Compositional 3D-aware Video Generation with LLM Director

Hanxin Zhu[1]*, **Tianyu He**[2], **Anni Tang**[3], **Junliang Guo**[2], **Zhibo Chen**[1], **Jiang Bian**[2]
[1]University of Science and Technology of China
[2]Microsoft Research Asia
[3]Shanghai Jiao Tong University
hanxinzhu@mail.ustc.edu.cn, tianyuhe@microsoft.com, memory97@sjtu.edu.cn,
junliangguo@microsoft.com, chenzhibo@ustc.edu.cn, jiang.bian@microsoft.com

Abstract

Significant progress has been made in text-to-video generation through the use of powerful generative models and large-scale internet data. However, substantial challenges remain in precisely controlling individual concepts within the generated video, such as the motion and appearance of specific characters and the movement of viewpoints. In this work, we propose a novel paradigm that generates each concept in 3D representation separately and then composes them with priors from Large Language Models (LLM) and 2D diffusion models. Specifically, given an input textual prompt, our scheme consists of three stages: 1) We leverage LLM as the director to first decompose the complex query into several sub-prompts that indicate individual concepts within the video (*e.g.*, scene, objects, motions), then we let LLM to invoke pre-trained expert models to obtain corresponding 3D representations of concepts. 2) To compose these representations, we prompt multi-modal LLM to produce coarse guidance on the scales and coordinates of trajectories for the objects. 3) To make the generated frames adhere to natural image distribution, we further leverage 2D diffusion priors and use Score Distillation Sampling to refine the composition. Extensive experiments demonstrate that our method can generate high-fidelity videos from text with diverse motion and flexible control over each concept. Project page: `https://aka.ms/c3v`.

1 Introduction

Benefitting from large-scale data and the advancement of the generative models [1, 2], we have witnessed plenty of astonishing results across a wide array of tasks. For example, Large Language Models (LLM) pre-trained on web-scale datasets are revolutionizing machine learning with strong capability of zero-shot learning [3] and planning [4, 5], while diffusion models [6] empower text-to-image generation with a rapid surge in both quality and diversity [7–9].

To harness the power of text-to-image models in text-to-video generation, modern solutions directly view video as multiple images. In this way, tremendous efforts have been dedicated to extending text-to-image models with temporal interaction to ensure consistency between frames [10–17]. However, generating visual content conditioned on the textual prompt alone struggles to express multiple concepts with precise spatial layout control [18–20]. To tackle this issue, LVD [21] and VideoDirectorGPT [22] propose to first generate spatiotemporal bounding boxes of each object based on the textual prompt with LLM, and then condition the video generation on the obtained layouts. Although rough layout control can be realized, they still have inherent limitations for detailed concept control, *e.g.*, the motion and appearance of specific characters, and the movement of viewpoints.

*This work is accomplished in Microsoft, April 2024.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-4184

In nature, our understanding of the world is compositional [23, 24, 20], and the interaction with the world takes place in a 3D. Motivated by this, in contrast to the prior endeavors that *implicitly* learn different concepts in 2D space, we are interested in exploring an alternative solution that *explicitly* composes concepts in 3D space for video generation. To this end, we in particular identify two key technical challenges: 1) Since a textual prompt contains multiple concepts, how to coordinate the generation of various concepts? 2) Given the generated concepts, how to compose them to follow common sense in the real world?

In this work, we introduce text-guided compositional 3D-aware video generation (C3V), a novel paradigm that regards LLM as director and 3D as structural representation for video generation. C3V consists of three main stages: **1)** Given a textual prompt, to coordinate the generation of various concepts, we leverage LLM to disassemble the input prompt into sub-prompts, where each sub-prompt describes an individual concept, *e.g.*, the scene, objects, and motion. For each concept, a pre-trained expert model is assigned by LLM to generate its corresponding 3D representation (*e.g.*, 3D Gaussians [25], SMPL parameters [26]) according to the textual description. **2)** To provide coarse instruction for composition (*i.e.*, the scale and trajectory of each object in the scene), we further resort to the priors in multi-modal LLM by querying it with the rendered scene image and the textual goals. However, directly instructing multi-modal LLM to return the scale and trajectory of each object leads to unexpected results, as it is challenging for LLM to estimate visual dynamics. Therefore, we follow a step-by-step reasoning philosophy [27] by representing the object with the bounding boxes and dividing the trajectory estimation into sub-tasks, *i.e.*, estimating the starting points, ending points, and trajectories step-by-step. **3)** After obtaining the coarse trajectories from the language space, we also propose to refine the scales, rotations, and exact locations with priors from large-scale visual data. Specifically, taking inspiration from DreamFusion [28], which proposes to distill generative priors from pre-trained image diffusion models into 3D objects, we employ Score Distillation Sampling (SDS) [28] to optimize the transformation matrix of each object in 3D space.

Our system has three main advantages: 1) Because each concept is represented by individual 3D representations, it naturally supports flexible control and interaction of each concept. 2) It inherently excels at synthesizing complex and long videos such as drama, etc. 3) The viewpoint is controllable.

Extensive experiments demonstrate that our proposed method can generate 3D-aware videos with diverse motion and high visual quality, even from complex queries that contain multiple concepts and relationships. We also illustrate the flexibility of C3V by editing various concepts of the generated videos. The generated videos are presented on our project page. To the best of our knowledge, we make the first attempt towards text-guided compositional 3D-aware video generation. We hope it can inspire further explorations on the interplay between video and 3D generation.

2 Related Works

2.1 Video Generation with LLM

Recently, there have been substantial efforts in training text-to-video models on large-scale datasets with autoregressive Transformer [29, 30, 17] or diffusion models [10–13, 16]. A prominent approach for text-to-video generation is to extend a pre-trained text-to-image model by inserting temporal layers into its architecture, and fine-tuning models on video data. However, although effective, it remains challenging to generate objects with specific attributes or positions. To address this challenge, a series of studies proposed to exploit knowledge from LLM [31, 32] to achieve controllable generation [21, 19, 22, 33–35], zero-shot generation [36–39], or long video generation [40]. For example, Free-Bloom [36] and DirecT2V [38] used LLM to transform the input textual prompt into a sequence of sub-prompts that describe each frame. LVD [21] and VideoDirectorGPT [22] employed LLM to generate spatiotemporal bounding boxes to control the object-level dynamics in video generation.

In light of the above success of exploiting LLM to direct video generation in 2D space, we view LLM as a director in 3D, which differs from previous methods not only in terms of technical route but also in benefits: providing free interaction with individual concepts and flexible viewpoint control.

2.2 Compositional 3D Generation

Generating 3D assets from textual prompt has garnered significant attention owing to its promising applications in various fields such as AR [41], VR [42], and autonomous driving [43]. However, due to the lack of large-scale 3D data, it is challenging to apply 2D generative models to 3D directly. Therefore, building upon Dream Fields [44], DreamFusion introduced the Score Distillation Sampling (SDS) [28], a technique enhancing 3D generation by distilling 2D diffusion priors from pre-trained text-to-image generative models. Motivated by the success of DreamFusion [28], dedicated efforts have been made to improve SDS [45–47]. Though achieving remarkable results, these methods struggle to generate scenes with multiple distinct elements. To mitigate this issue, several techniques was proposed to guide 3D generation with additional conditions like layout priors, which we refer to as compositional 3D generation [48–50]. However, these works still focus on static compositional 3D generation and lack visual dynamic modeling.

Recently, two concurrent works Comp4D [51] and TC4D [52] also achieved compositional 4D generation (*i.e.*, dynamic 3D generation). However, they only considered composition between objects, and the trajectory of these methods is either formulated by kinematics-based equations or pre-defined by users. Differently, we explore 3D-aware video generation with integrated 3D scenes and compose various concepts with priors from both LLM and 2D diffusion models.

