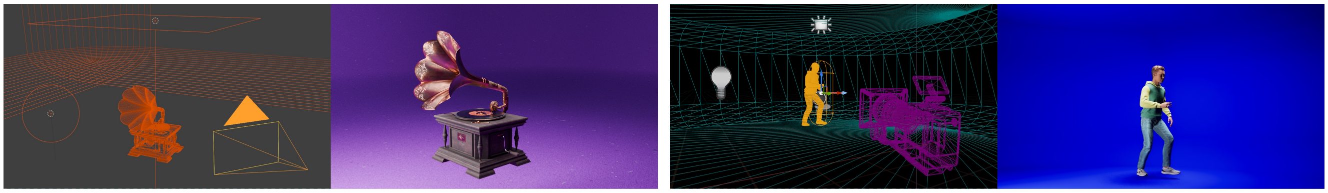
	Property Name	Choice	Description
Camera	Camera Focus Type	Follow	The camera focus follows the object.
		Fixed	The camera focus is static in the world space.
	Camera Focus Position	Upper, Center, Lower	The camera focus is at the upper/center/lower part of the object.
	Camera Movement Type	Truck, Dolly, Pedestal, Tilt, Pan, Spin, Following, Zoom	The basic camera movement types.
	Camera Movement Value	Scalar	How much the camera moves.
	Camera Initial Position	3D Position	The initial position of the camera.
	Camera Focal Length	Scalar	The scalar controls how much percentage of the object is visible on the screen.
Light and Environment	Scene Type	Env	The environment is given by a HDR environmental map. The map will also be used as the light source.
		Basic	The environment is an indoor room which color is controlled by "Scene Color" and has two light sources.
		Empty	The environment is empty but has two light sources or one environmental map as the light source.
	Scene Color	RGB color	The color for the indoor room when presented.
	Light Position	3D position	The position of the light when presented.
	Light Color	Scalar	The color temperature of the light when presented.
	Light Intensity	Scalar	The intensity of the light when presented.
	Ambient Light Intensity	Scalar	Ambient light intensity. The ambient light exists when the lights are used.
Render	Background Color	RGBA color	The background color of the location where the scene is empty.
	Render Engine	Blender/Unreal	
	Render Quality	High/Low	The quality of the rendering. We have two presets of rendering setting.

Table 10. The parameters used for controlling our rendering pipeline.



(a) Blender

(b) Unreal Engine

Figure 6. 3D scene setup in Blender and Unreal Engine. The wireframes and corresponding rendering outputs.

A. Appendix

A.1. Details of Synthetic Data Generation

Following a standard CGI production pipeline for creating videos, our synthetic video generation framework comprises two main modules: (1) *3D scene setup* and (2) *rendering*. Below, we provide a detailed overview of these modules and the specific parameters that govern them.

A.1.1. 3D Scene Setup

As discussed in Sec. 2, we focus on generating videos featuring a single object per scene. To achieve this, we build a procedural 3D scene generator driven by a carefully chosen set of parameters, enabling the production of a wide variety of synthetic videos. A typical 3D scene is composed of four main components: (1) the 3D object, (2) the camera, (3) the lighting conditions, and (4) the environment. We adopt this composition in our generator. Each component in our generator is controlled by a set of parameters, which we detail below.

3D Object. As we target single-object videos, we seek to include 3D assets that are both high-quality and highly varied. To this end, we collect assets from Objaverse 1.0 [15], Digital Twin Catalog [48], Blender Market [2], and Metahuman [18]. These sources collectively provide diverse asset categories and styles. We further filter assets from Objaverse based on categories, polygon count, view count, user ratings, and VLM to ensure overall quality. For other sources, we retain all assets since they are already curated with high fidelity.

Camera. We represent the camera using a set of parameters that capture real-world usage scenarios (see Table 10). These parameters include:

- *Camera movement type:* Determines the camera's trajectory around the object. In our experiments, we select one movement type at a time and quantify its extend using a parameter "Camera Movement Value".
- *Initial position and focus:* Specifies where the camera starts and how it focuses on the primary object.
- *Focal length:* Adjusts the camera's field of view relative to how much of the screen the object occupies.

