
	✓	✓	5.3	9.0	8.5
✓	✓	✓	**9.4**	**9.7**	**9.3**

Table 6: Graph construction ablation.

Settings	VBench Huang et al. (2024)		
	Consis. ↑	Dynam. ↑	Aesth. ↑
w/o Graph	86.24	87.15	46.08
w/. Graph	**97.83**	**99.74**	**59.37**

Table 7: Semantic encoders discussion.

Settings	VBench Huang et al. (2024)		
	Consis. ↑	Dynam. ↑	Aesth. ↑
ResNet50	88.45	90.82	49.64
CLIP ViT-L	95.62	97.10	56.21
Wan2.1	**97.83**	**99.74**	**59.37**

Table 8: Fusion strategy discussion.

Settings	VBench Huang et al. (2024)		
	Consis. ↑	Dynam. ↑	Aesth. ↑
Concat.	90.10	91.85	52.12
Weighted	93.85	95.60	55.20
Sem-Bridged	**97.83**	**99.74**	**59.37**

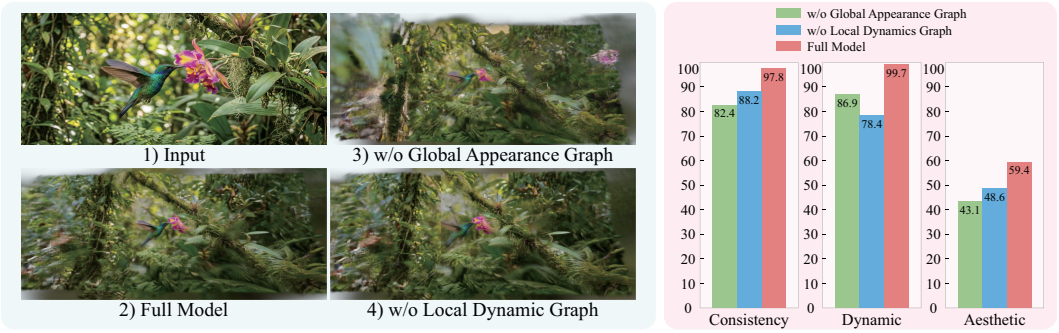


Figure 11: **Ablation of separate global and local graphs.** The visualization reveal that ablating global or local graphs leads to obvious appearance degradation or motion artifacts. Moreover, the performance metrics significantly declined when ablating either branch.

respective domain. This confirms that the global graph is essential for spatial appearance consistency, and the local graph is vital for temporal dynamic coherence.

5.4 Discussion

Influence of Semantic Backbone. Replacing Wan2.1 (Wan et al., 2025) with ResNet50 and CLIP ViT-L confirms Wan2.1's superior semantic representations (Table 7).

Effectiveness of Semantic-Bridged Fusion Strategy. We replace our semantic-bridged fusion strategy with basic concatenation and weight fusion baselines to verify its effectiveness. Table 8 highlights that semantic-bridged fusion facilitates plausible 4D generation.

Limitations. Despite its robust 4D generation capabilities, our framework faces challenges with extremely long-duration dynamics. The autoregressive inference of spatiotemporal cognition graphs allows minor early topological inaccuracies to accumulate, potentially causing structural degradation over extended sequences. To mitigate error propagation and extend the generation horizon, future work will explore bidirectional temporal constraints and global trajectory refinement modules.

6 Conclusion

In this work, we propose ST-Gen4D, a cognition-guided framework that embeds spatiotemporal cognition into a world model for 4D generation. By integrating disentangled spatial appearance and temporal dynamics representations, our method effectively resolves global apparent collapse and local dynamic artifacts prevalent in existing 4D generation models. Specifically, we construct a spatiotemporal cognition graph from foundational multimodal tokens to capture complex topological dependencies. An action-conditioned world model then autoregressively infers future cognition states, which effctively steer a latent flow-matching diffusion process for 4D Gaussian generation. Extensive experiments across 3D and 4D generation tasks demonstrate the superiority of our approach, affirming its exceptional capability in modeling plausible spatial appearance and coherent temporal dynamics.