3 Preliminaries

3.1 3D Gaussian Splatting

3D Gaussian Splatting (3DGS) [25] has been attracting a lot of interest for novel view synthesis, due to its photorealistic visual quality and real-time rendering. 3DGS utilizes a set of anisotropic ellipsoids (*i.e.*, 3D Gaussians) to encode 3D properties, in which each Gaussian is parameterized by position $\mu \in \mathbb{R}^3$, covariance $\mathbf{\Sigma} \in \mathbb{R}^{3\times3}$ (obtained from scale $\mathbf{s} \in \mathbb{R}^3$ and rotation $\mathbf{r} \in \mathbb{R}^3$), opacity $\alpha \in \mathbb{R}$, and color $\mathbf{c} \in \mathbb{R}^3$.

To render a novel view, 3DGS adopts a tile-based rasterization, where 3D Gaussians are projected onto the image plane as 2D Gaussians. The final color $\mathbf{c}(\mathbf{p})$ of pixel $\mathbf{p}$ is denoted as:

$$\mathbf{c}(\mathbf{p}) = \sum \hat{\mathbf{c}}\hat{\sigma} \prod (1 - \hat{\sigma}), \tag{1}$$

where $\hat{\mathbf{c}}$ and $\hat{\sigma}$ represent the individual color and opacity values of a series of 2D Gaussians contributing to this pixel. 3DGS are then optimized using L1 loss and SSIM [53] loss in a per-view optimization manner. Thanks to the nature of modeling 3D scenes explicitly, optimized 3D Gaussians can be easily controlled and edited.

3.2 Score Distillation Sampling

Different from text-to-image generation which benefits from a large number of text-image pairs available, text-to-3D generation suffers from a severe lack of data. To mitigate this issue, Score Distillation Sampling (SDS) [28] was proposed to distill generative priors from pretrained diffusion-based text-to-image models ϕ. Specifically, for a 3D representation parameterized by θ, SDS is served as a way to measure the similarity between the rendered images $x = g(\theta)$ and the given textual prompts y, where g represents the rendering operation. As a result, the gradients used to update θ are computed as follows:

$$\nabla_\theta \mathcal{L}_{SDS}(\phi, x = g(\theta)) = \mathbb{E}_{t,\epsilon}[w(t)(\hat{\epsilon}_\phi(x_t; y, t) - \epsilon)], \tag{2}$$

where t is the noise level, ϵ is the ground-truth noise, $w(t)$ is a weighting function, $\hat{\epsilon}_\phi$ is the estimated noise given noised images x_t with text embeddings y. Please refer to DreamFusion [28] for details.

4 Method

Overview. To achieve text-guided compositional 3D-aware video generation (C3V), we regard LLM as director and 3D as structural representation. To this end, our method consists of three stages. To begin with, we utilize LLMs to decompose the input textual prompts into three sub-prompts, each

 https://doi.org/10.52202/079017-4184

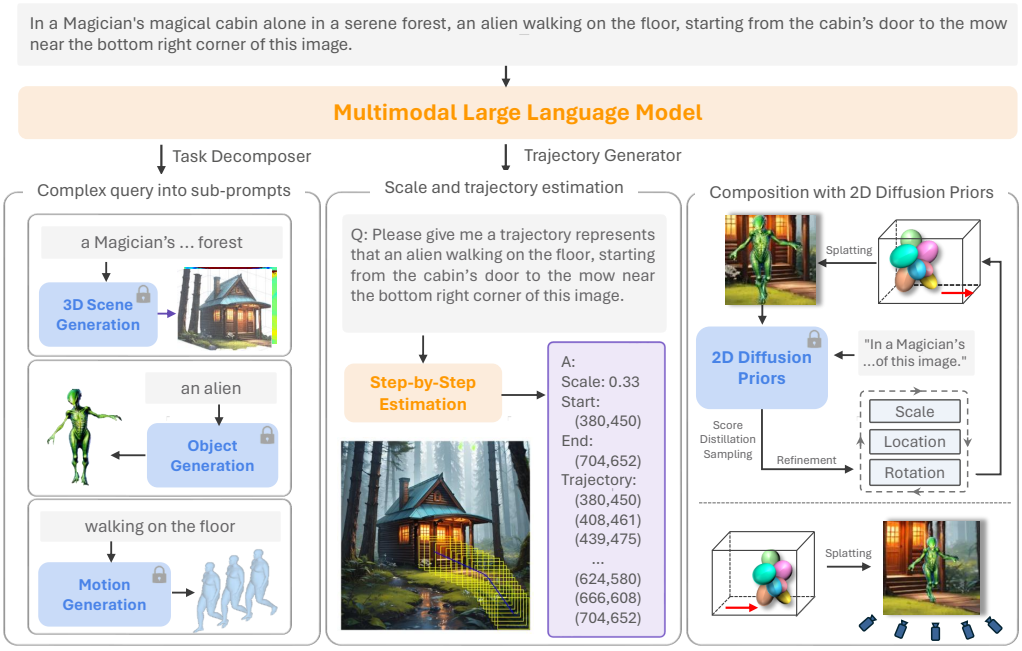


Figure 1: Illustration of our method. It consists of three stages: 1) The input textual prompt is decomposed into individual concepts by the LLM. Then we generate each concept in the form of 3D with the corresponding pre-trained expert model (*left* & Sec. 4.1). 2) We leverage knowledge in multi-modal LLM to estimate the 2D trajectory of objects step-by-step (*middle* & Sec. 4.2). 3) After lifting the estimated 2D trajectory into 3D as initialization, we refine the scales, locations, and rotations of objects within the 3D scene using 2D diffusion priors (*right* & Sec. 4.3).

of which provides a description for generating a corresponding concept (*i.e.*, scene, object, motion, etc.) respectively (Sec. 4.1). Subsequently, we leverage multi-modal LLM to obtain coarse-grained scales and trajectories for each animatable object (Sec. 4.2). Finally, we employ 2D diffusion priors to refine the objects' location, scale, and rotation for a fine-grained composition (Sec. 4.3).

4.1 Task Decomposition with LLM

Task Instructions. Given a textual prompt, we invoke LLM (*e.g.*, GPT-4V [32]) to decompose it into several sub-prompts. Each sub-prompt describes an individual concept such as the scene, object, and motion. Specifically, for an input prompt y, we query LLM with the instruction like: "*Please decompose this prompt into several sub-prompts, each describing the scene, objects in the scene, and the objects' motion.*", from which we obtain the corresponding sub-prompts.

3D Representation. After obtaining the sub-prompt for each concept, we aim to generate its corresponding 3D representations using the pre-trained expert models. In this work, we build structural representation on 3DGS [25], which is an explicit form and therefore flexible enough to compose or animate. Concerning concepts like motion, our framework can generalize to arbitrary animatable 3D Gaussian-based objects. For simplicity, we take human motion as an instantiation because it is general for various scenarios. In order to obtain diverse human motions, we take a retrieval-augmented approach [54] to acquire motion in the form of SMPL-X parameters [55] from large motion libraries [56] according to the motion-related sub-prompt.