Such parameterization allows us to mimic various camera

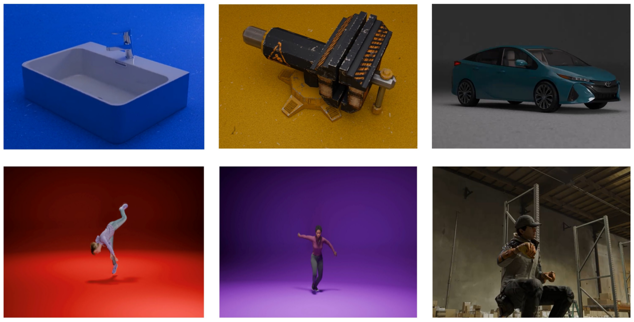


Figure 7. Examples of our synthetic video data. We render the synthetic videos with diverse background to alleviate the potential biases in synthetic videos.

behaviors from the real world.

Lighting and Environment. For simplicity, we jointly model the environment and its lighting conditions (see Table 10). Our parameterization supports three main configurations:

- *HDR environment map:* Provides both the background and primary light source. We use environment maps from Poly Haven [51].
- *Solid-color indoor room:* Uses two light sources (Figure 6) for illumination: one positioned above the object and another placed elsewhere in the scene.
- *Empty scene:* Lit by either an environment map or two lights for more controlled illumination with empty surroundings.

Although these settings may appear simple, they cover a wide range of lighting conditions and backdrop variations, thereby maintaining diversity while keeping the primary object prominent.

A.1.2. Rendering Setup

We employ two open-source rendering engines to generate high-quality video outputs:

- *Unreal Engine (Lumen):* We use Unreal Engine 5.4.4 with Lumen as our renderer with maximal render-quality settings to achieve realistic rendering effects [19].
- *Blender (Cycles):* We use Blender 4.2 and Cycles renderer configured with carefully chosen parameters to balance rendering speed and visual fidelity [62]

These engines offer robust rendering pipelines and physically based shading models, ensuring that our synthetic data closely reflects real-world lighting conditions.

A.1.3. Random Sampling of Parameter Space

To produce a large and diverse set of synthetic videos, we define a configuration ("config") file containing all relevant parameters described above. Figure 7 show some examples of synthetic videos with diverse setups. Our 3D scene generator parses this config file and sets up the scene. Then, the rendering engines render the scene into a video. For large-scale generation, we employ random sampling over each parameter's prescribed probability distribution, guided by the key insights from Sec. 3. Each sampling step produces a unique config file, which is then rendered into a separate synthetic video. This process enables us to generate a vast set of diverse synthetic videos with minimal manual intervention.

A.2. Implementation Details on Building the Reference Model

In building captions for the reference model, our goal is to omit the visual concepts we wish the mixed model to transfer and let the reference model to only captures the visual patterns of the synthetic data. We do that when we synthesis captions using VLM by removing tags. For example, if we were to feed these tags <human><motion><scene><metadata> into the mixed model and the desired visual concept is <motion>, then we would only feed <human><scene><metadata> to the VLM when we synthesis captions for the reference model. During inference, we also use captions that have tags omitted only for the reference model.

A.3. More Ablation Experiments and Visualizations

In this section, we provide additional visualizations of the data curation experiments and the ablation studies. Figure 8 and Figure 9 show the effect of using poor quality asset and rendering respectively. Figure 10 shows the effect of excessive training on synthetic data. Color patterns are introduced into the generation model. Figure 11 gives an example of fine-grained and generic captions and an example of using special tags. Figure 12 and Figure 13 show the comparison between videos from generation with and without *SimDrop*. Lastly, Figure 14 showcases the layer decompostion videos can use to separate out dynamic objects (*e.g.* animals, fluids) to enable video matting. Finally, Figure 15 shows more generated videos across all three tasks.

A.4. Evaluation Prompts

Large Human Motion

Dancing:

- A dancer practicing at home
- In a street setting, a teenager is performing breakdance moves, including leaning back, balancing on one leg, and rhythmically moving arms.
- An attractive man energetically dances, featuring lively movements. He crosses his arms and vigorously moves his legs, imitating horse riding and other whimsical actions.
- A young woman gracefully pirouettes on one foot, her other leg bent elegantly and arms outstretched for balance and flair. She transitions through various spins, show-

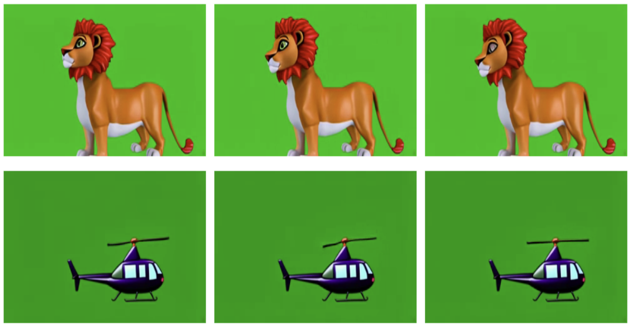


Figure 8. Example outputs from video generation models trained on synthetic datasets with low-quality assets. The resulting objects frequently exhibit cartoonish or animated characteristics, diverging from the intended original visual style.

