Alignment Is All You Need For X-to-4D Generation

Qiaowei Miao, *Student Member, IEEE,* Kehan Li, Yawei Luo*, Yi Yang, *Fellow, IEEE*

***Abstract*—Generative diffusion models excel at synthesizing high-quality images, videos, and 3D content under multimodal control. However, arbitrary user-defined modality-to-4D (X-to-4D) generation remains challenging due to the high cost of constructing diverse datasets and the limited scalability of existing methods. This paper presents Align4D, a flexible framework that translates any-modal input into coherent video–3D pairs, using video to guide 4D motion and 3D data to shape 4D geometry. Align4D introduces three key techniques: (1) Object Distance Alignment, which searches Video-Aligned and Multiview-Aligned Object Distances (VAOD/MAOD) respectively, to reconcile 4D renderings to video and the priors of multiview diffusion models; (2) Motion-Geometry Joint Alignment, which constrains known and unknown views through synchronized video and 3D inputs, ensuring consistent 4D generation; and (3) Asynchronous Optimization, which decouples Gaussian attribute and deformation network training to enhance motion and geometry fidelity. We further propose the X4D dataset, integrating prompt, image, video, and 3D data for benchmarking. Experiments on X4D and Consistent4D demonstrate that Align4D achieves state-of-the-art quality and consistency in X-to-4D generation. Project page: https://miaoqiaowei.github.io/Align4D/.**

***Index Terms*—4D generation, object generation, multimedia content creation.**

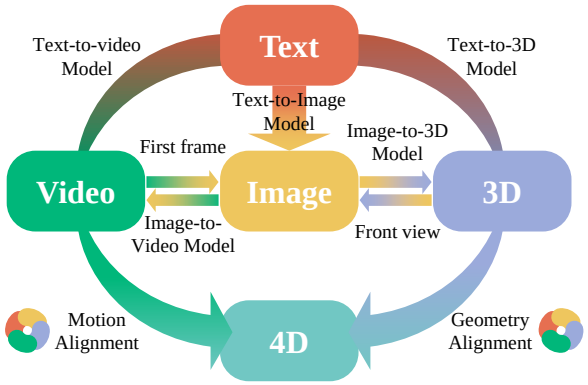


Fig. 1. **Align4D transforms an arbitrary input into a coherent *Video-3D pair* by sequentially leveraging multiple off-the-shelf models.** Within this X-to-4D generation framework, Align4D focuses on rigorously aligning the 4D object's temporal motion with the video prior and its spatial geometry with the 3D representation, thereby achieving powerful synthesis capabilities.

I. Introduction

4D content generation is essential for creating realistic and dynamic content in applications such as virtual reality, the metaverse, computer games, and cinematic visual effects. However, existing 4D generation methods [1]–[10] typically focus on single-modal inputs, which limits their generative flexibility and capabilities. Text-to-4D methods [1]–[3] provide broad accessibility, but the generated objects often lack a strong sense of realism and purpose. Image-to-4D methods [7], [11], [12] enhance appearance quality but fail to effectively provide motion guidance. Video-to-4D methods [5], [8]–[10], [13] improve motion realism but compromise structural consistency. Meanwhile, 3D-to-4D methods [6] preserve spatial geometry but lack temporal coherence. Recent efforts [14], [15] explore constructing datasets to train diffusion models for generating multiview and multi-timestamp images. However, scaling these datasets to enable multimodal inputs remains infeasible due to the prohibitive data collection and computational resource requirements. These limitations highlight the urgent need for a flexible and unified framework that enables arbitrary modality-to-4D (X-to-4D) generation.

In this paper, we revisit the 4D generation task and present a key insight: **4D content synthesis can be intrinsically decoupled into 3D geometry generation and temporal motion generation.** Building on this idea, we propose **Align4D**, a unified framework that reformulates arbitrary-modality-guided 4D generation into a task of aligning 4D assets with synchronized video and 3D data through matched object distances. Instead of training an end-to-end model from scratch, our framework leverages the robust inference capabilities of off-the-shelf pretrained diffusion models to bridge the gap between input modalities and 4D outputs. As illustrated in Figure 1, Align4D establishes flexible generative pathways where diversity input (text, image, video, or 3D) is converted into a coherent *Video-3D pair* via video generation models [16]–[19] and 3D generation models [20]–[22]. Align4D then focuses on rigorously aligning the temporal motion of the 4D target with the video input, and ensuring its spatial geometry is consistent with the 3D representation. This dual alignment process is critical, as it guarantees the synthesized 4D results exhibit both action coherence and structural fidelity. This modular design effectively circumvents the data scarcity issue, thereby offering a generalizable and flexible solution for X-to-4D generation with powerful synthesis capabilities, as shown in Figure 2.

However, unifying these disparate priors presents a significant challenge: **aligning the 4D output with both the motion posture from the generated video and the geometric structure from the generated 3D content.** Since the video and 3D priors are derived from independent models, they often

This work was supported by National Natural Science Foundation of China (62293554, U2336212), "Pioneer" and "Leading Goose" R&D Program of Zhejiang (2025C02022, 2024C01161), Zhejiang Provincial Natural Science Foundation of China (LZ24F020002), Ningbo Innovation "Yongjiang 2035" Key Research and Development Programme (2024Z292), and Young Elite Scientists Sponsorship Program by CAST (2023QNRC001). The author gratefully acknowledges the support of Zhejiang University Education Foundation Qizhen Scholar Foundation. This research was supported by HUAWEI's AI Hundred Schools Program and was carried out using the Ascend AI technology stack.

Qiaowei Miao, Kehan Li, Yawei Luo, and Yi Yang are with the Zhejiang University, Hangzhou 310027, China (e-mail: qiaoweimiao@zju.edu.cn; kehan.li@zju.edu.cn;yaweiluo@zju.edu.cn; yangyics@zju.edu.cn).

Yawei Luo is the corresponding author.

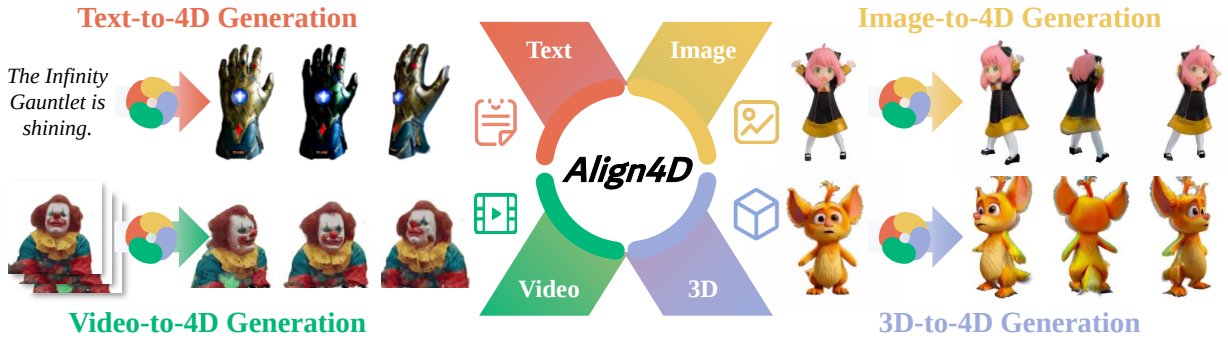


Fig. 2. **Align4D is a novel X-to-4D framework that enables users to input text, images, videos, or 3D objects to generate 4D targets.** Align4D transforms arbitrary modality inputs into corresponding video-3D pairs, synchronizing the motion of the generated 4D objects with the videos and aligning their structures with the 3D data through matched object distances , achieving dynamic objects with high temporal dynamics and precise geometry.

exhibit spatial or temporal discrepancies. We identify two critical aspects of this alignment problem. First, for **Known Spatiotemporal Viewpoints**, the 4D target must align with the 3D geometry at the initial state and the video content across time. A major hurdle here is determining the appropriate object distance; existing methods [5], [10], [23] often rely on manual, empirical settings, which lead to floating artifacts or distortions when applied to arbitrary inputs. Second, for **Unknown Spatiotemporal Viewpoints**, the model must maintain geometric consistency across unseen views while adhering to the motion dynamics observed in the video. Optimizing 4D targets using multiple diffusion priors simultaneously is notoriously unstable and time consuming [7], [8], [24], often resulting in degraded quality due to parameter sensitivity.

To tackle these challenges, Align4D introduces a novel alignment-driven optimization strategy designed to synthesize 4D objects with smooth motion and consistent geometry. First, we propose **Object Distance Alignment**, which automatically searches for the Video-Aligned Object Distance (VAOD) to match front-view renderings with video frames, and the Multiview-Aligned Object Distance (MAOD) to calibrate the 4D model with multiview diffusion priors. Furthermore, we develop a **Motion-Geometry Joint Alignment (MGJA)** module. For known viewpoints, we utilize VAOD to strictly align the 4D model with the input video. For unknown viewpoints, rather than relying on unstable multi-model optimization, we employ a single multiview diffusion model guided by MAOD and condition it on the video frames to transfer motion and geometry priors to non-front views. To ensure robust convergence, we introduce an **Asynchronous Optimization** strategy that refines geometry and motion parameters separately. Extensive experiments on our newly collected X4D dataset and the Consistent4D dataset [5] demonstrate that Align4D effectively harmonizes conflicting priors, generating high-fidelity 4D targets that often surpass the quality of the initial video-3D pairs. Our contributions are as follows:

- We propose an X-to-4D generation framework that supports arbitrary input modalities, including text, images, videos, and 3D data, as conditioning inputs to generate 4D assets.
- We introduce a novel object distance alignment method designed to search the matched object distances for aligning 4D renderings with video and the prior of a multiview diffusion model. Building on this foundation, we employ a single multiview diffusion model to jointly align the motion and geometry of 4D renderings with video and 3D data in an asynchronous way.
- We construct X4D, a first quadruple dataset to benchmark X-to-4D generation capabilities via inputs generated by pretrained diffusion models.
- Extensive experiments demonstrate that Align4D excels in generating 4D objects, producing fine textures, precise geometry, and seamless motion.

II. Related Works

Generative diffusion models. They revolutionize the landscape of visual generation, demonstrating remarkable performance across tasks such as image, video, and 3D content generation [25]. (a) Image generation models [26]–[29] leverage their impressive text-guided generative capabilities to produce high-resolution and highly creative images, sparking significant interest in extending diffusion models to video and 3D generation domains. (b) Video generation models [19], [30]–[38], including text-to-video and image-to-video generation, garner increasing attention. For instance, text-to-video models [30], [39] rely on large-scale, high-quality text-to-video datasets for training, enabling a deeper understanding of verbs and generating rich, creative sequences of temporally coherent video frames. On the other hand, image-to-video models [19], [33] infer the subsequent actions of a target object based solely on a given initial frame. However, these models lack flexible control over the generated actions. (c) 3D generation models [21], [40]–[47] allow users to set text or images as control conditions to generate static 3D targets based on meshes or 3D Gaussian representations. These models achieve robust generalization through large-scale training datasets and can rapidly generate 3D objects for unseen inputs using inference alone. The rapid advancements in image, video, and 3D diffusion models lay a solid foundation for constructing an X-to-4D generation framework.