AI USE STATEMENT

No generative AI tools were used in the research, implementation, experimentation, analysis, or preparation of this work. The authors take full responsibility for all content presented in this work.

REFERENCES

Andreea Ardelean, Mert Özer, and Bernhard Egger. Gen3dsr: Generalizable 3d scene reconstruction via divide and conquer from a single view. In *2025 International Conference on 3D Vision (3DV)*, pp. 616–626. IEEE, 2025.

Sherwin Bahmani, Ivan Skorokhodov, Victor Rong, Gordon Wetzstein, Leonidas Guibas, Peter Wonka, Sergey Tulyakov, Jeong Joon Park, Andrea Tagliasacchi, and David B Lindell. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7996–8006, 2024.

Sherwin Bahmani, Tianchang Shen, Jiawei Ren, Jiahui Huang, Yifeng Jiang, Haithem Turki, Andrea Tagliasacchi, David B Lindell, Zan Gojcic, Sanja Fidler, et al. Lyra: Generative 3d scene reconstruction via video diffusion model self-distillation. *arXiv preprint arXiv:2509.19296*, 2025.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021.

Michael J Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8726–8737, 2023.

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*, pp. 22563–22575, 2023.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Leo Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024.

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.

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020.

Minglin Chen, Longguang Wang, Sheng Ao, Ye Zhang, Kai Xu, and Yulan Guo. Layout2scene: 3d semantic layout guided scene generation via geometry and appearance diffusion priors. *arXiv preprint arXiv:2501.02519*, 2025a.

Yabo Chen, Yuanzhi Liang, Jiepeng Wang, Tingxi Chen, Junfei Cheng, Zixiao Gu, Yuyang Huang, Zicheng Jiang, Wei Li, Tian Li, et al. Teleworld: Towards dynamic multimodal synthesis with a 4d world model. *arXiv preprint arXiv:2601.00051*, 2025b.

Zhaoxi Chen, Tianqi Liu, Long Zhuo, Jiawei Ren, Zeng Tao, He Zhu, Fangzhou Hong, Liang Pan, and Ziwei Liu. 4dnex: Feed-forward 4d generative modeling made easy. *arXiv preprint arXiv:2508.13154*, 2025c.

Jaeyoung Chung, Suyoung Lee, Hyeongjin Nam, Jaerin Lee, and Kyoung Mu Lee. Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. *arXiv preprint arXiv:2311.13384*, 2023.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5828–5839, 2017.

Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, et al. Objaverse-xl: A universe of 10m+ 3d objects. *Advances in Neural Information Processing Systems*, 36:35799–35813, 2023.

Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10933–10942, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Yuze He, Yushi Bai, Matthieu Lin, Wang Zhao, Yubin Hu, Jenny Sheng, Ran Yi, Juanzi Li, and Yong-Jin Liu. T3 bench: Benchmarking current progress in text-to-3d generation. *arXiv preprint arXiv:2310.02977*, 2023.

Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21807–21818, 2024.

Yanqin Jiang, Li Zhang, Jin Gao, Weimin Hu, and Yao Yao. Consistent4d: Consistent 360 {\deg} dynamic object generation from monocular video. *arXiv preprint arXiv:2311.02848*, 2023.

Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Dynamicstereo: Consistent dynamic depth from stereo videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13229–13239, 2023.

Yann LeCun et al. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2041–2050, 2018.

Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 300–309, 2023.

Huan Ling, Seung Wook Kim, Antonio Torralba, Sanja Fidler, and Karsten Kreis. Align your gaussians: Text-to-4d with dynamic 3d gaussians and composed diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8576–8588, 2024a.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22160–22169, 2024b.

Tianqi Liu, Zihao Huang, Zhaoxi Chen, Guangcong Wang, Shoukang Hu, Liao Shen, Huiqiang Sun, Zhiguo Cao, Wei Li, and Ziwei Liu. Free4d: Tuning-free 4d scene generation with spatial-temporal consistency. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 25571–25582, 2025.

Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 21013–21022, 2022.

Guanxing Lu, Baoxiong Jia, Puhao Li, Yixin Chen, Ziwei Wang, Yansong Tang, and Siyuan Huang. Gwm: Towards scalable gaussian world models for robotic manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9263–9274, 2025.

Yanzhe Lyu, Chen Geng, Karthik Dharmarajan, Yunzhi Zhang, Hadi Alzayer, Shangzhe Wu, and Jiajun Wu. Choreographing a world of dynamic objects. *arXiv preprint arXiv:2601.04194*, 2026.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. Leworld-model: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026.

Zhenxing Mi, Yuxin Wang, and Dan Xu. One4d: Unified 4d generation and reconstruction via decoupled lora control. *arXiv preprint arXiv:2511.18922*, 2025.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2446–2454, 2020.

Wenqiang Sun, Shuo Chen, Fangfu Liu, Zilong Chen, Yueqi Duan, Jun Zhang, and Yikai Wang. Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion. *arXiv preprint arXiv:2411.04928*, 2024.

InSpatio Team, Xiaoyu Zhang, Weihong Pan, Zhichao Ye, Jialin Liu, Yipeng Chen, Nan Wang, Xiaojun Xiang, Weijian Xie, Yifu Wang, et al. Inspatio-worldfm: An open-source real-time generative frame model. *arXiv preprint arXiv:2603.11911*, 2026.

Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12619–12629, 2023a.

Haonan Wang, Hanyu Zhou, Haoyue Liu, and Luxin Yan. 4d-vggt: A general foundation model with spatiotemporal awareness for dynamic scene geometry estimation. *arXiv preprint arXiv:2511.18416*, 2025a.

Haonan Wang, Hanyu Zhou, Haoyue Liu, Tao Gu, and Luxin Yan. Cog2gen3d: Sculpturing 3d semantic-geometric cognition for 3d generation. *arXiv preprint arXiv:2603.05845*, 2026.

Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Yi Zhang, Chang Zeng, Yanxi Zhou, et al. Spatialvid: A large-scale video dataset with spatial annotations. *arXiv preprint arXiv:2509.09676*, 2025b.

Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4909–4916. IEEE, 2020.

Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Zun Wang, Yansong Shi, et al. Internvideo2: Scaling foundation models for multimodal video understanding. In *European conference on computer vision*, pp. 396–416. Springer, 2024.

Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in neural information processing systems*, 36:8406–8441, 2023b.

Yuxue Yang, Lue Fan, Ziqi Shi, Junran Peng, Feng Wang, and Zhaoxiang Zhang. Neoverse: Enhancing 4d world model with in-the-wild monocular videos. *arXiv preprint arXiv:2601.00393*, 2026.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12–22, 2023.

Taoran Yi, Jiemin Fang, Junjie Wang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. Gaussiandreamer: Fast generation from text to 3d gaussians by bridging 2d and 3d diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6796–6807, 2024.

Heng Yu, Chaoyang Wang, Peiye Zhuang, Willi Menapace, Aliaksandr Siarohin, Junli Cao, Laszlo A Jeni, Sergey Tulyakov, and Hsin-Ying Lee. 4real: Towards photorealistic 4d scene generation via video diffusion models. *Advances in Neural Information Processing Systems*, 37:45256–45280, 2024.

Guangyao Zhai, Evin Pınar Örnek, Dave Zhenyu Chen, Ruotong Liao, Yan Di, Nassir Navab, Federico Tombari, and Benjamin Busam. Echoscene: Indoor scene generation via information echo over scene graph diffusion. In *European Conference on Computer Vision*, pp. 167–184. Springer, 2024.

Yuyang Zhao, Zhiwen Yan, Enze Xie, Lanqing Hong, Zhenguo Li, and Gim Hee Lee. Animate124: Animating one image to 4d dynamic scene. *arXiv preprint arXiv:2311.14603*, 2023.

