A. More Qualitative Results

A.1. The Properties of Appearance Noise Prior.

we fist examine the influence of the Appearance Noise Prior on the quality of generated video results by varying the parameters λ and γ. As illustrated in Fig. 3, videos generated with the integration of the Appearance Noise Prior display heightened coherence and superior quality in comparison to those generated without this prior. The introduced prior proves beneficial by endowing the model with enhanced capabilities to preserve distinctive characteristics of input images, even when they deviate from the training data in WebVid-10M.

Higher resolution generation. Besides, our empirical findings indicate that adjusting the ratio of the Appearance Noise Prior contributes to the production of high-resolution videos by our model. As illustrated in Fig. 4, the model demonstrates effective generation of 512x512 resolution videos, surpassing its original training resolution of 320x320, thanks to the integration of the Appearance Noise Prior.

Reduce the number of sampling steps. Additionally, we discover that the Appearance Noise Prior plays a crucial role in enhancing the efficiency of the diffusion process. As illustrated in Fig. 5, in situations involving simpler motion patterns, the integration of the Appearance Noise Prior empowers the network to produce satisfactory results even with a reduced sampling step count, set at 5. This decrease in steps significantly improves the efficiency of video production. For instance, employing our base image&text-to-video model to generate a 9-frame video at 2 fps now requires only 1.3 seconds.

Adjust the amplitude of motion. During the inference stage, the introduction of γ serves as a controllable parameter for adjusting the intensity of motion in the generated videos. As illustrated in Fig. 6, setting $\gamma = 0.02$ produces results with small yet consistent movements, while $\gamma = -0.01$ results in larger but less stable motions. Considering that the FVD metric favors stable yet discernible motion, we opted for $\gamma = 0.02$ in our FVD evaluation, striking the best tradeoff.

A.2. More qualitative Results of MicroCinema

In this section, we present additional video generation results Fig. 7 and Fig. 8. We utilize Midjourney as the initial stage text-to-video model. It is evident that the videos generated through our method not only maintain aesthetic quality in imagery but also exhibit clear and coherent motion.

A.3. Qualitative Comparison with Previous Work.

We provide additional examples for comparison with previous works in Fig. 9, Fig. 10, and Fig. 11. Our approach demonstrates the ability to generate visually stunning videos, akin to cinematic quality. In comparison to prior work, it showcases superior image quality, enhanced temporal consistency, greater stylistic diversity, and improved textual coherence.

B. Proof of Appearance Noise Prior

In this section, we present a proof of the compatibility of the Appearance Noise Prior with all ODE samplers. We demonstrate that incorporating the Appearance Noise Prior and employing new noise as supervision does not necessitate alterations to the sampler process itself. Instead, it only requires modifications to the initial noise during sampling.

B.1. Denoising Diffusion Probabilistic Models

Firstly, we introduce the standard framework of Denoising Diffusion Probabilistic Models (DDPM). The forward process in DDPM, when articulated in discrete form, is as follows:

$$\boldsymbol{z}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{z}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \quad t = 1, \dots, T, \tag{1}$$

$$\boldsymbol{z}_t = \sqrt{\alpha_t}\boldsymbol{z}_{t-1} + \sqrt{\beta_t}\boldsymbol{\epsilon}_{t-1}. \tag{2}$$

The corresponding Stochastic Differential Equation (SDE) process of the DDPM can be represented by a unified expression, given by the following equation:

$$\mathrm{d}\boldsymbol{z} = f(\boldsymbol{z}, x)\,\mathrm{d}x + g(x)\mathrm{d}\boldsymbol{w}, \quad x \in [0, 1], \tag{3}$$

where $\boldsymbol{w}$ is a standard Wiener process. To derive the expressions for $f(\boldsymbol{z}, x)$ and $g(x)$, as T approaches infinity, two continuous functions, $\bar{\alpha}(t)$ and $\beta(t)$, can be defined:

$$\bar{\alpha}(x), x \in [0, 1], \bar{\alpha}(x = \frac{t}{T}) = \bar{\alpha}_t, \tag{4}$$

$$\beta(x), x \in [0, 1], \beta(x = \frac{t}{T}) = T\beta_t, \tag{5}$$

where $\bar{\alpha}_t$ and β_t are the coefficients corresponding to those in equations Eq. (1) and Eq. (2), respectively. By substituting Eq. (5) into Eq. (2), utilizing $\alpha_t = 1 - \beta_t$, and considering the limit as $T \to \infty, \beta_t \to 0$, and subsequently applying a Taylor series expansion for approximation, equation Eq. (2) can be reformulated as follows:

$$\boldsymbol{z}_t = (1 - \beta(\frac{t}{T})\frac{1}{2T})\boldsymbol{z}_{t-1} + \sqrt{\frac{\beta(\frac{t}{T})}{T}}\,\boldsymbol{\epsilon}_{t-1}. \tag{6}$$