4D Generation. The goal of 4D generation is to synthesize dynamic 3D objects with consistent geometry and motion across views. However, existing approaches still face

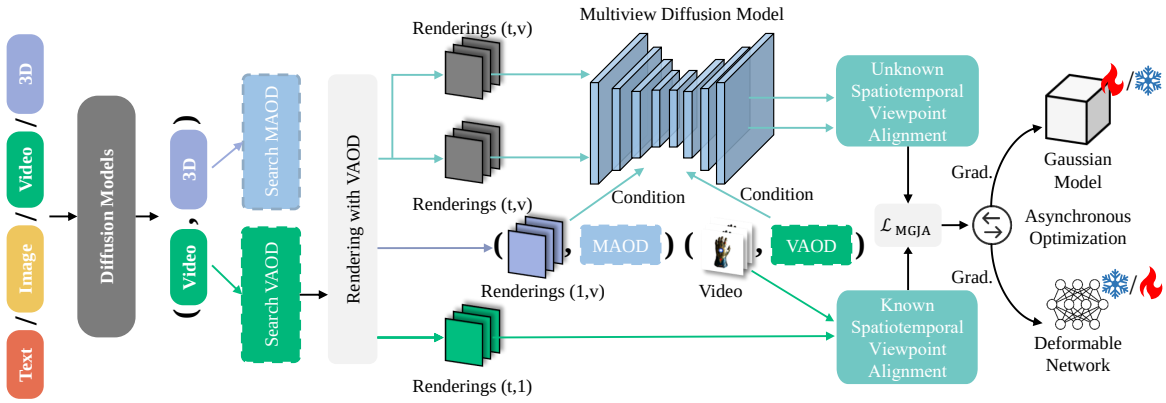


Fig. 3. **Overview of the Align4D framework.** Given arbitrary input modalities, pretrained video and 3D diffusion models are used to construct a unified video–3D pair. We then search for two key object distances: the Video-Aligned Object Distance (VAOD), which aligns the 4D front-view renderings with the video, and the Multiview-Aligned Object Distance (MAOD), which aligns multiview geometry with the multiview diffusion prior [53]. Using these distances, Align4D performs Motion–Geometry Joint Alignment under both known and unknown spatiotemporal viewpoints. Finally, an asynchronous optimization stage refines the Gaussian representation and the deformation network to produce high-quality 4D content.

several critical limitations across different dimensions. Regarding **control modalities**, current methods primarily rely on single-modal inputs: while text-driven methods [2], [4] and image-to-4D pipelines [7] offer partial multimodal support, dedicated video-to-4D [8], [10], [48] and 3D-to-4D [6] methods cannot jointly accommodate text, image, 3D, and video as unified inputs. Concerning **representations**, NeRF-based approaches [2]–[4], [49] provide smooth temporal modeling but suffer from prohibitive training times. Conversely, Gaussian Splatting variants [1], [7], [48], [50]–[52] improve efficiency yet struggle to provide fine-grained motion control over thousands of Gaussians, compromising temporal coherence. For **guidance and optimization**, many approaches rely on proprietary video diffusion models [30], [39] for motion guidance, which severely limits reproducibility. Open-source attempts [2], [7] often suffer from inaccurate object distances and an entangled motion–geometry optimization process, leading to degraded 4D consistency. Furthermore, the reliance on optimization from random initialization via SDS contributes to slow convergence and prolonged generation times. In contrast, our **Align4D** introduces a 3DGS-based 4D framework with a deformable motion field and novel multimodal alignment strategies. This design enables efficient generation while preserving both geometric fidelity and video-consistent motion.

III. Methodology

In this paper, we introduce Align4D, an X-to-4D generation framework that aligns motion with video priors and geometry with 3D inputs. Leveraging pretrained models, Align4D converts arbitrary modalities into unified video–3D pairs for 4D synthesis. We first outline the preliminaries (Section III-A). To ensure accurate motion–geometry alignment, we propose object distance alignment, which identifies the video-aligned object distance and multiview-aligned object distance (Section III-B). Based on these distances, we design a motion–geometry joint alignment module for both known and unknown spatiotemporal viewpoints (Section III-C). Finally, an asynchronous optimization strategy further refines the 4D representation (Section III-D), and we introduce the constructed multimodal dataset X4D (Section III-E). Our framework is illustrated in Figure 3.

A. Preliminaries

Score Distillation Sampling (SDS). SDS is a widely adopted mechanism in 3D and 4D generation [7], [24], [40], [54], enabling the transfer of 2D diffusion priors into 3D representations. During diffusion model training, a noised sample x_τ is produced by adding Gaussian noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ to data x at a randomly sampled timestep τ, and the UNet ϕ is trained to predict the added noise ϵ. Given a 3D scene representation with parameters θ, its rendering $x = g(\theta)$ is treated as the diffusion input. The SDS gradient [24] is then computed as:

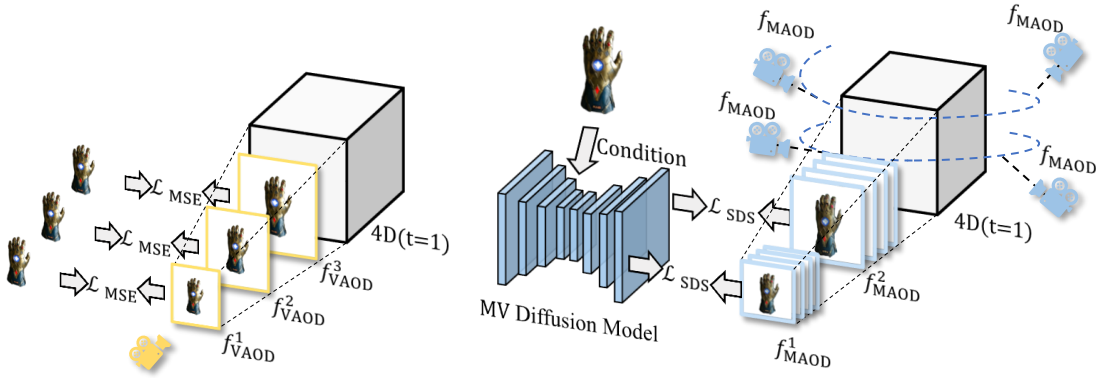
$$\nabla_\theta \mathcal{L}_{\text{SDS}}(x = g(\theta)) = \mathbb{E}_{\tau,\epsilon}\left[w(\tau)\big(\hat{\epsilon}_\phi(\mathbf{z}, v, \tau) - \epsilon\big)\frac{\partial x}{\partial \theta} \right], \quad (1)$$

where $\mathbf{z}$ is the latent corresponding to x, v denotes conditioning signals, and $w(\tau)$ is the timestep-dependent weighting. This gradient is backpropagated through the differentiable renderer g to update the 3D parameters θ.

B. Object Distance Alignment

Theoretical analysis. Whether using explicit video supervision or SDS-based implicit supervision, consistent object-distance alignment is fundamental to 4D generation. In video-conditioned settings, the camera captures the target at a fixed object distance, whereas pretrained diffusion models are trained on images rendered with a canonical focal length but varying distances. This mismatch makes 4D generation challenging: the reconstructed 4D object must simultaneously match the video's object distance and conform to the spatial priors encoded in the diffusion model. Thus, estimating both the video-aligned object distance and the multiview diffusion–aligned object distance becomes essential.

We introduce the pinhole camera model to illustrate how to calculate the object distance. For a given camera with a fixed focal length f, an image x of a 4D target θ is captured at an



(a) Search Video-Aligned Object Distance (b) Search Multiview-Aligned Object Distance

Fig. 4. **Object distance alignment.** (a) We search for the Video-Aligned Object Distance (VAOD) to align the known front-view renderings of the 4D object with the input video. Specifically, we render front-view images under different object distances and compute $\mathcal{L}_{\text{MSE}}$ with respect to the video frames; the distance that yields the global minimum $\mathcal{L}_{\text{MSE}}$ is selected as the VAOD. (b) We then search for the Multiview-Aligned Object Distance (MAOD), which aligns multiview renderings with the geometry prior of the multiview diffusion model. Using the first frame as the condition, we render four orthogonal views and compute $\mathcal{L}_{\text{SDS}}$. The object distance corresponding to a local minimum of $\mathcal{L}_{\text{SDS}}$, and that is smaller than the VAOD, is selected as the MAOD.

object distance d. If the true physical width of the 4D target θ is W, and its projected width in image x is w, then by similar triangles, we get:

$$\frac{w}{f} = \frac{W}{d}, \quad d = \frac{W}{w} f. \tag{2}$$

However, a 4D target isn't a physical entity, so its physical dimensions cannot be directly measured to derive the object distance. Therefore, we propose a search method to find the optimal object distance. Based on Equation 2, the ratio of any object distance d_s to the video-aligned object distance d_v is equal to the ratio of the target size in the rendered image w_s to the target size in the video frame w_v:

$$\frac{d_v}{d_s} = \frac{w_s}{w_v}. \tag{3}$$

Thus, by varying d_s to generate rendered images x_s, we compute the corresponding widths w_s and compare them with the target width w_v in the video frame. When w_s closely matches w_v, Equation 3 indicates that $d_v = d_s$, *i.e.*, d_s is identified as the video-aligned object distance.

For the object distance associated with the multiview diffusion model, the matched value should yield rendered inputs whose output distribution matches the model's training-time distribution. Since diffusion models enforce alignment between their outputs and a Gaussian prior during training, we can assess how well a candidate object distance aligns with the training distribution by fixing condition and varying only the object distance. In this setting, the similarity between the UNet-predicted noise and the Gaussian prior serves as an indicator of distance alignment. Motivated by this observation, we treat the diffusion UNet as a distributional discriminator and use the SDS loss—an effective proxy for the KL divergence between the model's output distribution and the prior [24]—to evaluate the degree of alignment at each object distance. This allows us to reliably search for the multiview diffusion model–aligned object distance. Building on this theoretical basis, we then develop concrete search procedures for the Video-Aligned Object Distance (VAOD) and the Multiview-Aligned Object Distance (MAOD).

Video-Aligned Object Distance (VAOD). To determine the object distance that best aligns the 4D model with the input video, we search for the Video-Aligned Object Distance (VAOD). As shown in Figure 4 (a), we initialize the 4D representation θ using the provided 3D model ψ at $t = 1$. We then sweep the object distance $d' \in \mathcal{D} = [d_{\min}, d_{\max}]$ and render the front-view images $\{x^{d'}_{\theta_1}\}$. Each rendering is compared with the first video frame I_1 using the Mean Squared Error (MSE), and the distance yielding the global minimum is selected as:

$$d_{\text{VA}} = \arg\min_{d'} \|x^{d'}_{\theta_1} - I_1\|^2_2. \tag{4}$$

The rationale for choosing the global minimum is illustrated in Figure 5. When d' is too small, the object becomes overly magnified, producing saturated renderings with uniformly low MSE. As d' increases, the object gradually becomes fully visible, and the MSE decreases until reaching the best geometric and motion correspondence. Beyond this point, further increases in d' cause the object to shrink and eventually vanish, driving the rendered image toward a white background and increasing the MSE toward its upper bound.