Yuyang Zhao, Chung-Ching Lin, Kevin Lin, Zhiwen Yan, Linjie Li, Zhengyuan Yang, Jianfeng Wang, Gim Hee Lee, and Lijuan Wang. Genxd: Generating any 3d and 4d scenes. *arXiv preprint arXiv:2411.02319*, 2024.

Yang Zheng, Adam W Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J Guibas. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19855–19865, 2023.

Yufeng Zheng, Xueting Li, Koki Nagano, Sifei Liu, Otmar Hilliges, and Shalini De Mello. A unified approach for text-and image-guided 4d scene generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7300–7309, 2024.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.

Ziyue Zhu, Zhanqian Wu, Zhenxin Zhu, Lijun Zhou, Haiyang Sun, Bing Wan, Kun Ma, Guang Chen, Hangjun Ye, Jin Xie, et al. Worldsplat: Gaussian-centric feed-forward 4d scene generation for autonomous driving. *arXiv preprint arXiv:2509.23402*, 2025.

A TECHNICAL APPENDICES AND SUPPLEMENTARY MATERIAL

In this supplementary material, we conduct some additional ablation study and discussions, and provide detailed limitation analysis and possible solution in Sec. A.1. Moreover, we demonstrate more visualization results of text-to-4D and image-to-4D generation in Sec. A.2.

A.1 ABLATION STUDY AND DISCUSSION

Impact of 3D Geometry Embedding. We ablate the 3D geometry embedding to validate its necessity for the global appearance and local dynamic graphs in ST-Gen4D. As Table 9 reveals, omitting this spatial prior causes severe degradation in reconstructing unstructured primitives. Its inclusion is therefore essential for optimal generative performance, proving that explicit 3D geometry embedding are crucial for the network to capture geometric dependencies.

Table 9: Impact of 3D geometry embedding.

Settings	3D-Front (Fu et al., 2021)		
	Chamfer Distance ↓	F-Score ↑	IoU ↑
w/o 3D Geometry Embedding	0.075	51.24	0.531
w/. 3D Geometry Embedding	**0.051**	**62.15**	**0.713**

Influence of Cognition Graph Nodes. To assess the representational capacity of our cognition graphs, we investigate the influence of the number of graph nodes. As Table 10 reveals, this parameter dictates a clear performance trade-off. While too few nodes restrict the degrees of freedom needed to model intricate spatial appearance and local dynamics, an excessive count incurs redundant computation and risks overfitting, ultimately degrading performance. Therefore, an optimal intermediate node count best balances capacity and generalization for the sculpturing of 4D spatiotemporal cognition.

Table 10: Influence of the number of nodes in spatiotemporal cognition graph.

Number of Graph Nodes	VBench (Huang et al., 2024)		
	Consistency ↑	Dynamics ↑	Aesthetic ↑
32	90.15	91.42	49.28
64	94.33	95.81	54.65
128	96.72	98.45	57.91
256	**97.83**	**99.74**	**59.37**
512	97.05	98.66	58.24
1024	95.18	96.73	56.41

Discussion on Computational Efficiency. To evaluate the practical applicability of our approach, we compare its computational efficiency against recent baselines on the VBench (Huang et al., 2024) dataset. As summarized in Table 11, our approach avoids the excessive memory consumption typical of methods like Dream-in-4D (Zheng et al., 2024), while overcoming the slow inference speeds often seen in lightweight alternatives like Free4D (Liu et al., 2025). By delivering rapid inference times alongside a moderate memory footprint, our proposed method effectively balances generation quality and resource utilization. These results demonstrate that our framework maintains competitive computational efficiency while achieving state-of-the-art 4D scene generation performance.

Limitations and Possible Solutions. While our framework demonstrates robust capabilities in 4D scene synthesis, it encounters specific challenges when scaling to extremely long-duration generation. We detail these limitations and outline promising future directions below.