By setting $x = t/T$, and incorporating $\boldsymbol{w}$ into the equation, we obtain:

$$\boldsymbol{z}(x) - \boldsymbol{z}(x - \mathrm{d}x) = -\frac{\beta(x)}{2}\boldsymbol{z}(x - \mathrm{d}x)\mathrm{d}x + \sqrt{\beta(x)}\mathrm{d}\boldsymbol{w}. \tag{7}$$

Upon simplification, we obtain the Stochastic Differential Equation (SDE) formulation of DDPM:

$$\mathrm{d}\boldsymbol{z}(x) = -\frac{\beta(x)}{2}\boldsymbol{z}(x)\mathrm{d}x + \sqrt{\beta(x)}\mathrm{d}\boldsymbol{w}, \quad x \in [0,1]. \tag{8}$$

Comparing with Eq. (3), we can deduce:

$$f(\boldsymbol{z}, x) = -\frac{\beta(x)}{2}\boldsymbol{z}, \quad g(x) = \sqrt{\beta(x)}. \tag{9}$$

For the SDE process described in Eq. (3), the corresponding reverse Ordinary Differential Equation (ODE) process is represented by the following equation:

$$\mathrm{d}\boldsymbol{z} = f(\boldsymbol{z}, x)\,\mathrm{d}x - \frac{1}{2}g(x)^2\nabla_{\boldsymbol{z}}\log p_x(\boldsymbol{z})\,\mathrm{d}x. \tag{10}$$

Given that $\boldsymbol{z}(x)$ follows a Gaussian distribution $N(\sqrt{\bar{\alpha}(x)}\boldsymbol{z}(0), (1-\bar{\alpha}(x))I)$, its score function can be related to the noise as follows:

$$\nabla_{\boldsymbol{z}}\log p_x(\boldsymbol{z}) = -\frac{\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{z}(x), x)}{\sqrt{1-\bar{\alpha}(x)}}, \tag{11}$$

where $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{z}(x), x)$ is estimated using the following loss function:

$$\mathcal{L}_{\boldsymbol{\theta}} = \mathbb{E}_{q(\boldsymbol{z}(x))}\left[\|\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{z}(x), x) - \boldsymbol{\epsilon}\|^2\right], \tag{12}$$

where $q(\boldsymbol{z}(x))$ denotes the noisy data distribution of $\boldsymbol{z}(x)$ and $x \sim U[0,1]$. By utilizing equation Eq. (11), for DDPM models that implement the ϵ-prediction, the reverse ODE process is articulated as follows:

$$\mathrm{d}\boldsymbol{z} = f(\boldsymbol{z}, x)\,\mathrm{d}x + \frac{\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{z}(x), x)}{2\sqrt{1-\bar{\alpha}(x)}}g(x)^2\mathrm{d}x. \tag{13}$$

By incorporating Eq. (9), the final form can be derived as follows:

$$\mathrm{d}\boldsymbol{z} = -\frac{\beta(x)}{2}\boldsymbol{z}\mathrm{d}x + \frac{\boldsymbol{\epsilon}_{\boldsymbol{\theta}}(\boldsymbol{z}(x), x)}{2\sqrt{1-\bar{\alpha}(x)}}\beta(x)\mathrm{d}x. \tag{14}$$

B.2. Appearance Noise Prior

To simplify notation, let $\boldsymbol{\mu} = \lambda[\boldsymbol{z}^c, \boldsymbol{z}^c, ..., \boldsymbol{z}^c]$, where $\boldsymbol{z}^c$ represents the center frame of $\boldsymbol{z}_0$ and varies with the input video $\boldsymbol{z}_0$.This formulation explicitly demonstrates that $\boldsymbol{\mu}$ is a function of $\boldsymbol{z}_0$, thereby directly linking it to the variations present in the input video. Then the forward process of Appearance Noise Prior is change to:

$$\boldsymbol{z}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{z}_0 + \sqrt{1-\bar{\alpha}_t}(\boldsymbol{\epsilon} + \boldsymbol{\mu}), \quad t = 1, \ldots, T, \tag{15}$$

$$\boldsymbol{z}_t = \sqrt{\alpha_t}\boldsymbol{z}_{t-1} + (\sqrt{1-\bar{\alpha}_t} - \sqrt{\alpha_t - \bar{\alpha}_t})\boldsymbol{\mu} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{t-1}. \tag{16}$$