Multiview-Aligned Object Distance (MAOD). Selecting a matched object distance is crucial when using the SDS method with multiview diffusion models to provide geometric priors. This necessity stems from the limited scope of pretrained data in these models, often resulting in the classification of images for 4D generation as out-of-distribution data, thereby making generative outcomes heavily dependent on the diffusion model's ability to generalize. The aligned object distance, as a key control parameter, significantly enhances the model's capacity to adapt to out-of-distribution data and strengthens its ability to provide geometric priors within the SDS framework, ultimately improving the quality of 4D generation.

To find the optimal object distance parameter, we search for the Multiview-Aligned Object Distance (MAOD) to align multiview renderings of 4D objects with the multiview diffusion

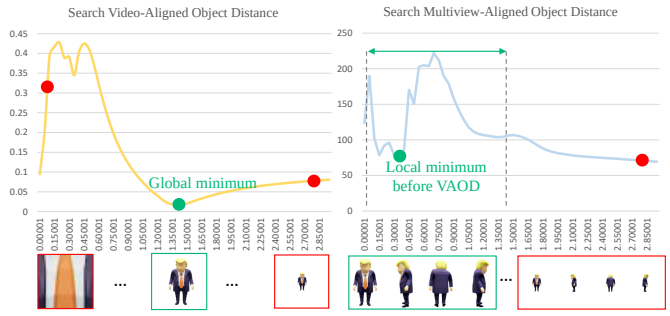


Fig. 5. **Searching strategies for video-aligned object distance and multiview-aligned object distance.**

model's prior. As illustrated in Figure 4 (b), we obtain orthogonal four-view rendered images corresponding to azimuth angles $[-90°, 0°, 90°, 180°]$. Within the range $[d_{\min}, d_{\max}]$, we render a set of images $\left\{\left\{x_{\theta_1}^{c,d'}\right\}_{c\in C}\right\}_{d'\in\mathcal{D}}$, where C denotes the coordinates. We then use the video frame I_1 at time step $t = 1$ as the image control condition. It is important to note that I_1 is equivalent to the 4D front rendered image $x_{\theta_1}^{d_{\text{VA}}}$ corresponding to the VAOD. Next, the four orthogonal rendered images at each object distance d' are used as inputs to compute the modified SDS loss $\mathcal{L}_{\text{SDS}}$ for searching:

$$\mathcal{L}_{\text{SDS}} = \frac{1}{|C||\mathrm{T}|}\sum_{c\in C}\sum_{\tau\in\mathrm{T}} w(\tau)\left\|\epsilon_\phi(z_{\theta_1}^c; I_1, c, d', \tau) - \epsilon\right\|_2^2, \quad (5)$$

where $z_{\theta_1}^c = \alpha_\tau x_{\theta_1}^c + \sigma_\tau \epsilon_{\text{fix}}$. We sample a noise latent ϵ_{fix} and add it to each object distance and each viewpoint rendering $x_{\theta_1}^c$ to calculate $\mathcal{L}_{\text{SDS}}$. Here, c represents the coordinate of the rendering $x_{\theta_1}^c$, τ is the timestep of the multiview diffusion model. We use diffusion timesteps $\mathrm{T} = \{700, 800, 900\}$.

The choice of the diffusion timestep τ is closely related to the noise schedule coefficients $w(\tau)$, which control the gradual noise injection in diffusion models—ensuring that early steps remain close to the data distribution while later steps approach a Gaussian prior [55], [56]. A smaller τ increases the mismatch between the predicted noise and the injected Gaussian noise, making the SDS loss less reliable for reflecting how well the diffusion prior aligns with the rendered inputs. We provide the SDS–object-distance curves in Figure 6. (a) When $\tau \in \{100, 200, 300\}$ display no stable pattern, resulting in unreliable MAOD estimation. (b) Intermediate timesteps $\tau \in \{400, 500, 600\}$ begin to exhibit a decreasing-then-increasing trend but remain inconsistent across distances. (c) In contrast, $\tau \in \{700, 800, 900\}$ produce highly consistent curves with a pronounced local minimum near the true object distance, enabling robust MAOD identification. Therefore, we adopt $\tau \in \{700, 800, 900\}$ and use the averaged SDS loss across these timesteps, which reduces stochastic variability and yields stable, reliable MAOD estimates.

Compared with the MSE loss used for VAOD, the SDS loss exhibits fundamentally different behavior. A well-trained diffusion model produces a lower SDS loss when the input aligns well with its learned prior. Consequently, within the distance range smaller than VAOD, using the video frame I_1 as the conditioning view yields a clear local minimum, which corresponds to the object distance most compatible

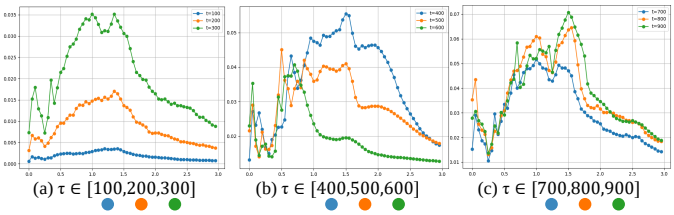


Fig. 6. **SDS loss versus object distance for different diffusion timesteps τ.** (a) $\tau \in \{100, 200, 300\}$, the curves exhibit irregular and unstable variations. (b) $\tau \in \{400, 500, 600\}$, a coarse decreasing–then–increasing trend emerges, but the minima remain inconsistent across distances. (c) $\tau \in \{700, 800, 900\}$, the curves show strong consistency with a clear local minimum near the matched object distance. Averaging the SDS losses across multiple large-τ timesteps yields a stable and reliable estimate of the MAOD.

with the multiview diffusion prior and is thus selected as the MAOD. For distances larger than VAOD, the rendered object progressively disappears, causing the SDS loss to drop toward a trivial lower bound associated with a blank background, as illustrated in Figure 4. Although this global minimum is numerically smaller, it does not reflect meaningful geometric alignment. Our experiments show that the correct MAOD is given by the local minimum located to the left of VAOD, whereas the global minimum is an artifact caused by background-dominated renderings. More analysis is provided in Section V-A.

C. Motion-Geometry Joint Alignment

Known Spatiotemporal Viewpoint Alignment. First, we align the given video with the multi-time-step front views of the 4D target. For a generated video consisting of $\mathcal{T}$ frames, $\{I_t\}_{\mathcal{T}}$, we render the corresponding $\mathcal{T}$ frames under the front view c' at the object distance d_{VA}, $\left\{x_{\theta_t}^{c'}\right\}_{\mathcal{T}}$. We calculate the MSE loss between them to inject motion into the 4D model:

$$\mathcal{L}_{\text{KSVA}} = \frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}} ||x_{\theta_t}^{c'} - I_t||_2^2 + ||m_{\theta_t}^{c'} - M_t||_2^2, \quad (6)$$

where θ_t denotes the 4D model at time t, $m_{\theta_t}^{c'}$ represents the alpha channel of $x_{\theta_t}^{c'}$ and M_t represents the mask of frame I_t.

Unknown Spatiotemporal Viewpoint Alignment. Beyond enforcing constraints on frontal actions, extending motion information to other unknown viewpoints remains a significant challenge. Previous approaches [1]–[4], [7] attempt to impose constraints using SDS optimization via image diffusion models, multiview diffusion models, and video diffusion models. However, this indiscriminate stacking of SDS not only incurs substantial computational overhead but also fails to guarantee cohesive results. To address this issue, we revisit the optimization strategy of multiview diffusion models for SDS and propose the Motion-Geometry Joint Alignment (MGJA).

MGJA leverages a single multiview diffusion model, referencing both video and 3D data, to effectively transfer motion and geometric performance from frontal views to unknown targets in nonfrontal views and non-initial timesteps for 4D objects. We utilize one multiview diffusion model [53] to simultaneously align 4D renderings with video motion and 3D geometry. The motion and posture of a 4D object in unknown

views need to align with the corresponding video frame at each time step. Specifically, as shown in Figure 3, for each time step t, we randomly select N viewpoints. These N viewpoints are used to render a set of corresponding images from the 4D model at the video-aligned object distance d_{VA}. Subsequently, we utilize $\mathcal{L}_{\text{mot}}$ to transfer the motion information from the front view to these rendered viewpoints:

$$\mathcal{L}_{\text{mot}} = \frac{1}{N}\sum_{n=1}^{N} w(\tau)\left\|\epsilon_\phi(\alpha_\tau x_{\theta_t}^{c_n} + \sigma_\tau\epsilon; I_t, c_n, d_{\text{VA}}, \tau) - \epsilon\right\|_2^2, \tag{7}$$

where ϵ is the randomly sampled noise, ϕ represents the U-Net of the multiview diffusion model, and $w(\tau)$ denotes the weight coefficient related to the timestep. The variable τ is a randomly sampled timestep within the multiview diffusion model. I_t is the t-th video frame, and c_n is the coordinate of the n-th viewpoint. $\mathcal{L}_{\text{mot}}$ ensures that the motion posture of the 4D object's position at each moment aligns with the motion posture of the front view.

Merely optimizing the motion alignment of the 4D object with the video might cause the generated target's geometric structure to deviate from the 3D target. Therefore, it is essential to ensure that the 4D target maintains geometric consistency with the 3D input. For unknown viewpoints of 4D object, another critical prior is the rendered images from the 3D representation at the same viewpoint. Although the rendered images of 3D and 4D from the same viewpoint may differ due to motion over time, their geometric structures remain highly correlated. By leveraging these same-view-rendered images from the 3D representation as geometric priors, we can ensure better alignment between the 4D output and the 3D target, avoiding undesirable modifications or artifacts introduced by the multiview diffusion model. This enhances the fidelity of the 4D geometry to the given 3D target. So, we propose $\mathcal{L}_{\text{geo}}$ for optimization:

$$\mathcal{L}_{\text{geo}} = \frac{1}{N}\sum_{n=1}^{N} w(\tau)||\epsilon_\phi(\alpha_\tau x_{\theta_t}^{c_n} + \sigma_\tau\epsilon; x_\psi^{c_n}, 0, d_{\text{MA}}, \tau) - \epsilon||_2^2. \tag{8}$$

Since both $x_{\theta_t}^{c_n}$ and $x_\psi^{c_n}$ are rendered from the same viewpoint c_n, the coordinate parameter in Equation 8 is zero. Thus, $x_\psi^{c_n}$ is regarded as the hypothetical front viewpoint and is utilized to refine the rendered image $x_{\theta_t}^{c_n}$ of the 4D model through SDS optimization, following motion deformation, from the viewpoint c_n. Here, x_ψ^c represents the rendered image of the 3D input ψ from the same viewpoint c. Additionally, considering that the structure of a 4D object can gradually diverge more from the 3D geometry over time, we introduce a temporal gradient coefficient $\frac{\mathcal{T}-t}{\mathcal{T}}\lambda$, where λ is a hyperparameter used to control the scales of $\mathcal{L}_{\text{mot}}$ and $\mathcal{L}_{\text{geo}}$. At this point, we can derive the $\mathcal{L}_{\text{USVA}}$ to optimize unknow spatiotemporal viewpoint expression of 4D object:

$$\mathcal{L}_{\text{USVA}} = \frac{1}{\mathcal{T}}\sum_{t=1}^{\mathcal{T}} \frac{t}{\mathcal{T}}\lambda\mathcal{L}_{\text{mot}} + \frac{\mathcal{T}-t}{\mathcal{T}}\lambda\mathcal{L}_{\text{geo}}. \tag{9}$$

From the above description, we can now derive the overall optimization objective $\mathcal{L}_{\text{MGJA}}$:

$$\mathcal{L}_{\text{MGJA}} = \mathcal{L}_{\text{KSVA}} + \mathcal{L}_{\text{USVA}}. \tag{10}$$

Fig. 7. **Samples from X4D dataset.** Each quadruplet consists of a prompt, a generated image, a generated video, and a generated 3D object, all created by off-the-shelf diffusion models.