Instantiation. To illustrate the scheme formally, consider the following example. We have sub-prompts y_1, y_2 and y_3 that describe scene, object, and motion respectively. Additionally, we have corresponding pre-trained text-guided expert models ϕ_1, ϕ_2, and ϕ_3 that are selected by the LLM. The concept generation can be formulated as follows:

$$G_1 = \phi_1(y_1),\ G_2 = \phi_2(y_2, M),\ M = \phi_3(y_3), \tag{3}$$

https://doi.org/10.52202/079017-4184

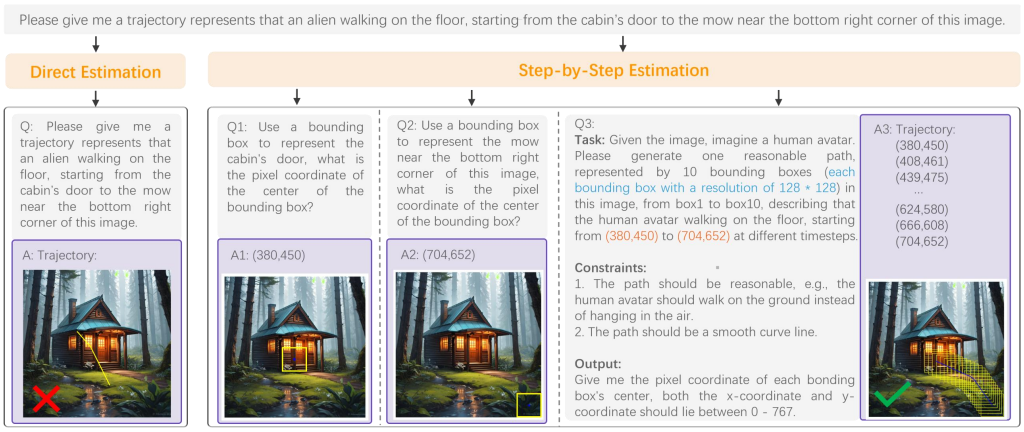


Figure 2: Illustration of coarse-grained trajectory generation with LLM. Instead of querying multi-modal LLM to estimate dynamic trajectory directly, we generate trajectory in a step-by-step manner: estimating the locations of starting and ending points first, then reasoning the path between them.

where G_1 and G_2 represent the 3D Gaussians, and M means the motions used to drive G_2. In the following sections, we will provide details on the composition of the generated concepts.

4.2 Coarse-grained Trajectory Generation with LLM

Given the generated concepts, we aim to compose them into a dynamic 3D representation to render videos that align with the input textual prompt. Achieving this requires scales and trajectories of the objects to indicate their sizes and locations within the scene. To this end, we propose to leverage knowledge encoded in multi-modal LLMs (*i.e.*, GPT-4V [32]) to provide priors.

For the scale of the object, we find that directly querying GPT-4V with the input prompt and rendered scene image can yield a reasonable estimation of its resolution (H_{2D} and W_{2D}). However, this is not the case for trajectory estimation. As demonstrated in Fig. 2, directly querying GPT-4V for trajectory will lead to a result that deviates conspicuously from common sense. Based on this observation, we conclude two issues: 1) it is too difficult for GPT-4V to generate the trajectory within a single query, as it lacks priors on visual dynamics; 2) since GPT-4V is trained to generate text, it has limitations on imagining visual content.

To mitigate this, we introduce two simple yet effective techniques. 1) Although GPT-4V lacks visual knowledge of the object, we can alleviate this by representing the object as a bounding box with the estimated resolution. 2) We follow a step-by-step reasoning philosophy [27] and propose to let GPT-4V estimate the locations of starting and ending points first, then reason the path between them.

Overall, we can formulate the above process as follows:

$$\begin{aligned} \{L_p^i\}_{i=1}^N &= \Phi(y_p, I, S, L_s, L_e), \\ S = \Phi(y, I), L_s &= \Phi(y_s, I), L_e = \Phi(y_e, I), \end{aligned} \tag{4}$$

where Φ represents the multi-modal LLM (*i.e.*, GPT-4V), I denotes the rendered scene image, S represents the estimated scale of the object given textual prompt y and I, L_s and L_e represent the locations of starting and ending points respectively, $\{L_p^i\}_{i=1}^N$ represent N locations indicating the path between L_s and L_e. Notably, all locations (*i.e.*, L_s, L_e,$\{L_p^i\}_{i=1}^N$) are represented by 2D pixel coordinates on I.

4.3 Fine-grained Composition with 2D Diffusion Priors

Lift Trajectory from 2D to 3D. In Sec. 4.2, we obtain the 2D pixel coordinates $L_p^i = (p_x^i, p_y^i)$ of the estimated trajectory. However, 2D trajectory is not enough for composition in 3D space. Therefore, we convert it into corresponding 3D world coordinate $L_{3D}^i = (x^i, y^i, z^i)$. Specifically, we first predict the depth map D of the rendered scene image with a monocular depth estimator [57].

https://doi.org/10.52202/079017-4184

Then, we use the depth value of the center point of the lower boundary of the bounding box as the trajectory's depth. As a result, we can transform 2D trajectory into 3D:

$$(x^i, y^i, z^i, 1)^T = R^{-1}K^{-1}[(p_x^i + \frac{H_{2D}}{2}, p_y^i, 1)^T \cdot D(p_x^i + \frac{H_{2D}}{2}, p_y^i)] - (\frac{H_{3D}}{2}, 0, 0, 0)^T, \quad (5)$$

where R and K represent camera extrinsic and intrinsic respectively, H_{2D} and W_{2D} represent the resolution of the 2D bounding box. H_{3D} represent the actual height of the 3D bounding box of this object within the scene.

Composition Refinement with 2D Diffusion Priors. With the lifted 3D trajectory, we then integrate the object into the scene. However, the trajectory estimated by LLM is still rough and may not obey natural image distribution. To address this, we propose to further refine the object's scale, location, and rotation by distilling generative priors from pre-trained image diffusion models [7] into 3D space. Specifically, we treat the parameters for these attributes as optimizable variables and use SDS (Eq. 2) to improve the fidelity of rendered images. As a result, scale refinement can be formulated as follows:

$$\nabla_{\hat{S}}\mathcal{L}_{SDS}^{Scale} = \mathbb{E}_{t,\epsilon}[w(t)(\hat{\epsilon}_\phi(x_t(L_{3D}^1, (S + \sigma(\hat{S}) \cdot \tau_s - \frac{\tau_s}{2}) \cdot G_2); y, t) - \epsilon)], \quad (6)$$

where $\hat{S}$ represents the optimizable variable, S represents the scale estimated by GPT-4(V), σ means the Sigmoid function, τ_s is a threshold, G_2 represents the 3D gaussians of the object, and x_t is the noised 2D image given L_{3D}^i and scaled G_2.

After obtaining a more precise scale, we then refine the locations of the estimated 3D trajectory similarly, where the location refinement is denoted as:

$$\nabla_{\hat{L^i}}\mathcal{L}_{SDS}^{Location} = \mathbb{E}_{t,\epsilon}[w(t)(\hat{\epsilon}_\phi(x_t(L_{3D}^i + \sigma(\hat{L^i}) \cdot \tau_L - \frac{\tau_L}{2}, (S + \sigma(\hat{S}) \cdot \tau_s - \frac{\tau_s}{2}) \cdot G_2); y, t) - \epsilon)], \quad (7)$$

where $\hat{L^i}$ represents the optimizable variable, τ_L is a threshold.

For the rotation of the object at different timesteps, we can directly compute the corresponding rotation matrix, based on the assumption that the object at the current time step should face the location of the object at the next time step. As a result, the rotation matrix $\hat{R^i}$ at time step i can be computed using the following equation:

$$\begin{aligned} \hat{R^i} &= \begin{bmatrix} tx^2 + c & txy - zs & txz + ys \\ txy + zs & ty^2 + c & tyz - xs \\ txz - ys & tyz + xs & tz^2 + c \end{bmatrix}, \\ t &= 1 - c, c = \cos(\theta), s = \sin(\theta), \mathbf{u} = (x, y, z)^T \end{aligned} \quad (8)$$

where θ and $\mathbf{u}$ represent the rotation angle and axis obtained through the cross product of $(L_{3D}^{i+1} + \sigma(\hat{L^{i+1}}) \cdot \tau_L - L_{3D}^i - \sigma(\hat{L^i}) \cdot \tau_L)$ and $(0, 0, 1)^T$.

Inference. After obtaining individual concepts in the form of 3D and the optimized parameters that indicate how to compose various concepts, we can render the 3D representation into 2D video with flexible camera control in real time [25].

5 Experiments

In this section, we instantiate C3V with three concepts: scene, humanoid object, and human motion, to generate 3D-aware video from text. We compare our proposed method with state-of-the-art text-to-4D models (4D-FY [58]), compositional 4D generation models (Comp4D [51]) and text-to-video models (VideoCrafter2 [59]). Videos are available on our anonymous project page.