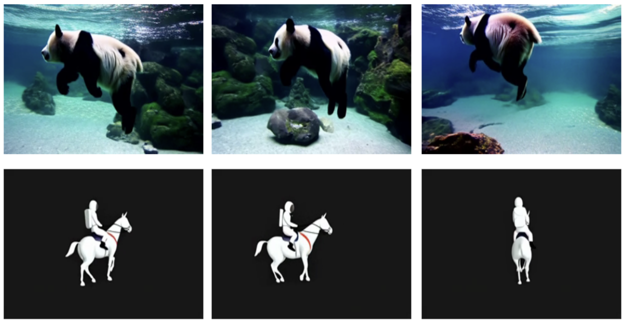


Figure 9. Visualization of generated outputs from video generation models trained with synthetic videos of low quality assets in large camera motion task. The objects in these generated videos more likely to appear static or animated.

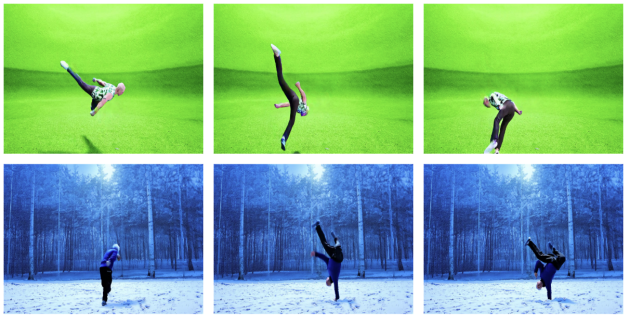
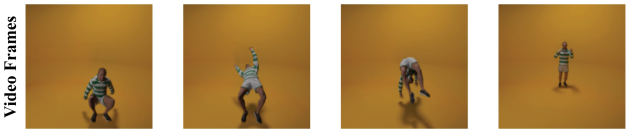


Figure 10. Visualization of over training video generation models trained with synthetic videos. Visual patterns such as color tone are more likely to appear in generated videos.

casing a dynamic dance routine that blends elements of northern soul dancing. She dances in a bustling urban plaza, or a serene beach at sunset, or a lively street festival, or, a beautifully lit dance studio. Each setting captures the fluidity and energy of her movements, adding depth and variety to her performance.



Generic Caption: The animated character performs backflip in a yellow background.

Fine-Grained Caption: The animated character executes a backflip by initially crouching low, launching itself upwards, rotating backwards in midair before returning to a standing position on its feet in a animated yellow background.

Special Tags: The animated character ... | **W.O. Special Tags**: A man in green and white hoodie ...

Figure 11. A comparison of generating captions for synthetic videos using existing methods (Generic Caption) and our method (Fine-Grained Caption). We also show a comparison of captions with special tags and without special tags.

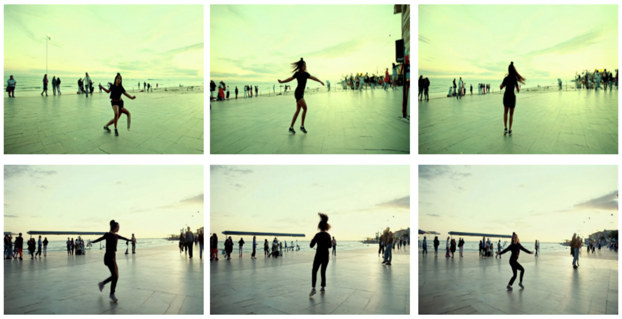


Figure 12. A comparison showcasing the effect of *SimDrop*. Row 1 is the result without *SimDrop* and Row 2 is the video with the method. The color tone in row two is significantly more better and without color pattern from the synthetic data.

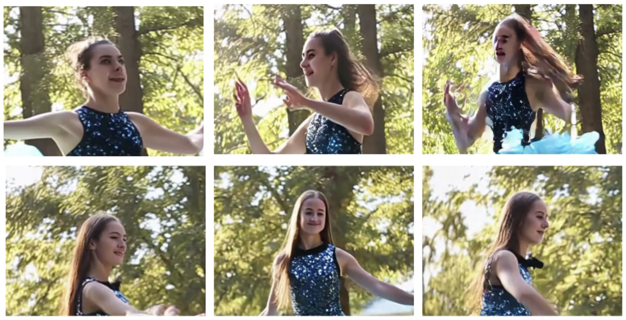


Figure 13. A comparison showcasing the effect of *SimDrop*. Row 1 is the result without *SimDrop* and Row 2 is the video with the method. The human faces in row two is significantly more realistic and appealing.