D. Asynchronous Optimization

The optimization objective $\mathcal{L}_{\text{MGJA}}$ aims to refine the 4D representation by integrating both known and unknown temporal perspectives to align effectively with video and 3D data. However, employing a synchronous optimization approach, as seen in previous works [7], can introduce instability, potentially leading to suboptimal generation results. To overcome this challenge, we propose an asynchronous optimization framework. As illustrated in Figure 3, our method strategically alternates between fixing the 3DGS and the deformation network while optimizing the other. This asynchronous strategy ensures that when the 3DGS geometry is suboptimal, the deformation network can compensate. When the deformation network's capacity to drive motion is inadequate, the 3DGS reinforces the geometric structure. This approach fosters a more balanced integration of geometry and motion.

E. X4D Dataset

Figure 7 presents representative samples from our X4D dataset, which is built from four types of aligned multimodal quadruplets (text, image, video, and 3D). We construct the dataset through four complementary pipelines: (a) **Prompt-driven pipeline.** A textual prompt (from MAV3D [4]) is first used to generate an image via SDXL [57]. The same prompt then drives SVD [19] to produce a video, and LGM [20] to reconstruct the corresponding 3D Gaussian representation. (b) **Image-driven pipeline.** Given a single input image, we retrieve visually related content from the web, synthesize a short video using SVD, reconstruct its 3D Gaussian model with LGM, and produce a paired textual description using image-to-prompt tool [58]. (c) **Video-driven pipeline.** For videos collected from Kling [59], we extract the first frame as the reference image for 3D reconstruction. Each video is also processed by an image-to-text model to obtain its textual prompt. (d) **3D-driven pipeline.** 3D assets from an open-source repository [22] are rendered from canonical viewpoints. The rendered image is then used for video synthesis and textual description generation, forming a complete text–image–video–3D quadruplet. Together, these four pipelines produce consistently aligned multimodal quadruplets, enabling the creation of a large-scale, diverse dataset tailored for advancing 4D generation research.

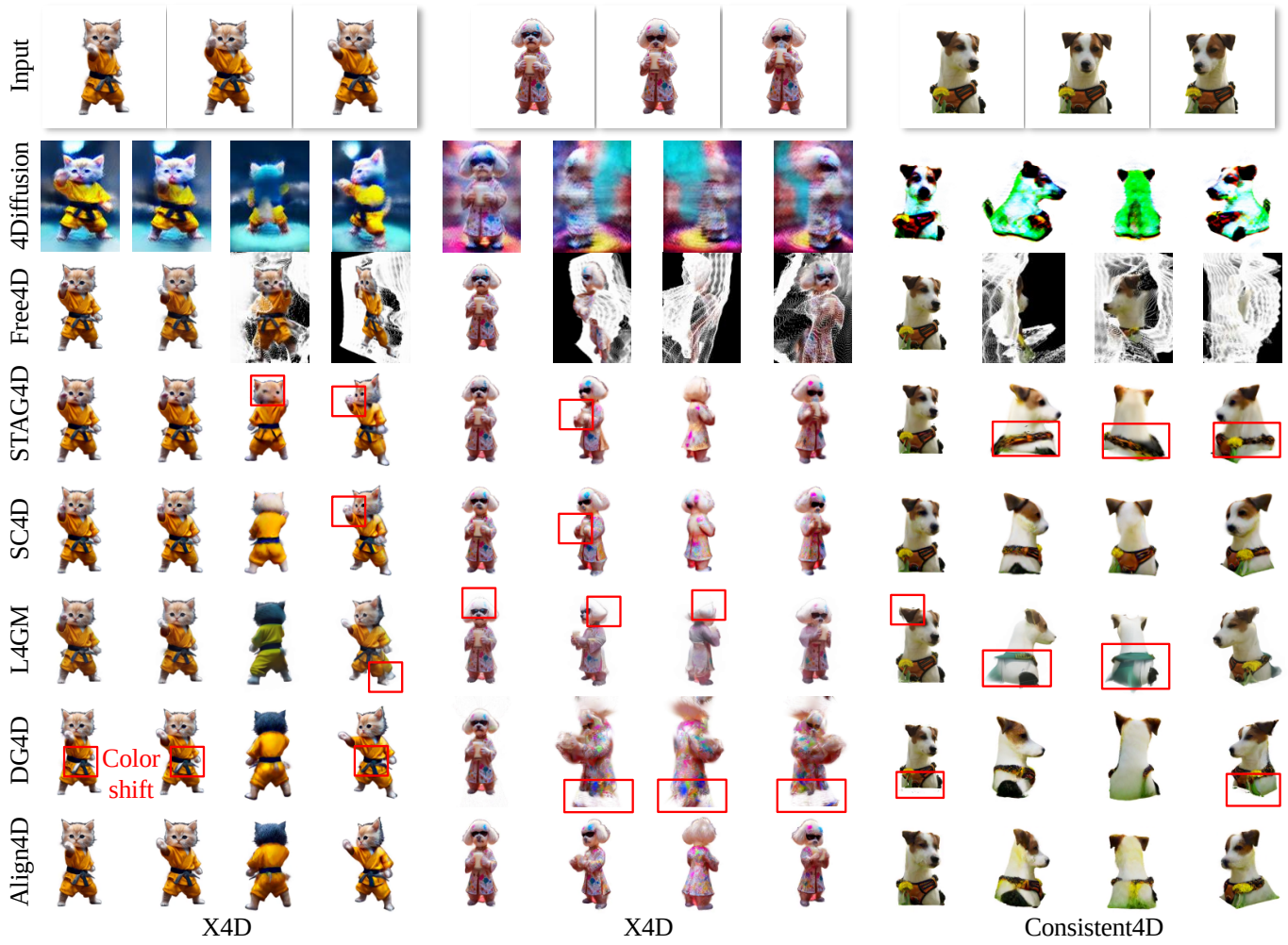


Fig. 8. **Qualitative comparisons between Align4D and other methods on our X4D and Consistent4D datasets.** Align4D is capable of generating detailed, geometrically consistent 4D targets that faithfully replicate the motion observed in the video. To better appreciate these qualities, please zoom in.

The construction of X4D inevitably inherits statistical biases from its upstream generative components—SDXL [27], SVD [19], and LGM [20]—each of which is trained on heterogeneous data sources. To mitigate cross-modal distribution mismatch, we intentionally combine models originating from diverse training corpora, with prompts generated via an image-to-prompt tool [58]. All assets, including text, images, videos, and 3D models, are collected from openly licensed community datasets. During cross-modal expansion using generative tools such as SDXL and Kling, all generated samples undergo strict manual filtering to ensure data quality and regulatory compliance.

IV. Experiment

A. Experimental Settings

Implementation Details. In Align4D, we set the dense percentage to 0.1, the densification interval to 100, and the densification gradient threshold to 0.05. All other settings follow DG4D [7]. In particular, DG4D uses a fixed object distance of 1.5 to align with the video, and sets the multi-view diffusion model parameter to 0 for SDS optimization, which we also adopt in our ablation experiments. During the object distance search, we set $d_{\text{min}} = 0.00001$ and $d_{\text{max}} = 3.00001$, performing searches at intervals of 0.05. The number of viewpoints is set to $N = 4$, and we use Zero123 [53] as the multiview diffusion model. All experiments are conducted on a system equipped with NVIDIA V100 GPUs with 32 GB memory each, an Intel Xeon processor (Skylake, IBRS), and 629 GB of RAM. The software environment includes Ubuntu 22.04.3 LTS with CUDA 12.4.

Dataset. We conduct experiments on the proposed X4D dataset, which contains aligned text–image–video–3D quadruplets generated through our unified multimodal pipeline, as shown in Figure 1. These quadruplets provide consistent input conditions across tasks, allowing fair comparisons among text-to-4D, image-to-4D, video-to-4D, and 3D-to-4D generation. To further evaluate performance, we also use the Consistent4D dataset [5], specifically designed for video-to-4D task. Both quantitative and qualitative experiments on X4D and Consistent4D assess the geometric consistency, motion coherence, and fidelity of the generated 4D outputs relative to the input control conditions.

Metrics. We evaluate generation quality using PSNR [60], SSIM [61], LPIPS [62], CLIP similarity [63], and FVD [64], following established practice in 4D generation. PSNR, SSIM,

TABLE I
QUANTITATIVE RESULTS OF ALIGN4D ON OUR X4D DATASET.

Methods	Human Evaluation				VBench			
	Appearance% ↑	Structure% ↑	Motion% ↑	Fidelity% ↑	Subject Consistency ↑	Bckground Consistency ↑	Aesthetic Quality ↑	Imaging Quality ↑
L4GM [8]	9.2	5.5	7.9	6.6	0.80	0.88	0.46	0.40
SC4D [10]	5.3	5.3	4.0	5.3	0.81	0.91	0.43	0.38
STAG4D [48]	19.7	17.1	26.3	18.4	0.83	0.91	0.50	0.41
DG4D [7] (baseline)	4.0	2.6	3.9	3.9	0.71	0.85	0.40	0.31
Align4D (ours)	**61.8**	**69.5**	**57.9**	**65.8**	**0.85**	**0.93**	**0.55**	**0.43**

LPIPS, and CLIP similarity are adopted as image-level metrics to assess perceptual and semantic alignment between rendered images and references. FVD serves as a video-level metric widely used in generative video evaluation, capturing both frame-level fidelity and temporal coherence. Additionally, we incorporate the CLIP-F score [65], defined as the average CLIP similarity between adjacent frames, to evaluate 4D rendering and quantify temporal semantic consistency. We further incorporate VBench [66] to assess 360° surround-view renderings. VBench provides model-based scores along four dimensions—subject consistency, background consistency, aesthetic quality, and imaging quality—allowing a comprehensive evaluation of rendering quality.