Implementation Details. We use LucidDreamer [60], HumanGaussian [61] and Motion-X [56] to generate 3D scenes, humanoid objects and motions respectively. To realize SDS, we utilize Stable Diffusion [7] as the image diffusion model. All the videos of our proposed method are rendered at a resolution of 512×512 in real time. Please refer to the appendix for more details.

(a) Text prompt: "*In a Magician's magical cabin alone in a serene forest, an alien walking on the floor, starting from the cabin's door to the mow near the bottom right corner of this image*".

(b) Text prompt: "*Four characters stood on the stage. In front of the stage, a man and a woman are performing Kung Fu and dancing respectively. On the right side of the stage, a skeleton man is dancing, and behind them, a clown is performing*".

Figure 3: Qualitative comparisons with baselines. When prompting complex queries, the baseline methods fail to follow the queries in terms of the number of objects and the corresponding motion. In contrast, our method excels in yielding both diverse motion and high visual quality.

Metrics. Following Comp4D [51], we choose Q-Align [62] as the referee to measure the quality and aesthetics of the video. The Q-Align score is a number ranging from 1 (worst) to 5 (best) where a higher score indicates a better performance. We also report the CLIP score [63] to measure the alignment between the generated videos and the input texts.

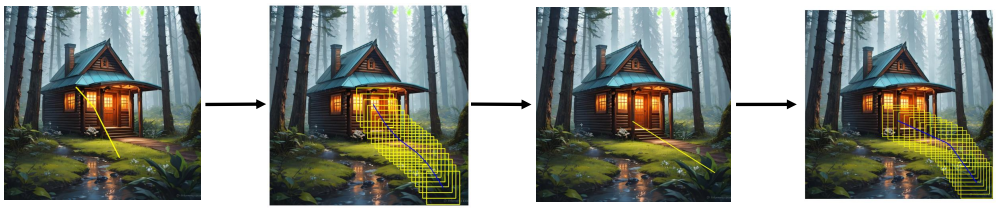
5.1 Comparison with Competitors

In Fig. 3, we conduct a comparative analysis of our method against 4D-FY [58], Comp4D [51], and VideoCrafter2 [59] with the same textual prompt. It can be observed that all three baselines fail to provide diverse motion from the textual prompt, while our method excels in yielding large motion and high visual quality. For example, our scheme successfully obeys the complex query in terms of

https://doi.org/10.52202/079017-4184

Table 1: Quantative comparisons with competitors. Our method consistently outperforms all baseline methods in terms of both the video quality and the alignment with textual prompts.

Metric	4D-FY [58]	Comp4D [51]	VideoCrafter2 [59]	Ours
QAlign-img-quality ↑ [62]	1.681	1.687	3.839	**4.030**
QAlign-img-aesthetic↑ [62]	1.475	1.258	3.199	**3.471**
QAlign-vid-quality↑ [62]	2.154	2.142	3.868	**4.112**
QAlign-vid-aesthetic↑ [62]	1.580	1.425	3.159	**3.723**
CLIP Score↑ [63]	30.47	27.50	35.20	**38.36**



(I) Direct trajectory estimation. (II) Trajectory estimation with bounding box indicating objects. (III) Trajectory estimation in a step-by-step manner. (IV) Ours.

(a) Ablation studies on trajectory estimation with multi-modal LLM.

(I) Without SDS refinement. (II) With scale refinement. (III) With trajectory refinement. (IV) With rotation refinement.

(b) Ablation studies on composition with 2D diffusion models.

Figure 4: Ablation studies on framework design. Each ablation is prompted with the same text.

the number of objects and the corresponding motion. In addition, since 4D-FY and Comp4D focus on object-centric generation, they fail to generate videos with natural backgrounds. In Tab. 3, we perform quantitative comparisons by utilizing Q-Align Score [62] and CLIP Score [63] to assess the quality of generated videos. Our method consistently outperforms the baseline models in terms of both the video quality and the alignment with textual prompts. More results are available in the appendix.

5.2 Ablation Studies

Ablations on Trajectory Estimation with Multi-modal LLM. As shown in Fig. 4(a)(I), a direct prompt of GPT-4V will lead to obvious unsatisfactory trajectory estimation. When only depending on bounding boxes to indicate the location of objects within the scene (Fig. 4(a)(II)), though a roughly better trajectory can be achieved, it still leads to unreasonable results, such as several floating bounding boxes. Similarly, using only the step-by-step estimation strategy described in Sec. 4.2 typically results in a trajectory that is merely a simple straight line connecting the starting and ending points (Fig. 4(a)(III)). With both of the two techniques, we can achieve the best performance, with a more reasonable and smooth trajectory (Fig. 4(a)(IV)).

Ablations on Composition with 2D Diffusion Models. To figure out whether it is necessary to conduct fine-grained composition with 2D generative priors, we gradually refine the scales, locations, and rotations with SDS and visualize the results in Fig. 4(b). All results are generated with the same textual prompt: "*An alien walking on the floor in front of the cabin's door.*". It shows that when we optimize the attributes with SDS, we can obtain consistently improved performance with a reasonable scale (Fig. 4(b)(II), accurate locations that are aligned with the input prompt (Fig. 4(b)(III), and orientation that accords with common sense (Fig. 4(b)(IV)).

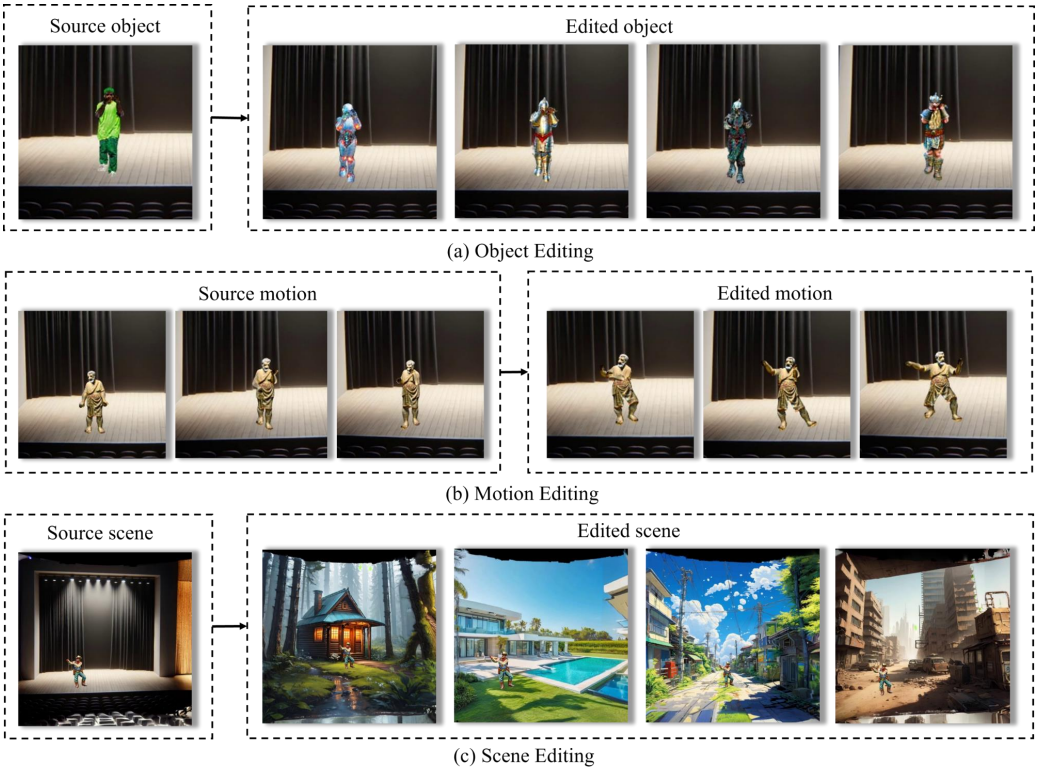


Figure 5: Our method offers flexible control of individual concepts. We demonstrate this by editing different concepts: the appearance and motion of the actors, and the scenes.