Applying a transformation to the coefficient preceding $\boldsymbol{\mu}$ in Eq. (16) yields:

$$\boldsymbol{z}_t = \sqrt{\alpha_t}\boldsymbol{z}_{t-1} + \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t} + \sqrt{\alpha_t - \bar{\alpha}_t}}\boldsymbol{\mu} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{t-1}. \tag{17}$$

Similar to equation (6), considering the limit conditions $T \to \infty, \beta_t \to 0, \alpha_t \to 1$, equation Eq. (17) can be reformulated as follows:

$$\boldsymbol{z}_t = (1 - \beta(\frac{t}{T})\frac{1}{2T})\boldsymbol{z}_{t-1} + \frac{\beta(\frac{t}{T})}{2T\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\mu} + \sqrt{\frac{\beta(\frac{t}{T})}{T}}\boldsymbol{\epsilon}_{t-1}. \tag{18}$$

By setting $x = t/T$, and incorporating $\boldsymbol{w}$ into the equation, we obtain:

$$\mathrm{d}\boldsymbol{z}(x) = -\frac{\beta(x)}{2}\boldsymbol{z}(x)\mathrm{d}x + \frac{\beta(x)}{2\sqrt{1-\bar{\alpha}(x)}}\boldsymbol{\mu}\mathrm{d}x + \sqrt{\beta(x)}\mathrm{d}\boldsymbol{w}. \tag{19}$$

Comparing with Eq. (3), we can deduce:

$$f(\boldsymbol{z}, x) = -\frac{\beta(x)}{2}\boldsymbol{z} + \frac{\beta(x)}{2\sqrt{1-\bar{\alpha}(x)}}\boldsymbol{\mu}, \quad g(x) = \sqrt{\beta(x)}. \tag{20}$$

Reverse ODE process can be represented by the following equation:

$$\mathrm{d}\boldsymbol{z} = f(\boldsymbol{z}, x)\,\mathrm{d}x - \frac{1}{2}g(x)^2\nabla_{\boldsymbol{z}}\log p_x(\boldsymbol{z})\,\mathrm{d}x. \tag{21}$$

As we employ the following form of the loss function:

$$\mathcal{L}_{\boldsymbol{\theta}} = \mathbb{E}_{q(\boldsymbol{z}(x))}\left[\|\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{z}(x), x) - (\boldsymbol{\epsilon} + \boldsymbol{\mu})\|^2\right]. \tag{22}$$

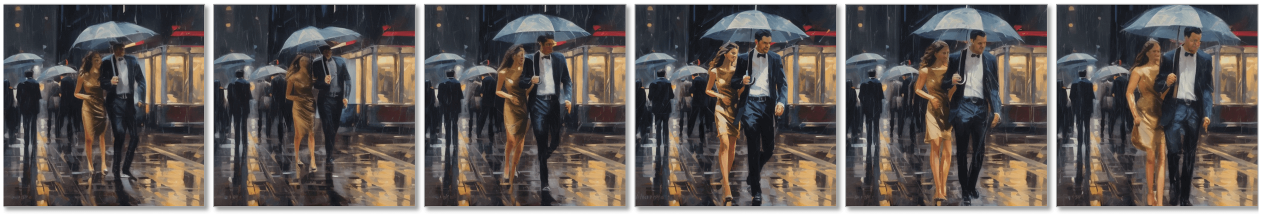
Therefore, the relationship between the score function and the network's estimated value $\boldsymbol{f}_{\boldsymbol{\theta}}$ becomes:

$$\nabla_{\boldsymbol{z}}\log p_x(\boldsymbol{z}) = -\frac{\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{z}(x), x) - \boldsymbol{\mu}}{\sqrt{1-\bar{\alpha}(x)}}. \tag{23}$$

By substituting Eq. (23) and Eq. (20) into Eq. (21), and noting that the coefficient preceding $\boldsymbol{\mu}$ is eliminated, we obtain the final form of the Reverse ODE:

$$\mathrm{d}\boldsymbol{z} = -\frac{\beta(x)}{2}\boldsymbol{z}\mathrm{d}x + \frac{\boldsymbol{f}_{\boldsymbol{\theta}}(\boldsymbol{z}(x), x)}{2\sqrt{1-\bar{\alpha}(x)}}\beta(x)\mathrm{d}x. \tag{24}$$

In the context of the Appearance Noise Prior, $\boldsymbol{f}_{\boldsymbol{\theta}}$ functions as the network's output, paralleled by $\boldsymbol{\epsilon}_{\boldsymbol{\theta}}$ in the DDPM framework. Notably, Eq. (24) and Eq. (14) exhibit identical forms. This similarity enables the straightforward integration of existing ODE sampling algorithms, with the only requisite modification being the adjustment of the initial sampling noise.