Baselines. To ensure fair cross-modal comparison, all baselines are adapted to our unified X-to-4D framework, receiving their respective modality from the structured X4D quadruplets as input. This allows diverse methods to generate the same 4D targets under consistent conditions. We include representative image-to-4D models [3], [7], [67] and state-of-the-art video-to-4D methods [8]–[10], [48], [68]. Text-to-4D [2] is excluded due to substantial differences in motion and geometric control. The standard 3D-to-4D method [6] is impractical due to extreme memory demands; instead, we use 3D tools like CRM [21] and MeshyAI [22] with rigging for a feasible comparison. Video generation models such as Kling [59] are additionally included, producing sequences under identical text or image conditions to ensure evaluation consistency. All methods are evaluated on X4D using the same quadruplets, ensuring fair comparison across modalities. We also report results on the Consistent4D dataset [5], which provides curated sequences for standardized assessment of motion fidelity, appearance consistency, and adherence to input conditions.

B. Comparisons to State-of-the-Art Methods

Qualitative evaluation. We provide extensive visual comparisons on the X4D dataset, where the input data is generated by pretrained generative models. As shown in Figure 8, Align4D produces high-quality, sharp renderings with strong temporal and multiview consistency. In comparison, 4Diffusion [68], despite being a large-scale multiview video diffusion model, struggles to generalize to the diverse test scenarios. Free4D [67], while integrating the powerful 4D reconstruction network MonST3R [71], often fails to generate geometrically detailed and temporally coherent 4D targets that are viewable from arbitrary angles. STAG4D [48] generally produces geometric artifacts, whereas SC4D [10] reduces geometric errors but lacks fine-grained detail. L4GM [8] achieves fast generation, yet exhibits noticeable geometric inconsistencies. DG4D [7] can approximate the frontal-view video visually; however, it occasionally deviates from the video or introduces substantial errors in other viewpoints. By contrast, Align4D consistently delivers the most refined 4D outputs, preserving dynamic motion and geometric fidelity across all views.

TABLE II
QUANTITATIVE RESULTS OF ALIGN4D ON THE CONSISTENT4D DATASET. THE BEST SCORES ARE HIGHLIGHTED IN **BOLD**. ∗ INDICATES GENERATING VIDEOS USING THE FIRST AND LAST FRAMES OF THE TRAINING VIDEO. + INDICATES GENERATING VIDEOS USING THE FIRST AND LAST FRAMES OF THE TEST VIDEO.

Methods	Consistent4D Dataset [5]					
	PSNR↑	SSIM↑	LPIPS↓	FVD↓	CLIP↑	CLIP-F↑
CRM [69] + Rigging	13.5	0.88	0.17	1258.5	0.79	-
Meshyai [22] + Rigging	12.7	0.87	0.19	1179.2	0.85	-
Kling* [59]	11.0	0.81	0.28	-	0.81	-
Kling+ [59]	11.4	0.80	0.26	-	0.78	-
4Diffusion [68]	12.2	0.84	0.26	1522.9	0.83	0.912
Free4D [67]	6.4	0.45	0.41	2513.6	0.77	0.809
L4GM [8]	14.2	0.84	0.20	1217.1	0.90	0.990
SC4D [10]	16.8	0.86	0.16	1132.2	0.91	0.988
4DGen [70]	12.7	0.87	0.19	1258.5	0.80	0.981
Efficient4D [9]	12.8	0.85	0.21	1304.2	0.89	0.954
STAG4D [48]	17.0	0.87	0.14	1251.7	0.90	0.988
DG4D [7] (baseline)	10.7	0.78	0.28	1262.0	0.89	0.978
Align4D (ours)	**17.8**	**0.90**	**0.11**	**1088.9**	**0.94**	**0.992**

TABLE III
QUANTITATIVE RESULTS FROM ABLATION STUDIES OF ALIGN4D ON CONSISTENT4D DATASET.

Methods	Consistent4D Dataset [5]				
	PSNR↑	SSIM↑	LPIPS↓	FVD↓	CLIP↑
STAG4D [48]	17.0	0.87	0.14	1251.7	0.90
STAG4D [48] + ODA + MGJA	17.5	0.88	0.12	1109.4	0.93
STAG4D [48] + AO	17.2	0.88	0.14	1129.2	0.91
DG4D [7] (baseline)	10.7	0.78	0.28	1262.0	0.89
Align4D W/O. ODA	15.9	0.86	0.16	1111.5	0.90
Align4D W/O. MGJA	15.5	0.85	0.17	1113.1	0.91
Align4D W/O. AO	17.6	0.89	0.13	1139.7	0.93
Align4D (ours)	**17.8**	**0.90**	**0.11**	**1088.9**	**0.94**

Quantitative evaluation. We first conduct comparisons on the X4D dataset. All models generate 4D targets conditioned on the same quadruplet inputs and render full 360° surround-view videos. A user study with 30 participants evaluates the generated results across four dimensions: Appearance, Structure, Motion, and Fidelity. Align4D consistently receives the highest human preference. Additionally, the rendered videos are evaluated using VBench on subject consistency, background consistency, aesthetic quality, and imaging quality,

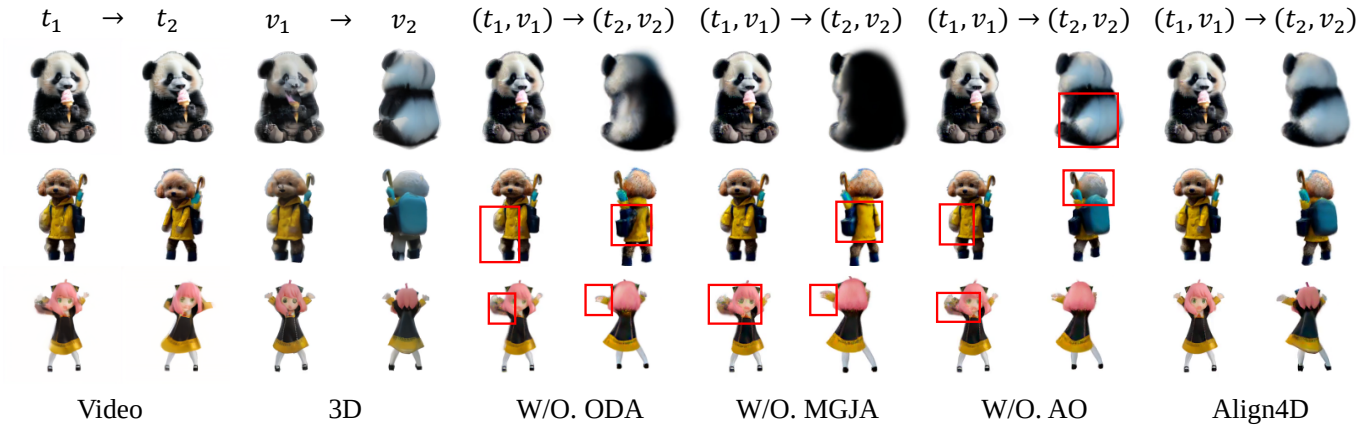


Fig. 9. **Qualitative ablation experiments on the X4D dataset for Align4D.** ODA is crucial for precisely injecting action and geometric data into video and multiview diffusion models. MGJA significantly enhances the rendering of 4D non-frontal, multi-moment views by meticulously aligning 4D actions to video and geometry to 3D inputs. Additionally, AO refines the intricate details of the generated 4D targets.

TABLE IV
COMPARISON OF LOSS REDUCTION BETWEEN ASYNCHRONOUS OPTIMIZATION (AO) AND JOINT OPTIMIZATION (JO).

Methods \ Step	100	200	300	400	500
Align4D + JO	72.3	174.2	155.2	44.0	3.5
Align4D + AO	34.4	87.7	85.7	24.8	3.2

TABLE V
RUNTIME COMPARISON BETWEEN ALIGN4D AND BASELINE METHODS. ONLY THE TIME SPENT ON GENERATING 4D TARGETS IS MEASURED, WHILE THE OVERHEAD FOR DATA PREPARATION IS EXCLUDED.

Methods	Time (Min.)	VRAM (GB)
L4GM [8]	1.5	24.6
SC4D [10]	35	8.9
STAG4D [48]	80	12.2
DG4D [7]	15	15.6
Align4D (ours)	25	19.4
ODA-VAOD	0.0025	2.7
ODA-MAOD	0.20	4.5

where Align4D also achieves the top scores. Furthermore, Table II reports the quantitative results on the Consistent4D dataset. Compared to state-of-the-art video-to-4D methods [3], [7]–[10], [48], Align4D attains the best performance across five key metrics, covering both image-level and video-level evaluations. This demonstrates that Align4D effectively integrates the geometric consistency of 3D data with the motion coherence of video data, successfully transferring these properties to novel, unseen views. In contrast, 3D generation models [21], [22] combined with manual rigging struggle to faithfully reproduce object motion, and video generation models [59] fail to maintain motion consistency under test-view conditions. Overall, Align4D exhibits clear advantages over these alternatives, producing more accurate, temporally coherent, and geometrically consistent 4D results.

C. Ablation Studies

The qualitative results are shown in Figure 9. As can be seen, the absence of object distance alignment leads not only to a blurred frontal view but also causes the back view to deviate from the 3D data priors when not using ODA for SDS optimization, resulting in erroneous coverage modifications. Without MGJA, unknown views, especially the rear view of the 4D generation target, exhibit more severe errors and inconsistencies. If asynchronous optimization is not used, although there is good alignment of dynamics and geometry with the input video and 3D data in both frontal and other views, some details lack refinement and there is a tendency to inherit imperfections from imperfect video or 3D data. Asynchronous optimization effectively alleviates these issues. When ODA, MGJA, and AO are used together, forming the complete Align4D, the 4D target achieves fidelity to video motion, 3D geometry, and exhibits detailed precision.

The quantitative results are shown in Table III. By synergistically employing ODA, MGJA, and AO, Align4D effectively addresses limitations in generated video-3D data, producing 4D targets with motion and geometry better aligned with human perception. ODA is a prerequisite for robust MGJA, disabling it significantly degrades performance (Table III W/O. ODA). MGJA then refines motion and geometry alignment using ODA's precise object distances; the effectiveness hinges on ODA's quality (Table III W/O. MGJA). AO further enhances MGJA's alignment by decoupling 3DGS and deformable network optimization, yielding finer and more stable convergence by mitigating component interference (Table III W/O. AO). Besides, Table V reports the computational costs of Align4D and baseline methods, highlighting the low overhead of VAOD and MAOD searches. In summary, the ODA module estimates accurate object distances, MGJA ensures robust multiview and temporal alignment, and AO facilitates optimal convergence. Together, these components allow Align4D to achieve state-of-the-art generation quality with balanced computational efficiency.