Limitation 1: Error Accumulation in Autoregressive Inference. The main challenge in long-sequence generation is autoregressive inference of the spatiotemporal cognition graph. When predicting future states from previous frames, minor early inaccuracies—such as misalignments in spatial graph nodes or local geometry embeddings—can compound. Over time, error propagation inevitably leads to structural degradation, semantic drift, and a loss of fine-grained details in later frames.

Table 11: Discussion on computational efficiency.

Methods	Memory Consumption (GB)	Inference Time (min)
4Dfy (Bahmani et al., 2024)	69.85	43.7
Dream-in-4D (Zheng et al., 2024)	78.28	32.6
Free4D (Liu et al., 2025)	42.64	16.9
NeoVerse (Yang et al., 2026)	74.32	3.8
WorldFM (Team et al., 2026)	47.76	9.7
Ours	67.35	6.5



Figure 12: Text-to-4D visualization on ST-4D dataset.

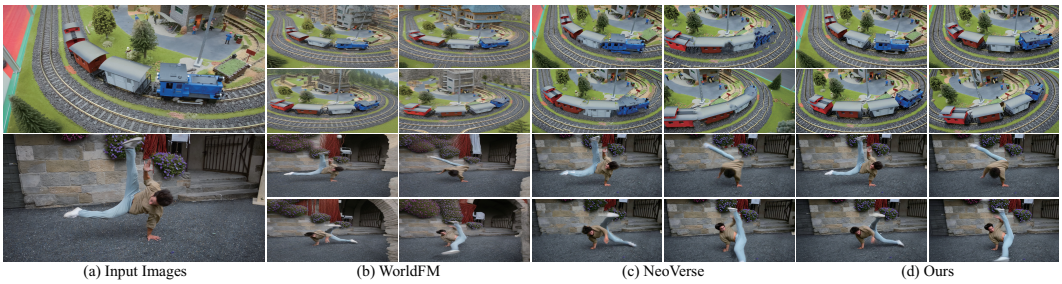


Figure 13: Image-to-4D visualization on DAVIS (Pont-Tuset et al., 2017) dataset.

Possible Solution 1: Global Trajectory Refinement Modules. To mitigate autoregressive drift, future work will explore integrating global trajectory refinement modules. Rather than relying solely on sequential generation, a secondary refinement pass could optimize the long-term motion trajectories of graph nodes holistically. By enforcing temporal consistency across the entire generation window, this approach would effectively suppress local error accumulation and preserve structural integrity.

Limitation 2: Sensitivity to Early Topological Errors in Complex Dynamics. While our graph efficiently captures spatiotemporal features, the initial topologies may struggle to accurately reflect underlying structures in scenarios with highly complex, rapid, or occluded dynamics. Because the current temporal evolution is strictly causal (forward-driven), the network lacks the capacity to self-correct these early topological errors once they are embedded into the dynamic graph.

Possible Solution 2: Bidirectional Temporal Constraints. To introduce self-correction mechanisms, we plan to implement bidirectional temporal constraints. By incorporating non-causal temporal modeling, such as bidirectional diffusion processes, the framework could leverage future context to correct past ambiguities. This would enable the spatiotemporal graph to dynamically adjust topological structures based on complete motion context, enhancing robustness.

A.2 Comparison Experiments

Text-to-4D Generation. We provide additional text-to-4D visualizations competing with SOTA methods (Bahmani et al., 2024; Zheng et al., 2024) on our proposed ST-4D datasets. Visually, competing approaches suffer from degraded overall appearances and fail to generate distinct local motions. Conversely, our method synthesizes 4D scenes with highly consistent global appearances and fluid, coherent local dynamics. This effectively demonstrates the superiority of our framework in ensuring both spatial plausibility and temporal coherence.

Image-to-4D Generation. We provide additional image-to-4D results against SOTA methods (Yang et al., 2026; Team et al., 2026) on DAVIS (Pont-Tuset et al., 2017) datasets. While baselines often exhibit inconsistent global backgrounds and blurred local motions, our method maintains high fidelity across both global spatial appearances and precise local dynamics. This highlights our advantage in transforming static images into realistic and temporally coherent 4D assets.