“An oil painting of a couple in formal evening wear going home get caught in a heavy downpour with umbrellas.”

Figure 1. When the facial features are extremely small, the model struggles to generate high-quality facial representations.

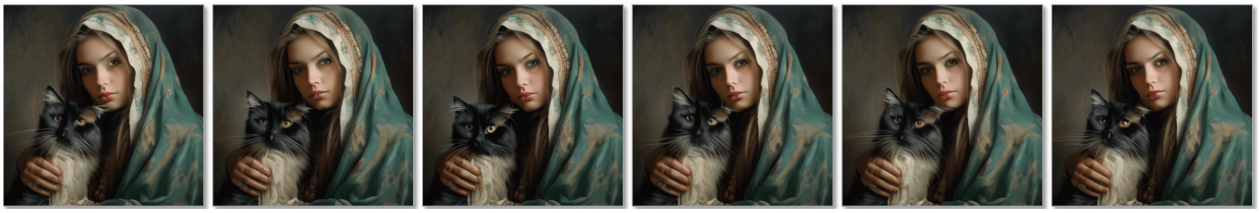


Figure 2. When the facial features are large, the model performs significantly better.

B.3. Implementation of Appearance Noise Prior

The implementation of Appearance Noise Prior in noise prediction models is straightforward. Traditionally, noise is added and trained using samples from a standard Gaussian distribution. With the Appearance Noise Prior, we modify this approach by superimposing an image prior $\lambda[\boldsymbol{z}^c, \boldsymbol{z}^c, ..., \boldsymbol{z}^c]$ onto the original noise, creating a new noise term for noise addition and supervision. During inference with ODE samplers, the initial sampling noise should be changed from $\mathcal{N}(\mathbf{0}, \boldsymbol{I})$ to $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{I})$, where $\boldsymbol{\mu} = \lambda[\boldsymbol{z}^c, \boldsymbol{z}^c, ..., \boldsymbol{z}^c]$. To achieve more consistent results, the strength of the prior can be appropriately enhanced by adjusting $\boldsymbol{\mu}$ to $(\lambda + \gamma)[\boldsymbol{z}^c, \boldsymbol{z}^c, ..., \boldsymbol{z}^c]$, thereby improving the consistency of the generated videos.

C. Implementation Details

For text-to-image stage, we use SD2.1-Base and SDXL for Quantitative Experiments. Specific details of the samples are provided in the Tab. 1. And we use Midjourney and DALL-E 2 for Qualitative Results. In the Tab. 3, we present the specific details of our image&text-to-video model. For the spatial layer, we utilized the SD2.1-Base model architecture and initial parameters. Additionally, we incorporated the VAE provided by SD2.1-Base, along with the CLIP text encoder, both of which were frozen during the training process. The Tab. 2 displays the parameter count for each component; the image&text-to-video model possesses 2.0 billion parameters, which were actively trained, while the spatial learning rate was set to one-tenth of the temporal learning rate.

Table 1. Text-to-Image Model sampling parameters, generation time test on one A100-80GB.

Sampling Parameters	SD2.1-Base	SDXL
Sampler	EulerEDM	
Steps	50	
Text guidance scale	7.5	
Image resolution	512x512	1024x1024
Generation time	2 s	9 s

Table 2. Number of Model Parameters

Model Name	# Params
Base image&text-to-video	2 B
CLIP text encoder	354 M
VAE	84 M