Module migratability. We successfully migrate our modules to the STAG4D method. We specifically integrate ODA, MGJA, and AO into STAG4D (see Table III). Notably, MGJA's synergistic function relies on the optimal object distance provided by ODA; thus, these two modules are inherently bound together for migration. This demonstrates that ODA,

Fig. 10. **Generation results comparing Align4D with text-to-4D method [2].** Given the same text prompts, Align4D delivers markedly superior geometric structure and a significantly wider motion range.

MGJA, and AO from Align4D can be directly transferred to other models like STAG4D, providing performance gains across multiple metrics compared to the STAG4D baseline.

Asynchronous optimization effectiveness. We further compared the effects of joint optimization (JO) and asynchronous optimization (AO) during the training phase to highlight their differences. As shown in Table IV, under the same number of steps, asynchronous optimization resulted in smaller fluctuations and a faster decrease in loss, using an SDS loss as an example. It also achieved a lower loss value at the same number of steps. This indicates that decoupling the optimization of the deformation network and 3DGS is more beneficial for aligning 4D objects with the input conditions.

D. Comparison with Text-to-4D generation method

Compared to the text-to-4D generation method 4D-Fy [2], the dynamic quality of 4D outputs controlled solely by text prompts is noticeably inferior to those guided by images or videos, even when using the same prompts, as illustrated in Figure 10. Moreover, text-to-4D methods cannot accept inputs from other modalities, such as images or videos, making both qualitative and quantitative evaluation on the Consistent4D dataset infeasible. In contrast, Align4D's flexible X-to-4D generation pipeline not only supports a wide range of input modalities but also leverages the smooth motion from the generated video and the fine-grained geometric guidance provided by the generated 3D object.

V. Disscussion

A. Why MAOD Should Be Smaller Than VAOD

MAOD is designed to approximate the intrinsic object distance preferred by the multiview diffusion model, which often

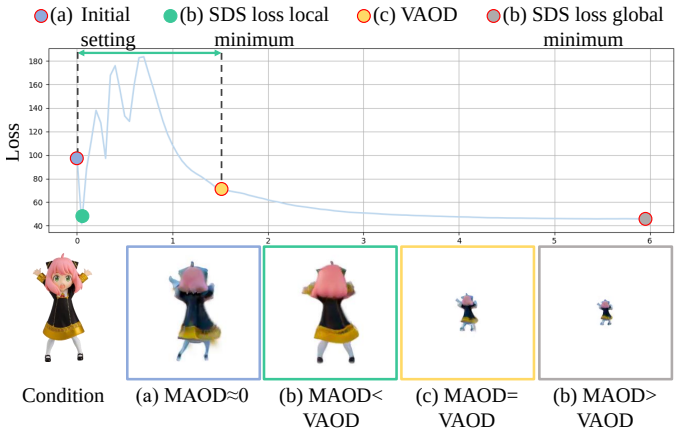


Fig. 11. **Generation results under different MAOD selections.** (a) A small object distance (OD) used as MAOD introduces color shifts and loss of fine details. (b) An SDS-loss local-minimum OD used as MAOD achieves the best input fidelity. (c) Using VAOD as MAOD produces an under-scaled object with reduced high-frequency detail. (d) An SDS-loss global-minimum OD used as MAOD likewise leads to under-scaling.

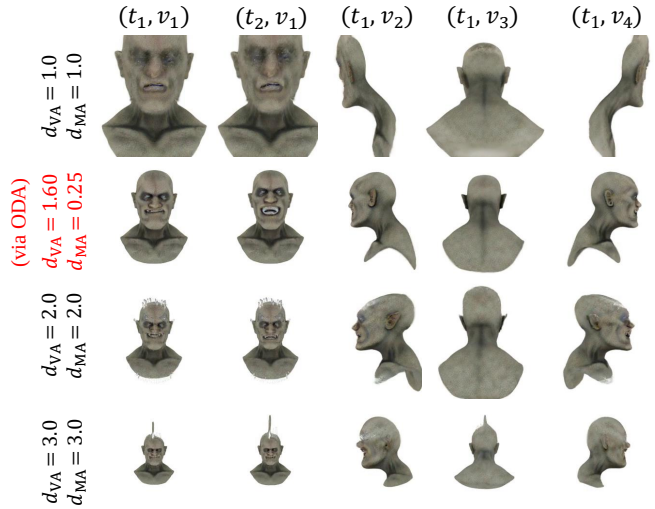


Fig. 12. **4D generation results under different object distances.** Only when the optimal values are found using the object distance alignment method for d_{VA} and d_{MA}, can we ensure the generation of 4D targets with faithful motion and accurate geometry.

differs from the viewpoint-aligned object distance (VAOD). Figure 11 illustrates this effect by varying MAOD while keeping VAOD fixed: (a) **MAOD** ≈ 0 **(initial setting)**: The generated results exhibit local blurring and overall color shifts, indicating that the multiview diffusion model cannot accurately infer the target geometry under the current object distance setting. (b) **MAOD** $<$ **VAOD (local SDS-loss minimum)**: The generated geometry is complete, and the rendered appearance closely matches the input, suggesting that this distance best aligns with the diffusion model's implicit distance prior. (c) **MAOD = VAOD**: The generated object appears underscaled with color inconsistencies, indicating that VAOD does not capture the diffusion model's internal understanding of object distance, and thus MAOD is required for proper alignment. (d) **MAOD** $>$ **VAOD (global SDS-loss minimum)**: Although this setting achieves the lowest SDS loss, it produces unrealistic object scale and appearance, demonstrating that global SDS-

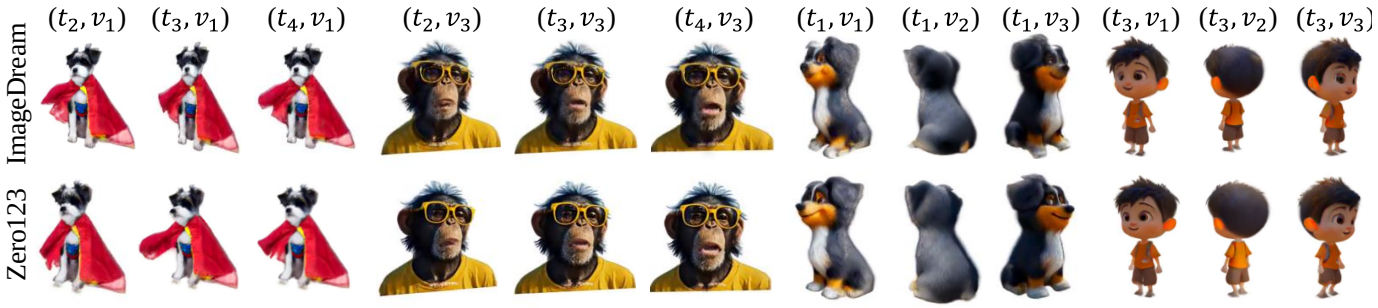


Fig. 13. **Align4D with different multiview diffusion models.** We incorporate two distinct multiview diffusion models, Zero123 [41] and ImageDream [72], both of which produce outputs with accurate geometric structures and temporally coherent motion. These results further demonstrate that Align4D is compatible with a variety of multiview diffusion backbones.

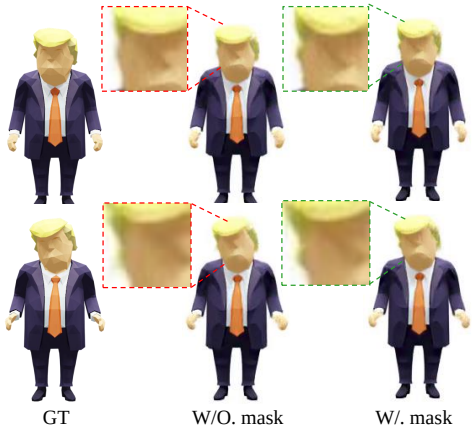


Fig. 14. **Effect of masks on 4D generation results.** Incorporating masks further enhances boundary fidelity and preserves fine edge details in the generated 4D content.

Fig. 15. **Visualization of diversity in generated results shows that by using different random seeds.** Align4D can produce 4D outputs with significant variations while faithfully adhering to the input conditions.

loss minimization does not correspond to a physically meaningful distance alignment. These observations indicate that the diffusion model favors an object distance slightly smaller than VAOD, typically near a local SDS-loss minimum, making it the most appropriate choice for defining MAOD.

B. Generation Quality Comparison Across Object Distances

Figure 12 shows a visual comparison of 4D generation results under different object distances. Using an unmatched video object distance causes noticeable blur in frontal-view actions and introduces conflicts with the multiview diffusion model supervision, resulting in geometric inaccuracies. Similarly, an unmatched multiview object distance leads to misalignment of the target geometry across viewpoints. By applying Object Distance Alignment (ODA), the matched object distances ensure consistent video motion and 3D object geometry, producing coherent and high-quality 4D assets.

C. Effect of Masks

Following DG4D [7], we employ RemoveBg [73], based on a pretrained U^{2}Net [74], to extract frame-wise masks that guide generation toward better alignment with the input video geometry. Figure 14 presents an ablation study on the mask loss. The results indicate that Align4D can generate high-quality geometry and maintain motion consistency even without masks. However, incorporating masks improves the preservation of fine local details and achieves a closer visual match to the input, particularly in the frontal view.

D. Stability of Generation Across Different Diffusion Models

MAOD exhibits broad compatibility with various diffusion models. Methods such as Zero123 [53] allow explicit control of the object distance by directly using camera parameters as inputs. In contrast, multiview diffusion models represented by ImageDream [72] employ a fixed distance schedule: they rely on several predefined camera parameters to generate multiview outputs, and their pretrained weights implicitly encode a fixed viewing distance. Consequently, only an extremely lightweight modification to Equation 5 is required for adaptation to these models. Specifically, within the distance range $\mathcal{D} = [d_{\min}, d_{\max}]$, we render image sets $\left\{\left\{x_{\theta_1}^{c,d'}\right\}_{c\in C}\right\}_{d'\in\mathcal{D}}$, where c corresponds to three horizontally varying azimuth angles $[-90°, 0°, 90°]$, which lie within the azimuth range covered by the ImageDream training data. The SDS loss for each candidate distance d' is computed as:

$$\mathcal{L}_{\text{SDS}}^{d'} = \frac{1}{|C|\,|\mathrm{T}|}\sum_{c\in C}\sum_{\tau\in\mathrm{T}} w(\tau)\left\|\epsilon_\phi\left(z_{\theta_1}^{c,d'}; I_1, c, \tau\right) - \epsilon\right\|_2^2. \tag{11}$$

We select the distance d' that is smaller than VAOD and corresponds to a local minimum of the SDS loss as the MAOD.