- A young woman is performing breakdance moves, including leaning back and balancing on one leg while engaging arms rhythmically.
- A woman dancing on grassland during sunset
- On a beach, an Ultraman from Japanese TV show is spinning around on one foot while keeping other leg bent and arms extended for balance and style. It performs multiple spins, emphasizing a dance move commonly associated with northern soul dancing.

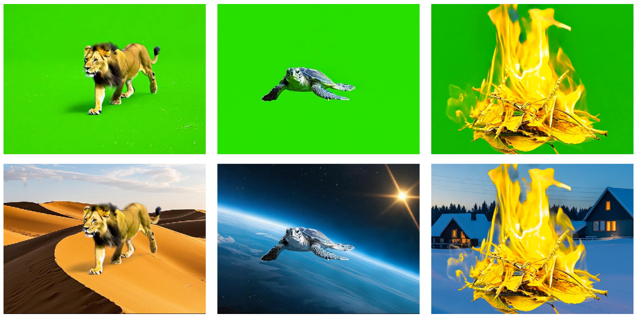


Figure 14. Example of background editing. Our layer generation enables easy background replacement via green-screen matting.

- In a bright dance room, a young woman is performing a dance with enthusiastic movements The person crosses arms and moves legs energetically, mimicking riding a horse and performing other playful gestures.
- A young woman is performing a breakdance move, starting with a dynamic step and then transitioning into a series of fluid body movements and rhythmic steps.
- A handsome man initiates with a dynamic step followed by a series of fluid body motions and rhythmic steps.

Gymnastics:

- In a bright dance room, A man executes a backflip by initially crouching low, launching himself upwards, rotating backwards in midair before returning to a standing position on his feet.
- In a well-lit dance studio, A woman performs a gymnastics moves to flip her body. Her backflip is to first crouch low, then rotating upwards and backward in midair, eventually landing back in a standing position.
- A man performs a backflip by first squatting down, then launching itself into the air, flipping backward, and finally landing back onfeet on grassland under sunshine.
- In a sunny grassland, a woman executes a backflip by initially crouching, then springing into the air, rotating backward, and ultimately landing on her feet.
- A female athlete performs a backflip by first squatting down, then launching itself into the air, flipping backward, and finally landing back onfeet during the floor execrise event at the Olympic Games.
- During the floor exercise event at the Olympic Games, a male athlete performs a stunning backflip. He begins by squatting down low, gathering his strength and focus. With a powerful burst of energy, he launches himself into the air, his body gracefully arching as he flips backward. The sunlight glints off his muscular form as he completes the rotation, and he lands solidly on his feet, his expression a mix of concentration and triumph.
- A man Moves with dynamic energy, shifting from a standing position to a deep crouch, then rotating her body midair before landing upright on the sunlit grassland.
- A woman is moving dynamically, transitioning from a standing position to a deep crouch and then rotating body mid-air before returning to an upright stance on grassland under sunshine.
- During the floor exercise event at the Olympic Games, a female athlete moves with dynamic precision. She transitions from a standing position to a deep crouch, then launches herself into the air, rotating her body mid-flight before landing gracefully back on her feet.
- At the Olympic Games' floor exercise event, a male athlete showcases his agility by swiftly dropping into a deep crouch from a standing position. He then propels himself into the air, executing a mid-air rotation, and lands back on his feet with precision and grace.

Large Camera Motion

- A lion standing on the grass. spin shot.
- An astronaut riding a horse, high definition, 4k. spin shot.
- A panda swimming underwater. spin shot.
- Video of sailboat on a lake during sunset. spin shot.
- Variety of succulent plants on a garden. spin shot.
- A birthday cake in the plate. spin shot.
- Big cargo ship passing on the shore. spin shot.
- Time lapse video, sunrise of the Great Wall. spin shot.
- A tree with Halloween decoration. spin shot.
- A Labrador dog wearing glasses and casual clothes is lying on the bed reading. spin shot.

Layer Decomposition

- A lion standing in a green background.
- A lion running in a green background.
- Turtle swimming in a green background.
- An african penguin walking in a green background.
- Variety of succulent plants in a green background.
- Leaves swaying in the wind in a green background.
- A stack of dried leaves burning in a green background.
- Big cargo ship like in the movies passing in a green background.
- Helicopter landing in a green background.
- A young woman is performing breakdance moves, including leaning back and balancing on one leg while engaging arms rhythmically in a light blue background.