Table 2: Quantative comparisons of ablation studies on trajectory estimation with multi-modal LLM.

Methods	Direct Estimation.	Estimation using bounding box.	Step-by-step estimation.	Ours
QAlign-img-quality ↑ [62]	2.056	2.894	3.752	**4.030**
QAlign-img-aesthetic ↑ [62]	1.568	2.156	3.047	**3.471**
QAlign-vid-quality ↑ [62]	2.178	3.043	3.904	**4.112**
QAlign-vid-aesthetic ↑ [62]	1.680	2.346	3.342	**3.723**
CLIP Score ↑ [63]	25.68	29.84	36.73	**38.36**

5.3 Applications on Controllable Generation

Due to our underlying 3D structural representation, our scheme has the natural merits of editing individual concepts. We illustrate this character in Fig. 5 by editing three different concepts: the appearance and motion of the actors, and the scenes. For the appearance and motion of the actor, we can seamlessly replace them in a zero-shot manner according to the textual prompt (Fig.5(a)(b)), while this is still challenging for implicit models [64, 65]. For scene editing, to ensure a smooth composition of objects within the target scene, we re-estimate the trajectory of the objects given the target scene. Kindly refer to appendix for more results.

6 Conclusion

In this paper, we present a novel paradigm for 3D-aware video generation by conceptualizing videos as compositions of independent concepts represented in 3D space. To this end, we leverage LLM as director to decompose the input textual prompts into individual concepts and then invoke pre-trained expert models to generate them separately. To compose various concepts, we first prompt multi-modal LLM in a step-by-step manner to provide coarse guidance on the scale and trajectory of objects,

https://doi.org/10.52202/079017-4184

Table 3: Quantative comparisons of ablation studies on composition with 2D diffusion models.

Methods	Without SDS.	With scale refinement.	With trajectory refinement.	Ours
QAlign-img-quality ↑ [62]	3.045	3.674	3.826	**4.030**
QAlign-img-aesthetic ↑ [62]	2.752	3.046	3.341	**3.471**
QAlign-vid-quality ↑ [62]	3.129	3.794	3.983	**4.112**
QAlign-vid-aesthetic ↑ [62]	2.704	3.468	3.603	**3.723**
CLIP Score ↑ [63]	31.35	35.27	37.04	**38.36**

then refine the composition with 2D generative priors. We verify our scheme in different scenarios, demonstrating its superiority over the baseline methods.

Limitations and Future Works. Although we demonstrate promising results in 3D-aware video generation, there still are limitations to be improved in the future. First, our framework is instantiated with limited concepts in this work, *i.e.*, scene, humanoid object, and human motion. It is exciting to generalize the framework to more concepts like animals, vehicles, etc. Second, the composition between concepts is conducted with priors from LLM and 2D diffusion priors in our method. However, it is still interesting to introduce physically grounded dynamics into 3D representation [66]. Third, though our method is naturally suitable for maintaining the consistency of actors across different scenes, it still needs further exploration on long video generation with multiple scenes, *e.g.*, a full-length film.

Ethics Statement. C3V is exclusively a research initiative with no current plans for product integration or public access. We are committed to adhering to Microsoft AI principles during the ongoing development of our models. The model is trained on AI-generated content, which has been thoroughly reviewed to ensure that they do not include personally identifiable information or offensive content. Nonetheless, as these generated data are sourced from the Internet, there may still be inherent biases. To address this, we have implemented a rigorous filtering process on the data to minimize the potential for the model to generate inappropriate content.

Acknowledgement. This work was supported in part by NSFC under Grant 62371434, 62021001.

References

[1] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:6840–6851, 2020.

[3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:1877–1901, 2020.

[4] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36, 2024.

[5] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. In *Conference on Robot Learning*, pages 540–562. PMLR, 2023.

[6] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems (NeurIPS)*, 34:8780–8794, 2021.

[7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.

[8] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[9] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:36479–36494, 2022.

[10] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv:2204.03458*, 2022.

[11] 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.

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

[13] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *arXiv preprint arXiv:2309.15818*, 2023.

[14] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, et al. Videocrafter1: Open diffusion models for high-quality video generation. *arXiv preprint arXiv:2310.19512*, 2023.

[15] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning. *arXiv preprint arXiv:2311.10709*, 2023.

[16] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Li Fei-Fei, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models. *arXiv preprint arXiv:2312.06662*, 2023.

[17] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Rachel Hornung, Hartwig Adam, Hassan Akbari, Yair Alon, Vighnesh Birodkar, et al. Videopoet: A large language model for zero-shot video generation. *arXiv preprint arXiv:2312.14125*, 2023.

[18] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.

[19] Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. Layoutgpt: Compositional visual planning and generation with large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

[20] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision*, pages 423–439. Springer, 2022.

[21] Long Lian, Baifeng Shi, Adam Yala, Trevor Darrell, and Boyi Li. Llm-grounded video diffusion models. In *International Conference on Learning Representations*, 2024.

[22] Han Lin, Abhay Zala, Jaemin Cho, and Mohit Bansal. Videodirectorgpt: Consistent multi-scene video generation via llm-guided planning. *arXiv preprint arXiv:2309.15091*, 2023.

[23] Noam Chomsky. *Aspects of the Theory of Syntax*. Number 11. MIT press, 2014.

[24] Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

 https://doi.org/10.52202/079017-4184

[25] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023.

[26] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: a skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):1–16, 2015.

[27] Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024.

[28] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations*, 2023.

[29] 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 descriptions. In *International Conference on Learning Representations*, 2023.

[30] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. In *International Conference on Learning Representations*, 2023.

[31] OpenAI. Chatgpt. *https://openai.com/chatgpt*, 2022.

[32] OpenAI. Gpt-4v(ision) system card. *https://openai.com/index/gpt-4v-system-card*, 2023.

[33] Sitong Su, Litao Guo, Lianli Gao, Hengtao Shen, and Jingkuan Song. Motionzero: Exploiting motion priors for zero-shot text-to-video generation. *arXiv preprint arXiv:2311.16635*, 2023.

[34] Yash Jain, Anshul Nasery, Vibhav Vineet, and Harkirat Behl. Peekaboo: Interactive video generation via masked-diffusion. *arXiv preprint arXiv:2312.07509*, 2023.

[35] Sixiao Zheng, Jingyang Huo, Yu Wang, and Yanwei Fu. Intelligent director: An automatic framework for dynamic visual composition using chatgpt. *arXiv preprint arXiv:2402.15746*, 2024.

[36] Hanzhuo Huang, Yufan Feng, Cheng Shi, Lan Xu, Jingyi Yu, and Sibei Yang. Free-bloom: Zero-shot text-to-video generator with llm director and ldm animator. *Advances in Neural Information Processing Systems*, 36, 2024.

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

[38] Susung Hong, Junyoung Seo, Heeseong Shin, Sunghwan Hong, and Seungryong Kim. Direct2v: Large language models are frame-level directors for zero-shot text-to-video generation. *arXiv preprint arXiv:2305.14330*, 2023.

[39] Gyeongrok Oh, Jaehwan Jeong, Sieun Kim, Wonmin Byeon, Jinkyu Kim, Sungwoong Kim, Hyeokmin Kwon, and Sangpil Kim. Mtvg: Multi-text video generation with text-to-video models. *arXiv preprint arXiv:2312.04086*, 2023.

[40] Shaobin Zhuang, Kunchang Li, Xinyuan Chen, Yaohui Wang, Ziwei Liu, Yu Qiao, and Yali Wang. Vlogger: Make your dream a vlog. *arXiv preprint arXiv:2401.09414*, 2024.

[41] Chenliang Chang, Kiseung Bang, Gordon Wetzstein, Byoungho Lee, and Liang Gao. Toward the next-generation vr/ar optics: a review of holographic near-eye displays from a human-centric perspective. *Optica*, 7(11):1563–1578, 2020.