TABLE VI
USER PREFERENCE FOR RESULTS GENERATED BY DIFFERENT MODELS INTEGRATED INTO THE ALIGN4D PIPELINE

3D model \ Video model	Kling [59]	VideoCrafter [17]
Meshai [22]	27.4%	22.5%
Tripo3d [75]	26.5%	23.6%

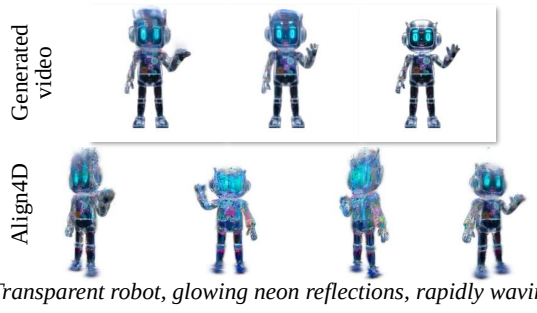


Fig. 16. **Failure case.** The generated 4D assets exhibit limited fine detail when handling transparent or rapidly flickering neon objects.

Figure 13 presents results generated using different diffusion models. Align4D produces highly consistent outputs across multiple diffusion backbones and remains stable across diverse examples, demonstrating that MGJA in Align4D is broadly applicable to various multiview diffusion models.

E. Align4D with Different Seeds

To investigate the impact of random seeds on 4D generation, we use the text-to-4D task as an example, since it involves the most extensive data processing through multiple diffusion models. Given a fixed prompt, an image is first generated via SDXL, followed by video synthesis with SVD and 3D generation with LGM to construct a video–3D pair. Align4D then generates the corresponding 4D object. As shown in Figure 15, while different seeds introduce stylistic variations in the video–3D pair, the resulting 4D objects consistently preserve the input text semantics and maintain stable geometric structure across seeds.

F. Robustness Across Diverse Video and 3D Generation Models

Align4D operates on video–3D pairs generated by various diffusion models, regardless of the underlying architecture. Table VI presents results obtained using different combinations of video and 3D generative models. For each configuration, the same conditioning input is used to synthesize the video–3D pair, which Align4D then processes to produce the final 4D output. A user study with 30 participants further evaluates these results. Across all combinations, Align4D receives comparable user preferences, demonstrating that the framework consistently delivers high-quality 4D generation, thereby further validating the model versatility of Align4D.

G. Failure Analysis

We further analyze scenarios where Align4D underperforms. When using a text-to-image model followed by an image-to-video model, with the resulting image-to-3D output to make a video-3D pair as input, the generated 4D assets inherit limitations from the underlying diffusion models. In particular, current text-to-image and image-to-video diffusion models struggle with transparent materials and rapidly changing neon effects. Consequently, Align4D's 4D outputs also reflect these deficiencies. We expect that as generative models improve in expressiveness, these limitations will be mitigated.

VI. CONCLUSION

In this paper, we propose Align4D, an alignment-based framework designed for flexible X-to-4D generation. This framework leverages multimodal inputs combined with pre-trained diffusion models to produce paired video-3D data . Align4D introduces object distance-based alignment, searching for the video-aligned object distance and the multiview-aligned object distance that achieves optimal generalization for multiview diffusion models. Based on this, the framework aligns the front-view renderings of 4D targets with the video and conditions unknown non-front view multi-timestep renderings on the same-timestep front-view video frames and the same-view initial-timestep 3D renderings. By employing SDS optimization, Align4D achieves joint alignment of motion and geometry. Furthermore, asynchronous optimization is utilized to refine the 4D target for better motion and geometry representation. To evaluate the open X-to-4D generation task, we propose a quadruplet dataset, X4D, consisting of (prompt, image, video, 3D). Through extensive testing in generated scenes, real-world scenarios, and synthetic environments, Align4D demonstrates exceptional generative capabilities.

REFERENCES

[1] H. Ling, S. W. Kim, A. Torralba, S. Fidler, and K. 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*, 2024, pp. 8576–8588.
[2] S. Bahmani, I. Skorokhodov, V. Rong, G. Wetzstein, L. Guibas, P. Wonka, S. Tulyakov, J. J. Park, A. Tagliasacchi, and D. 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*, 2024, pp. 7996–8006.
[3] Y. Yin, D. Xu, Z. Wang, Y. Zhao, and Y. Wei, "4dgen: Grounded 4d content generation with spatial-temporal consistency," *arXiv preprint arXiv:2312.17225*, 2023.
[4] U. Singer, S. Sheynin, A. Polyak, O. Ashual, I. Makarov, F. Kokkinos, N. Goyal, A. Vedaldi, D. Parikh, J. Johnson, and Y. Taigman, "Text-to-4d dynamic scene generation," in *International Conference on Machine Learning*, 2023.
[5] Y. Jiang, L. Zhang, J. Gao, W. Hu, and Y. Yao, "Consistent4d: Consistent 360° dynamic object generation from monocular video," in *International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=sPUrdFGepF
[6] Y. Zhao, Z. Yan, E. Xie, L. Hong, Z. Li, and G. H. Lee, "Animate124: Animating one image to 4d dynamic scene," *arXiv preprint arXiv:2311.14603*, 2023.
[7] J. Ren, L. Pan, J. Tang, C. Zhang, A. Cao, G. Zeng, and Z. Liu, "Dreamgaussian4d: Generative 4d gaussian splatting," *arXiv preprint arXiv:2312.17142*, 2023.
[8] J. Ren, C. Xie, A. Mirzaei, K. Kreis, Z. Liu, A. Torralba, S. Fidler, S. W. Kim, H. Ling *et al.*, "L4gm: Large 4d gaussian reconstruction model," *Annual Conference on Neural Information Processing Systems*, vol. 37, pp. 56 828–56 858, 2024.
[9] Z. Pan, Z. Yang, X. Zhu, and L. Zhang, "Efficient4d: Fast dynamic 3d object generation from a single-view video," *arXiv preprint arXiv:2401.08742*, 2024.

[10] Z. Wu, C. Yu, Y. Jiang, C. Cao, F. Wang, and X. Bai, "Sc4d: Sparse-controlled video-to-4d generation and motion transfer," in *European Conference on Computer Vision*. Springer, 2024, pp. 361–379.

[11] Z. Yang, Z. Pan, C. Gu, and L. Zhang, "Diffusion^2: Dynamic 3d content generation via score composition of video and multi-view diffusion models," in *International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=fectsEG2GU

[12] H. E. Pang, S. Liu, Z. Cai, L. Yang, T. Zhang, and Z. Liu, "Disco4d: Disentangled 4d human generation and animation from a single image," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 26 331–26 344.

[13] Y. Wang, X. Wang, Z. Chen, Z. Wang, F. Sun, and J. Zhu, "Vidu4d: Single generated video to high-fidelity 4d reconstruction with dynamic gaussian surfels," *Annual Conference on Neural Information Processing Systems*, vol. 37, pp. 131 316–131 343, 2024.

[14] Y. Xie, C.-H. Yao, V. Voleti, H. Jiang, and V. Jampani, "SV4d: Dynamic 3d content generation with multi-frame and multi-view consistency," in *International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=tJoS2d0Onf

[15] H. Liang, Y. Yin, D. Xu, H. Liang, Z. Wang, K. N. Plataniotis, Y. Zhao, and Y. Wei, "Diffusion4d: fast spatial-temporal consistent 4d generation via video diffusion models," in *Annual Conference on Neural Information Processing Systems*, 2024.

[16] H. Chen, M. Xia, Y. He, Y. Zhang, X. Cun, S. Yang, J. Xing, Y. Liu, Q. Chen, X. Wang *et al.*, "Videocrafter1: Open diffusion models for high-quality video generation," *arXiv preprint arXiv:2310.19512*, 2023.

[17] H. Chen, Y. Zhang, X. Cun, M. Xia, X. Wang, C. Weng, and Y. Shan, "Videocrafter2: Overcoming data limitations for high-quality video diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 7310–7320.

[18] T. Brooks, B. Peebles, C. Holmes, W. DePue, Y. Guo, L. Jing, D. Schnurr, J. Taylor, T. Luhman, E. Luhman, C. Ng, R. Wang, and A. Ramesh, "Video generation models as world simulators," 2024. [Online]. Available: https://openai.com/research/video-generation-models-as-world-simulators

[19] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts *et al.*, "Stable video diffusion: Scaling latent video diffusion models to large datasets," *arXiv preprint arXiv:2311.15127*, 2023.

[20] J. Tang, Z. Chen, X. Chen, T. Wang, G. Zeng, and Z. Liu, "Lgm: Large multi-view gaussian model for high-resolution 3d content creation," in *European Conference on Computer Vision*. Springer, 2024, pp. 1–18.

[21] Z. Wang, Y. Wang, Y. Chen, C. Xiang, S. Chen, D. Yu, C. Li, H. Su, and J. Zhu, "Crm: Single image to 3d textured mesh with convolutional reconstruction model," in *European Conference on Computer Vision*, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:268297409

[22] Meshy AI, "Meshy AI: The #1 ai 3d model generator for creators," https://www.meshy.ai/, 2025, accessed: 2025-04-17.

[23] Z. Li, Y. Chen, and P. Liu, "Dreammesh4d: Video-to-4d generation with sparse-controlled gaussian-mesh hybrid representation," in *Annual Conference on Neural Information Processing Systems*, 2024.

[24] B. Poole, A. Jain, J. T. Barron, and B. Mildenhall, "Dreamfusion: Text-to-3d using 2d diffusion," in *International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=FjNys5c7VyY

[25] Q. Miao, K. Li, J. Quan, Z. Min, S. Ma, Y. Xu, Y. Yang, P. Liu, and Y. Luo, "Advances in 4d generation: A survey," 2025. [Online]. Available: https://arxiv.org/abs/2503.14501

[26] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2022, pp. 10 684–10 695.

[27] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," in *Annual Conference on Neural Information Processing Systems*, 2022.

[28] Y. Xu, X. Xu, H. Gao, and F. Xiao, "Sgdm: An adaptive style-guided diffusion model for personalized text to image generation," *IEEE Transactions on Multimedia*, vol. 26, pp. 9804–9813, 2024.

[29] J. Xiao and X. Bi, "Model-guided generative adversarial networks for unsupervised fine-grained image generation," *IEEE Transactions on Multimedia*, vol. 26, pp. 1188–1199, 2024.

[30] A. Blattmann, R. Rombach, H. Ling, T. Dockhorn, S. W. Kim, S. Fidler, and K. 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*, 2023, pp. 22 563–22 575.

[31] S. Ge, S. Nah, G. Liu, T. Poon, A. Tao, B. Catanzaro, D. Jacobs, J.-B. Huang, M.-Y. Liu, and Y. Balaji, "Preserve your own correlation: A noise prior for video diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22 930–22 941.

[32] R. Girdhar, M. Singh, A. Brown, Q. Duval, S. Azadi, S. S. Rambhatla, A. Shah, X. Yin, D. Parikh, and I. Misra, "Factorizing text-to-video generation by explicit image conditioning," in *European Conference on Computer Vision*. Springer, 2024, pp. 205–224.