A.5. Human Evaluation Details

Our user study videos are available on the project website. We invite the community to also rate the videos.

Large Human Motion For large human motions, we asks our human raters to examine how many out of the generated videos in each video show no collapse in human body structure. Specifically, we ask them to focus on the limbs and torso areas. The detailed rules are as following: 1. Does the video include the full body of the person (all four limbs) for more than 2 seconds? 2. Is the video bascially showing what is specified by the prompt, including background and motion? 3. Does the person in the video looks animated? 4.

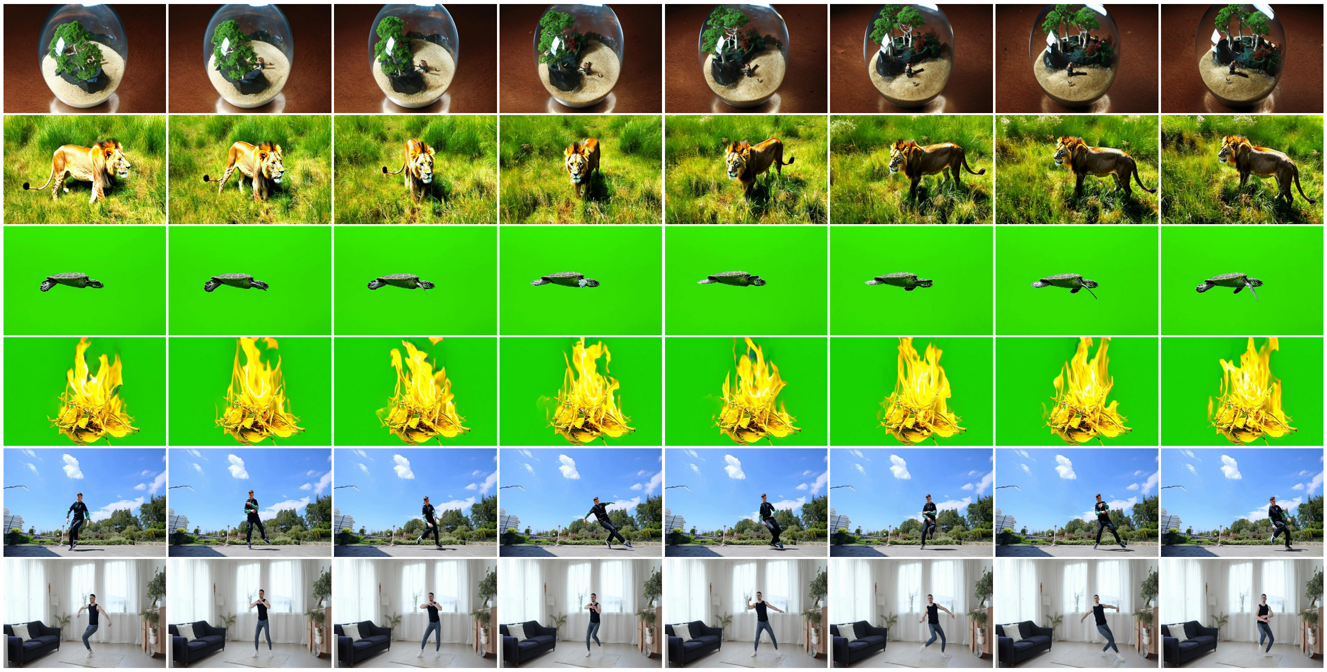


Figure 15. More visualization of generated videos for large camera motion (row 1,2), layer decomposition (row 3,4), and large human motion(row 5,6).

Is there limbs or torso addition/missing from the video? 5. Is there transition of body parts that are obviously unnatural (e.g. switching body parts at the same location)?

Please Note: 1. DO NOT focus your judgement on these part of the human body: hands, feet, or face 2. DO NOT judge the asethetics or naturalness of the human motion, please just focus on human body integrity

Large Camera Motion For Large camera motion, we instruct the human raters to focus on the object and the degree which the picture rotates. The detailed rules are as following: If any of the following question is yes, please mark the video as 0 1. If the object appear in the video is corrupt, unnatural, or animated 2. If the background is not of pure color as instructed by the prompt

Layer Decomposition For layer decompostion, we instruct the human raters to focus on the object and the background quality. The detailed rules are as following: If any of the following question is yes, please mark the video as 0 1. If the object appear in the video does not spin at all. 2. If the object appear in the video spins but the background does not move with the object 3. If the object appear in the video corrupts, becomes unnatural or looks animated.