[42] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023.

[43] Xiangyu Yue, Bichen Wu, Sanjit A Seshia, Kurt Keutzer, and Alberto L Sangiovanni-Vincentelli. A lidar point cloud generator: from a virtual world to autonomous driving. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, pages 458–464, 2018.

[44] Ajay Jain, Ben Mildenhall, Jonathan T Barron, Pieter Abbeel, and Ben Poole. Zero-shot text-guided object generation with dream fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 867–876, 2022.

[45] 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, 2024.

[46] Yukun Huang, Jianan Wang, Yukai Shi, Xianbiao Qi, Zheng-Jun Zha, and Lei Zhang. Dreamtime: An improved optimization strategy for text-to-3d content creation. *arXiv preprint arXiv:2306.12422*, 2023.

[47] Yixun Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, and Yingcong Chen. Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. *arXiv preprint arXiv:2311.11284*, 2023.

[48] Ryan Po and Gordon Wetzstein. Compositional 3d scene generation using locally conditioned diffusion. *arXiv preprint arXiv:2303.12218*, 2023.

[49] Gege Gao, Weiyang Liu, Anpei Chen, Andreas Geiger, and Bernhard Schölkopf. Graphdreamer: Compositional 3d scene synthesis from scene graphs. *arXiv preprint arXiv:2312.00093*, 2023.

[50] Xiaoyu Zhou, Xingjian Ran, Yajiao Xiong, Jinlin He, Zhiwei Lin, Yongtao Wang, Deqing Sun, and Ming-Hsuan Yang. Gala3d: Towards text-to-3d complex scene generation via layout-guided generative gaussian splatting. *arXiv preprint arXiv:2402.07207*, 2024.

[51] Dejia Xu, Hanwen Liang, Neel P Bhatt, Hezhen Hu, Hanxue Liang, Konstantinos N Plataniotis, and Zhangyang Wang. Comp4d: Llm-guided compositional 4d scene generation. *arXiv preprint arXiv:2403.16993*, 2024.

[52] Sherwin Bahmani, Xian Liu, Yifan Wang, Ivan Skorokhodov, Victor Rong, Ziwei Liu, Xihui Liu, Jeong Joon Park, Sergey Tulyakov, Gordon Wetzstein, et al. Tc4d: Trajectory-conditioned text-to-4d generation. *arXiv preprint arXiv:2403.17920*, 2024.

[53] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[54] Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*, 2020.

[55] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10975–10985, 2019.

[56] Jing Lin, Ailing Zeng, Shunlin Lu, Yuanhao Cai, Ruimao Zhang, Haoqian Wang, and Lei Zhang. Motion-x: A large-scale 3d expressive whole-body human motion dataset. *Advances in Neural Information Processing Systems*, 36, 2024.

[57] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021.

[58] 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. *arXiv preprint arXiv:2311.17984*, 2023.

 https://doi.org/10.52202/079017-4184

[59] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. *arXiv preprint arXiv:2401.09047*, 2024.

[60] 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.

[61] Xian Liu, Xiaohang Zhan, Jiaxiang Tang, Ying Shan, Gang Zeng, Dahua Lin, Xihui Liu, and Ziwei Liu. Humangaussian: Text-driven 3d human generation with gaussian splatting. *arXiv preprint arXiv:2311.17061*, 2023.

[62] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, et al. Q-align: Teaching lmms for visual scoring via discrete text-defined levels. *arXiv preprint arXiv:2312.17090*, 2023.

[63] 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*, pages 8748–8763. PMLR, 2021.

[64] Wilson Yan, Andrew Brown, Pieter Abbeel, Rohit Girdhar, and Samaneh Azadi. Motion-conditioned image animation for video editing. *arXiv preprint arXiv:2311.18827*, 2023.

[65] Jianhong Bai, Tianyu He, Yuchi Wang, Junliang Guo, Haoji Hu, Zuozhu Liu, and Jiang Bian. Uniedit: A unified tuning-free framework for video motion and appearance editing. *arXiv preprint arXiv:2402.13185*, 2024.

[66] Tianyi Xie, Zeshun Zong, Yuxin Qiu, Xuan Li, Yutao Feng, Yin Yang, and Chenfanfu Jiang. Physgaussian: Physics-integrated 3d gaussians for generative dynamics. *arXiv preprint arXiv:2311.12198*, 2023.

A Implementation Details

A.1 Experimental Settings

During the process of multi-modal LLMs-based trajectory estimation, we use 20 locations by default to indicate the trajectory between the starting point and the ending point, (*i.e.*, $N = 20$ in Eq. 4), where the path between adjacent locations is assumed as a straight line. For scale refinement (Eq. 6), τ_s is set to 0.1. For location refinement (Eq. 7), we apply it to refine all the twenty locations. The training iterations for each location is 1000 and τ_L is set to 0.1. All experiments are conducted using a single NVIDIA A100 GPU.

A.2 Pre-trained Expert Models

LucidDreamer [60]. As a powerful 3D scene generation method, LucidDreamer adopts an iterative view generation strategy, where a series of views are dreamed and aligned via depth-warping based inpainting networks. After obtaining these multiview-consistent images, 3D gaussians are optimized to construct a high-quality 3D scene, by means of typical training pipeline of 3DGS.

HumanGaussian [61]. We choose HumanGaussian as the method for human generation due to its capability of producing drivable avatars on the basis of 3DGS. Specifically, HumanGaussian starts with SMPL-X prior to densely sample Gaussians on the human mesh surface as initial center positions, followed by a texture-structure joint model and an annealed negative prompt guidance strategy to obtain high-fidelity outputs.

Motion-X [56]. Motion-X is a large-scale 3D expressive whole-body human motion dataset which comprises 15.6M precise 3D whole-body pose annotations (*i.e.*, SMPL-X) covering 81.1K motion sequences from massive scenes with sequence-level semantic labels. By calculating the similarity of text embeddings between the motion-related sub-prompt and sequence labels, we find the most matching sequence and acquire motion data in the form of SMPL-X parameters [55].

A.3 Metrics

Given the lack of ground truth videos for specific text queries, we utilize pre-trained quality-assessment models to evaluate the generated videos and their individual frames. In line with Comp4D [51], we employ Q-Align [62] as the benchmarking tool to assess the quality and aesthetics of the videos. The Q-Align rating, which ranges from 1 (worst) to 5 (best), is considered state-of-the-art, closely aligning with human judgments across established quality assessment benchmarks. Additionally, we include the CLIP score [63] to measure the alignment between the generated videos and the input texts, where higher scores signify better alignment.

B Results of Different Stages

As shown in Fig. 6, Fig. 7 and Fig. 8, we provide a detailed visualization of results obtained during different stages, including results of LLM-based task decomposition, results of coarse-grained trajectory generation with GPT-4V, and results of the final rendered videos.