[33] Y. Guo, C. Yang, A. Rao, Z. Liang, Y. Wang, Y. Qiao, M. Agrawala, D. Lin, and B. Dai, "Animatediff: Animate your personalized text-to-image diffusion models without specific tuning," in *International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=Fx2SbBgcte

[34] J. Ho, W. Chan, C. Saharia, J. Whang, R. Gao, A. Gritsenko, D. P. Kingma, B. Poole, M. Norouzi, D. J. Fleet *et al.*, "Imagen video: High definition video generation with diffusion models," *arXiv preprint arXiv:2210.02303*, 2022.

[35] L. Khachatryan, A. Movsisyan, V. Tadevosyan, R. Henschel, Z. Wang, S. Navasardyan, and H. Shi, "Text2video-zero: Text-to-image diffusion models are zero-shot video generators," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 15 954–15 964.

[36] A. Köksal, K. E. Ak, Y. Sun, D. Rajan, and J. H. Lim, "Controllable video generation with text-based instructions," *IEEE Transactions on Multimedia*, vol. 26, pp. 190–201, 2024.

[37] X. Zhang, C. Zhang, J. Xu, Y. Zhu, X. Shi, Y. Yang, and Y. Luo, "Video2roleplay: A multimodal dataset and framework for video-guided role-playing agents," in *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025, pp. 23 688–23 714.

[38] J. Zhu, H. Ma, J. Chen, and J. Yuan, "Motionvideogan: A novel video generator based on the motion space learned from image pairs," *IEEE Transactions on Multimedia*, vol. 25, pp. 9370–9382, 2023.

[39] U. Singer, A. Polyak, T. Hayes, X. Yin, J. An, S. Zhang, Q. Hu, H. Yang, O. Ashual, O. Gafni, D. Parikh, S. Gupta, and Y. Taigman, "Make-a-video: Text-to-video generation without text-video data," in *International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=nJfylDvgzlq

[40] Y. Shi, P. Wang, J. Ye, L. Mai, K. Li, and X. Yang, "MVDream: Multi-view diffusion for 3d generation," in *International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=FUgrjq2pbB

[41] R. Shi, H. Chen, Z. Zhang, M. Liu, C. Xu, X. Wei, L. Chen, C. Zeng, and H. Su, "Zero123++: a single image to consistent multi-view diffusion base model," *arXiv preprint arXiv:2310.15110*, 2023.

[42] X. Long, Y.-C. Guo, C. Lin, Y. Liu, Z. Dou, L. Liu, Y. Ma, S.-H. Zhang, M. Habermann, C. Theobalt *et al.*, "Wonder3d: Single image to 3d using cross-domain diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9970–9980.

[43] F. Yin, X. Chen, C. Zhang, B. Jiang, Z. Zhao, W. Liu, G. Yu, and T. Chen, "Shapegpt: 3d shape generation with a unified multi-modal language model," *IEEE Transactions on Multimedia*, vol. 27, pp. 4107–4120, 2025.

[44] Z. Ye, Y. Liu, and Y. Peng, "Maan: Memory-augmented auto-regressive network for text-driven 3d indoor scene generation," *IEEE Transactions on Multimedia*, vol. 26, pp. 11 057–11 069, 2024.

[45] J. Li, Y. Luo, Y. Li, X. Li, X. Li, Y. Hao, L. Wang, and Z. Li, "Dreamtexture: High-fidelity synthetic 3d data generation through decoupled geometry and texture synthesis," in *European Conference on Computer Vision*. Springer, 2024, pp. 303–320.

[46] Z. Min, Y. Luo, W. Yang, Y. Wang, and Y. Yang, "Entangled view-epipolar information aggregation for generalizable neural radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4906–4916.

[47] Z. Min, Y. Luo, J. Sun, and Y. Yang, "Epipolar-free 3d gaussian splatting for generalizable novel view synthesis," in *Annual Conference on Neural Information Processing Systems*, 2024, pp. 39 573–39 596.

[48] Y. Zeng, Y. Jiang, S. Zhu, Y. Lu, Y. Lin, H. Zhu, W. Hu, X. Cao, and Y. Yao, "Stag4d: Spatial-temporal anchored generative 4d gaussians," in *European Conference on Computer Vision*. Springer, 2025, pp. 163–179.

[49] Y. Zheng, X. Li, K. Nagano, S. Liu, O. Hilliges, and S. 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*, 2024, pp. 7300–7309.

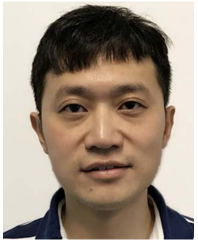
[50] J. Quan, Q. Miao, Y. Xu, Z. Lin, Y. Li, W. Yang, Z. Li, and Y. Luo, "Particlegs: Learning neural gaussian particle dynamics from videos

for prior-free physical motion extrapolation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026, pp. 8331–8341.
[51] Q. Miao, J. Quan, K. Li, Y. Xu, Y. Yang, and Y. Luo, "Frequency-aware dynamic gaussian splatting," in *The Fourteenth International Conference on Learning Representations*, 2026.
[52] Y. Xu, Q. Miao, J. Quan, W. Yang, Z. Li, and Y. Luo, "Langfield4d: Learning identity-adaptive and spatio-temporal continuous 4d language fields for dynamic scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026, pp. 9558–9569.
[53] R. Liu, R. Wu, B. Van Hoorick, P. Tokmakov, S. Zakharov, and C. Vondrick, "Zero-1-to-3: Zero-shot one image to 3d object," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 9298–9309.
[54] Z. Wang, C. Lu, Y. Wang, F. Bao, C. Li, H. Su, and J. Zhu, "Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation," *Annual Conference on Neural Information Processing Systems*, vol. 36, 2024.
[55] D. Kingma, T. Salimans, B. Poole, and J. Ho, "Variational diffusion models," *Annual Conference on Neural Information Processing Systems*, vol. 34, pp. 21 696–21 707, 2021.
[56] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=PxTIG12RRHS
[57] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "SDXL: Improving latent diffusion models for high-resolution image synthesis," in *International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=di52zR8xgf
[58] C. Holtz, "imagetoprompt," Feb. 2025. [Online]. Available: https://imagetoprompt.com/
[59] Kling, "Kling AI: Next-generation ai creative studio," https://klingai.com/, 2025, accessed: 2025-04-17.
[60] F. A. Fardo, V. H. Conforto, F. C. de Oliveira, and P. S. Rodrigues, "A formal evaluation of psnr as quality measurement parameter for image segmentation algorithms," *arXiv preprint arXiv:1605.07116*, 2016.
[61] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Asilomar Conference on Signals, Systems & Computers*, vol. 2. Ieee, 2003, pp. 1398–1402.
[62] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 586–595.
[63] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021. [Online]. Available: https://arxiv.org/abs/2103.00020
[64] T. Unterthiner, S. van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, "FVD: A new metric for video generation," 2019. [Online]. Available: https://openreview.net/forum?id=rylgEULtdN
[65] J. Bai, M. Xia, X. Wang, Z. Yuan, Z. Liu, H. Hu, P. Wan, and D. ZHANG, "Syncammaster: Synchronizing multi-camera video generation from diverse viewpoints," in *International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=m8Rk3HLGFx
[66] Z. Huang, Y. He, J. Yu, F. Zhang, C. Si, Y. Jiang, Y. Zhang, T. Wu, Q. Jin, N. Chanpaisit, Y. Wang, X. Chen, L. Wang, D. Lin, Y. Qiao, and Z. Liu, "VBench: Comprehensive benchmark suite for video generative models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.
[67] T. Liu, Z. Huang, Z. Chen, G. Wang, S. Hu, L. Shen, H. Sun, Z. Cao, W. Li, and Z. Liu, "Free4d: Tuning-free 4d scene generation with spatial-temporal consistency," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, October 2025, pp. 25 571–25 582.
[68] H. Zhang, X. Chen, Y. Wang, X. Liu, Y. Wang, and Y. Qiao, "4diffusion: Multi-view video diffusion model for 4d generation," in *Annual Conference on Neural Information Processing Systems*, 2024. [Online]. Available: https://openreview.net/forum?id=SFk7AMpyhx
[69] Z. Wang, Y. Wang, Y. Chen, C. Xiang, S. Chen, D. Yu, C. Li, H. Su, and J. Zhu, "Crm: Single image to 3d textured mesh with convolutional reconstruction model," in *European Conference on Computer Vision*. Springer, 2024, pp. 57–74.
[70] Y. Yin, D. Xu, Z. Wang, Y. Zhao, and Y. Wei, "4dgen: Grounded 4d content generation with spatial-temporal consistency," *arXiv preprint arXiv:2312.17225*, 2023.
[71] J. Zhang, C. Herrmann, J. Hur, V. Jampani, T. Darrell, F. Cole, D. Sun, and M.-H. Yang, "MonST3r: A simple approach for estimating geometry in the presence of motion," in *International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=lJpqxFgWCM
[72] P. Wang and Y. Shi, "Imagedream: Image-prompt multi-view diffusion for 3d generation," 2023. [Online]. Available: https://arxiv.org/abs/2312.02201
[73] D. Gatis, "rembg: Tool for removing image background," https://github.com/danielgatis/rembg, Feb. 2021.
[74] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U2-net: Going deeper with nested u-structure for salient object detection," *Pattern Recognition*, vol. 106, p. 107404, Oct. 2020. [Online]. Available: http://dx.doi.org/10.1016/j.patcog.2020.107404
[75] Tripo3D AI, "Tripo Studio 3D AI: Image to 3d model ai tool," https://studio.tripo3d.ai/, 2025, accessed: 2025-04-17.

Qiaowei Miao received a bachelor's degree in computer science and technology from Hebei University, China, in 2021. He is working toward a PhD at the School of Software Technology, Zhejiang University, China. His research interests include 4D vision and deep learning.

Kehan Li received a bachelor's degree in computer science and technology at Chongqing University, China, in 2024. He is pursuing a master's degree at Zhejiang University, focusing on Artificial Intelligence and Computer Vision.

Yawei Luo received the PhD degree from the Huazhong University of Science and Technology, in 2020. He is a ZJU 100 young professor with the School of Software Technology, Zhejiang University. He was a postdoctoral researcher with CCAI, College of Computer Science and Technology in Zhejiang University from 2020 to 2023. He was a visiting Ph.D student with ReLER lab, AAII, University of Technology Sydney, from 2017 to 2019. His research interests include knowledge engineering, domain adaptation, and 3D reconstruction.



Yi Yang received the PhD degree in computer science from Zhejiang University, Hangzhou, China, in 2010. He is currently a distinguished professor with Zhejiang University, China. He was a professor and director with the ReLER Lab, Australian Artificial Intelligence Institute (AAII), University of Technology Sydney, Australia. He was a postdoctoral research with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA. His current research interest include machine learning and its applications to multimedia content analysis and computer vision, such as multimedia indexing and retrieval, surveillance video analysis and video semantics understanding.