C More Results of Controllable Generation

In this section, we present additional results on controllable generation. As demonstrated in Fig. 9, Fig. 10 and Fig. 11, we can achieve fine-grained control of the target without affecting other concepts in the 3D space.

https://doi.org/10.52202/079017-4184

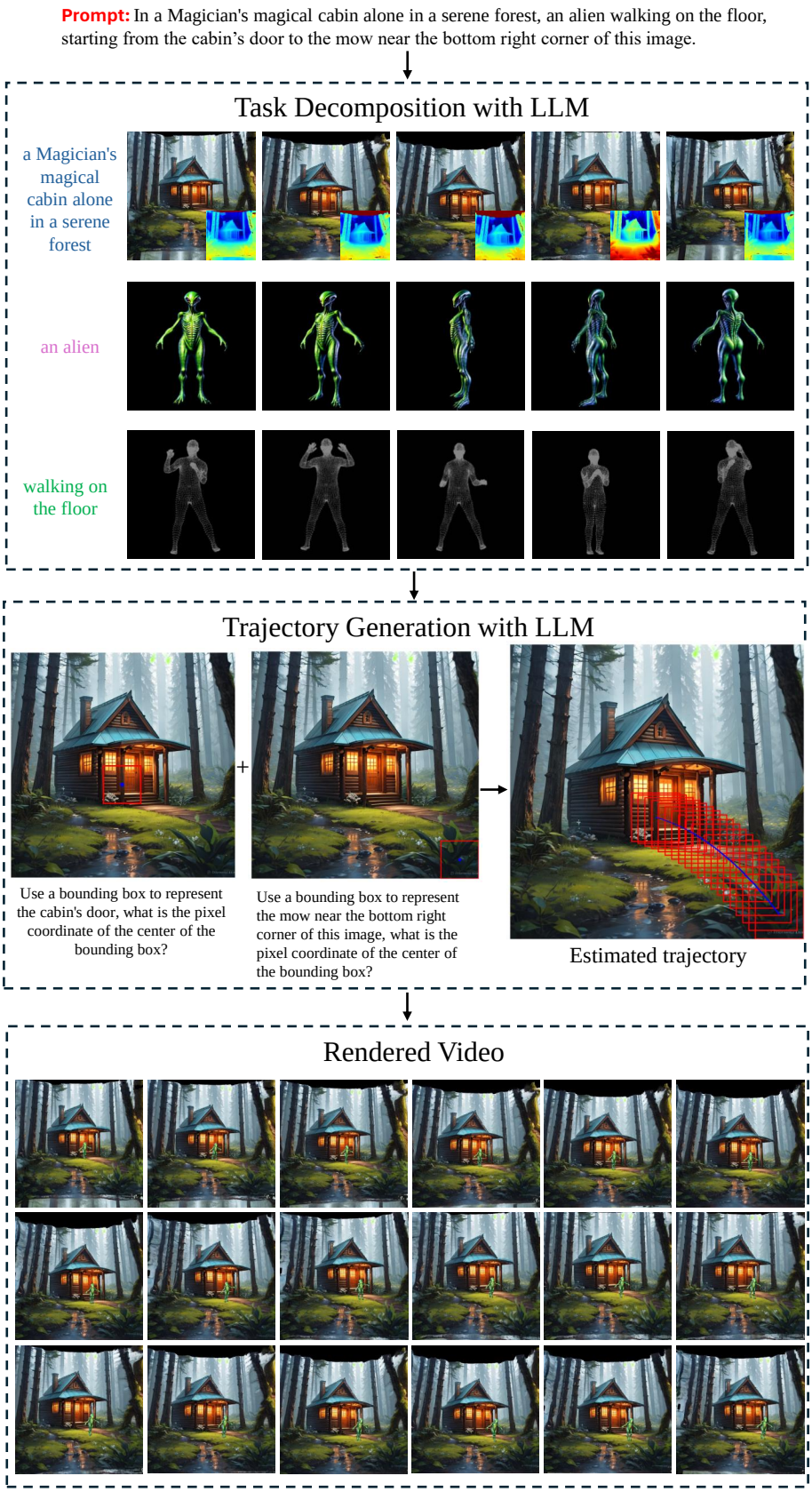


Figure 6: Results of different stages given the textual prompt: "*In a Magician's magical cabin alone in a serene forest, an alien walking on the floor, starting from the cabin's door to the mow near the bottom right corner of this image.*".

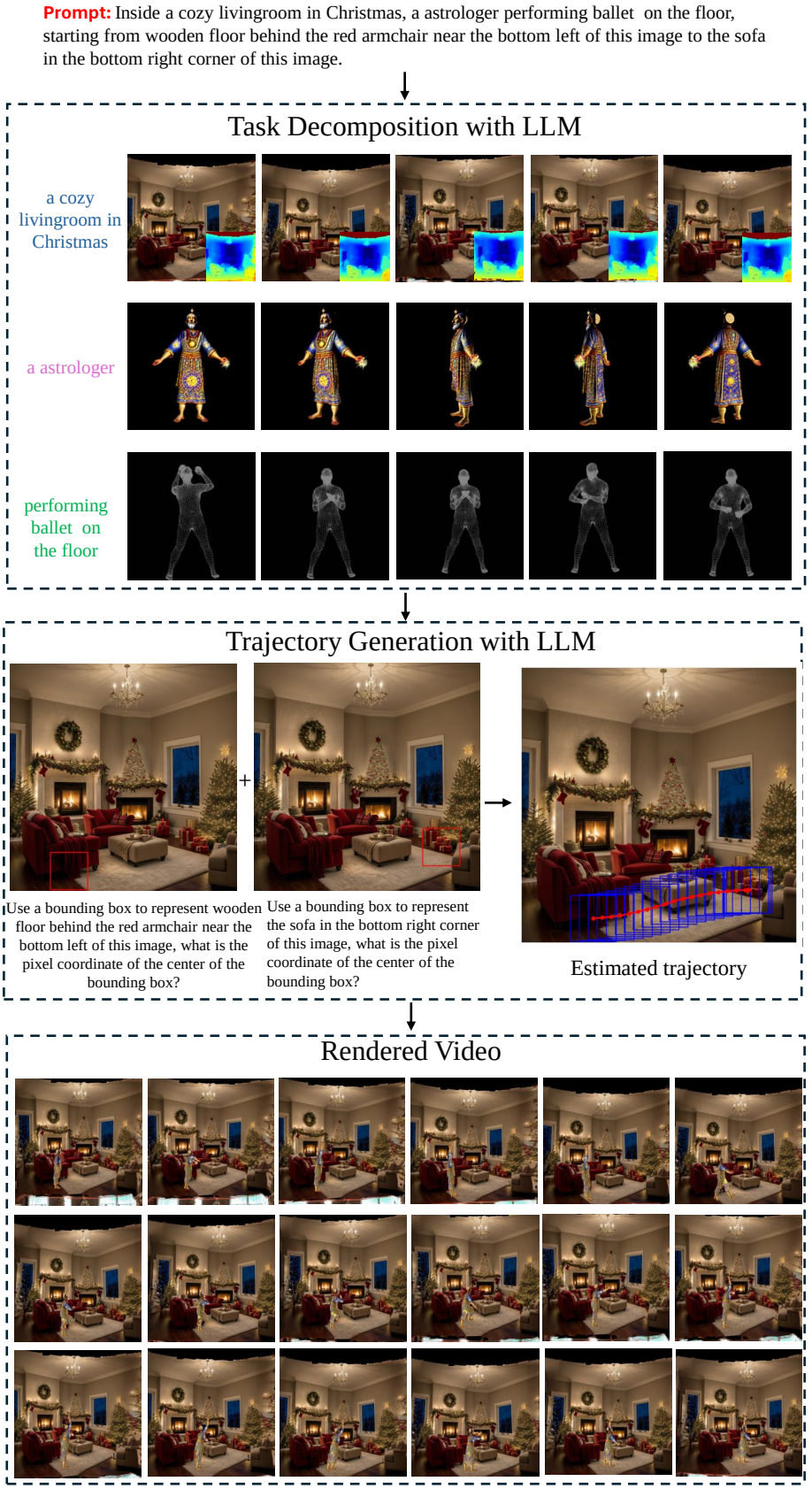


Figure 7: Results of different stages given the textual prompt: "*Inside a cozy livingroom in Christmas, a astrologer performing ballet on the floor, starting from wooden floor behind the red armchair near the bottom left of this image to the sofa in the bottom right corner of this image.*".

https://doi.org/10.52202/079017-4184

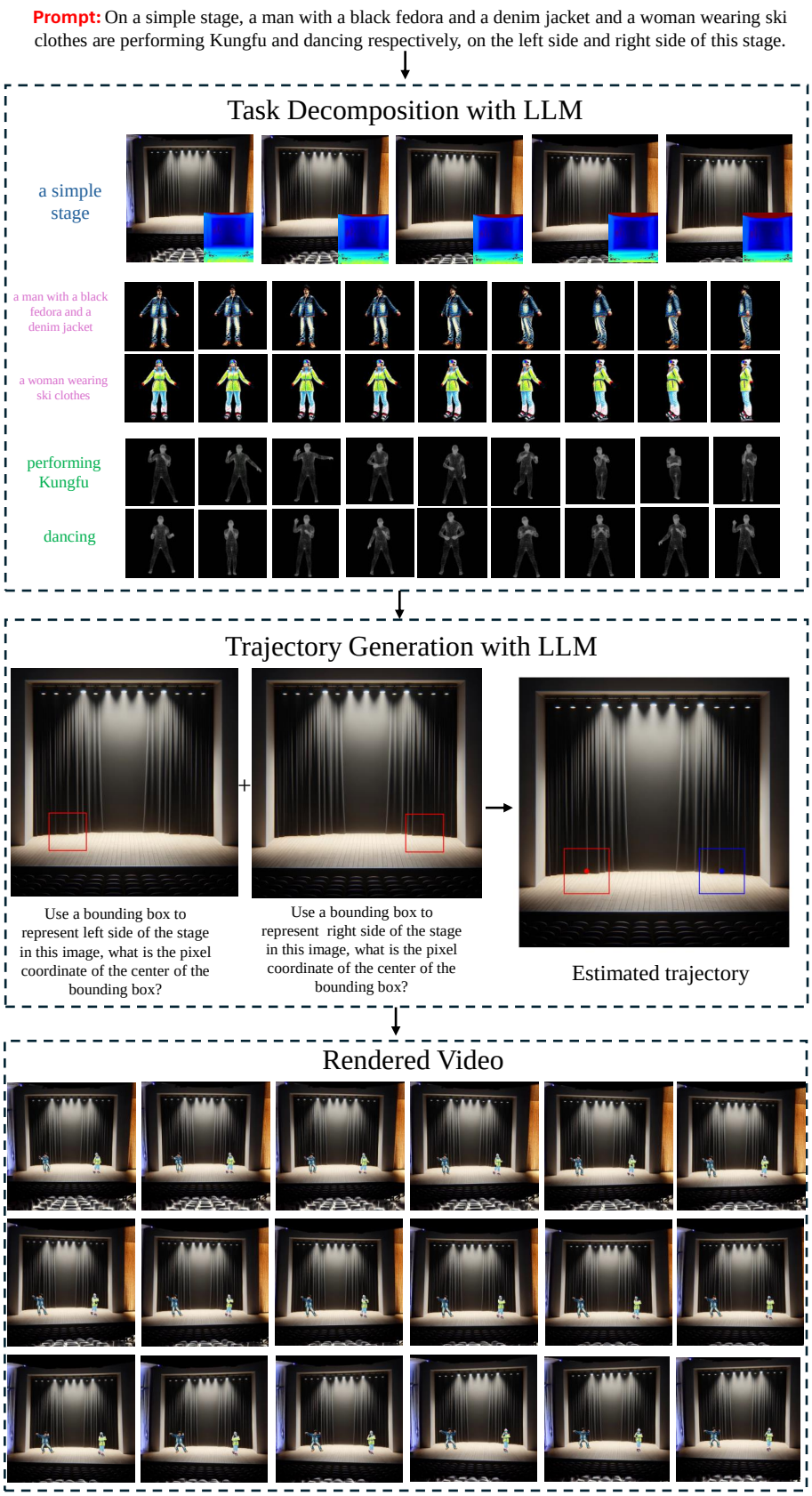
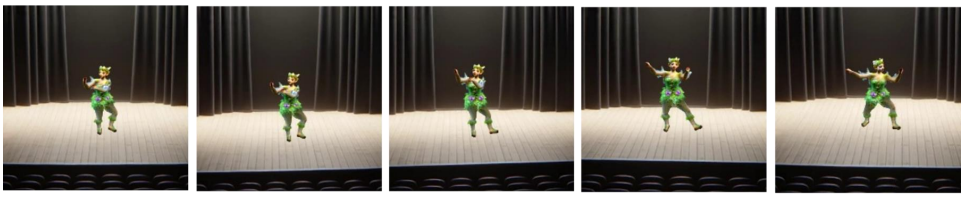
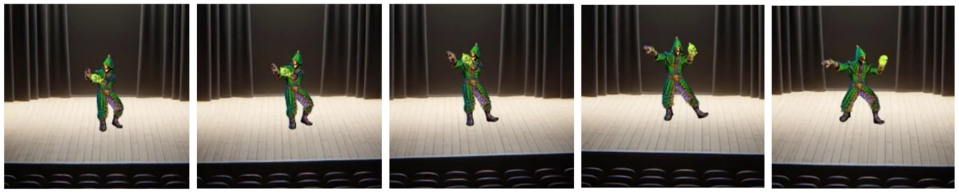


Figure 8: Results of different stages given the textual prompt: "*On a simple stage, a man with a black fedora and a denim jacket and a woman wearing ski clothes are performing Kungfu and dancing respectively, on the left side and right side of this stage.*".

(a) Prompt: A black man wearing a green t-shirt playing Kungfu on the stage.



(b) Prompt: Turn the character into a fairy.



(c) Prompt: Turn the character into a warlock.

Figure 9: Results of actor's appearance editing.

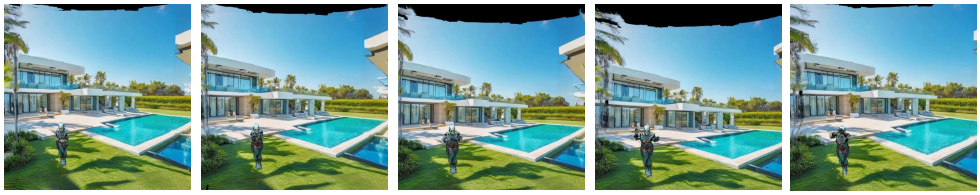
(a) Prompt: A demon slayer performing a han and tang dance on the stage.

(b) Prompt: Turn the motion of the character into a ballet.

(c) Prompt: Turn the motion of the character into a kin kong exotic dance.

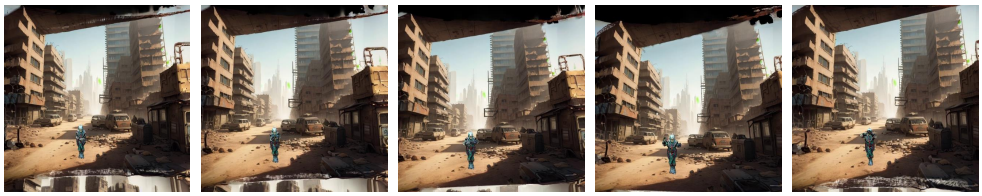
Figure 10: Results of actor's motion editing

https://doi.org/10.52202/079017-4184



(a) Prompt: A revenant dancing Daily Lovedive in a ultra-modern mega villa by the sea with swimming pool.

(b) Prompt: Turn the scene into a long anime-style road with anime-blocks and little anime-grass.



(c) Prompt: Turn the character into a postapocalyptic city in desert.

Figure 11: Results of scene editing.

https://doi.org/10.52202/079017-4184

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We clarify our contributions in Sec. 1.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We clarify our limitations in Sec. 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

https://doi.org/10.52202/079017-4184

Justification: This paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide implementation details in Sec. 5 and Sec. A.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We do not provide open access to the data and code in supplemental material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We provide implementation details in Sec. 5 and Sec. A.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Error bars are not reported because they would be too computationally expensive.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

https://doi.org/10.52202/079017-4184

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide computer resources used in Sec. A.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We report possible impacts of our method in Sec. 1 and Sec. 6.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite all models and datasets mentioned in the paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

 https://doi.org/10.52202/079017-4184

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The paper proposes a novel paradigm for text-to-video generation, named C3V. We plan to release the model's code along with comprehensive documentation to facilitate its use and replication.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The human avatars showed in this paper are all virtual humans that have no relationships with real human beings.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The human avatars showed in this paper are all virtual humans that have no relationships with real human beings.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-4184