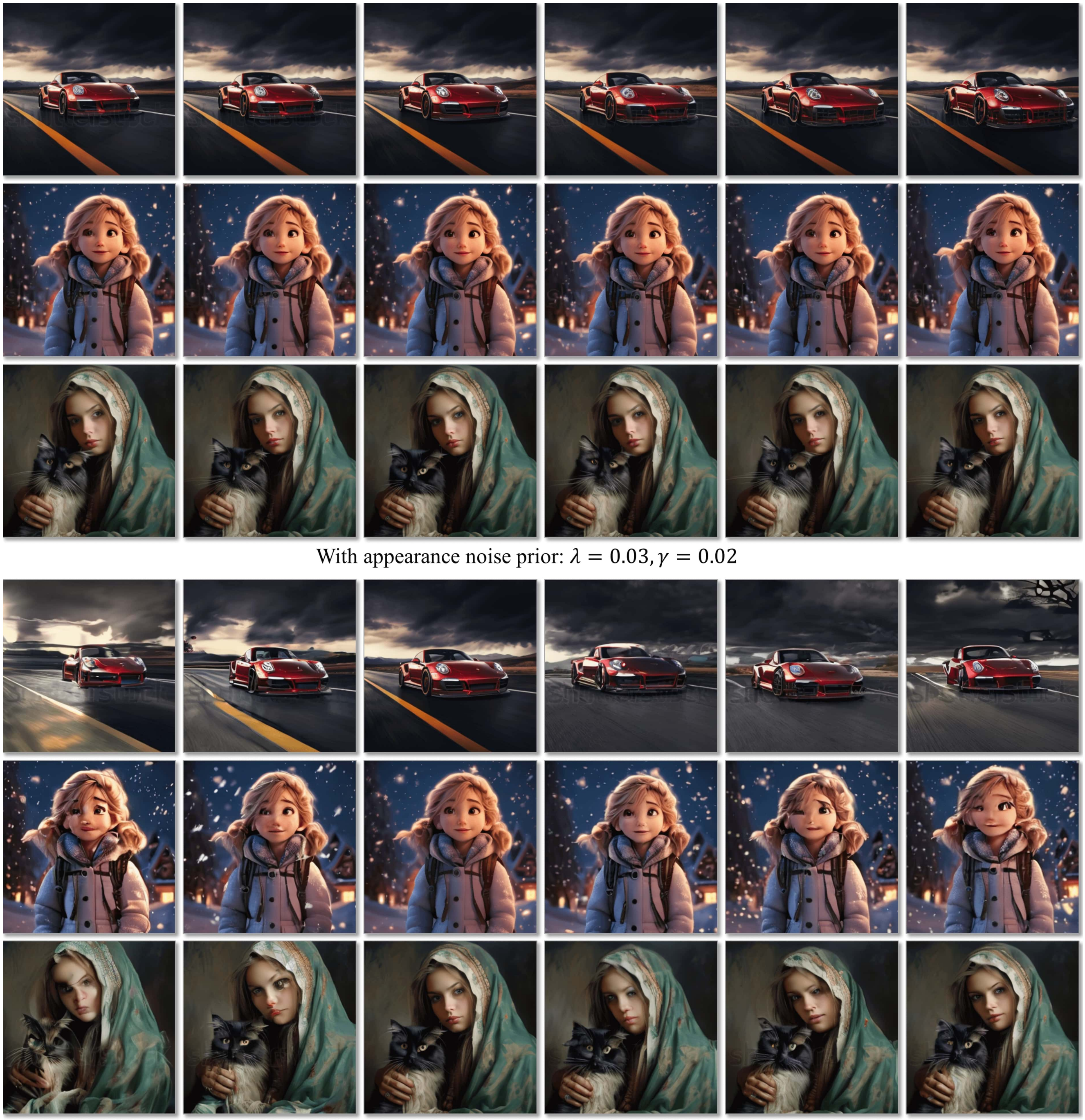
D. Limitations

Our method is based on the latent diffusion approach of SD2.1, utilizing an SD-pretrained VAE to encode images into the latent space. Currently, the VAE exhibits limited reconstruction capabilities for small objects, particularly small faces, leading to sub-optimal performance in these cases, as illustrated in the Fig. 1. Conversely, the model performs significantly better with larger faces, also demonstrated in the Fig. 2. To address this issue, it is necessary to re-train the VAE with increased channel size.

Another limitation of our approach is that we focused solely on temporal super-resolution (TSR) without incorporating spatial super-resolution (SSR). Ideally, a joint spatial-temporal super-resolution process could potentially achieve further improvements in the quality of the generated videos. This will be one of our future work.

Table 3. Hyperparameters for our diffusion models are detailed as follows. In the spatial layers, we utilize the pretrained SD2.1-Base, as previously discussed. The term $\gamma^\dagger$ represents a hyperparameter specific to the EulerEDM sampler, with $\gamma^\dagger = 0$ indicating the use of an ODE sampler.

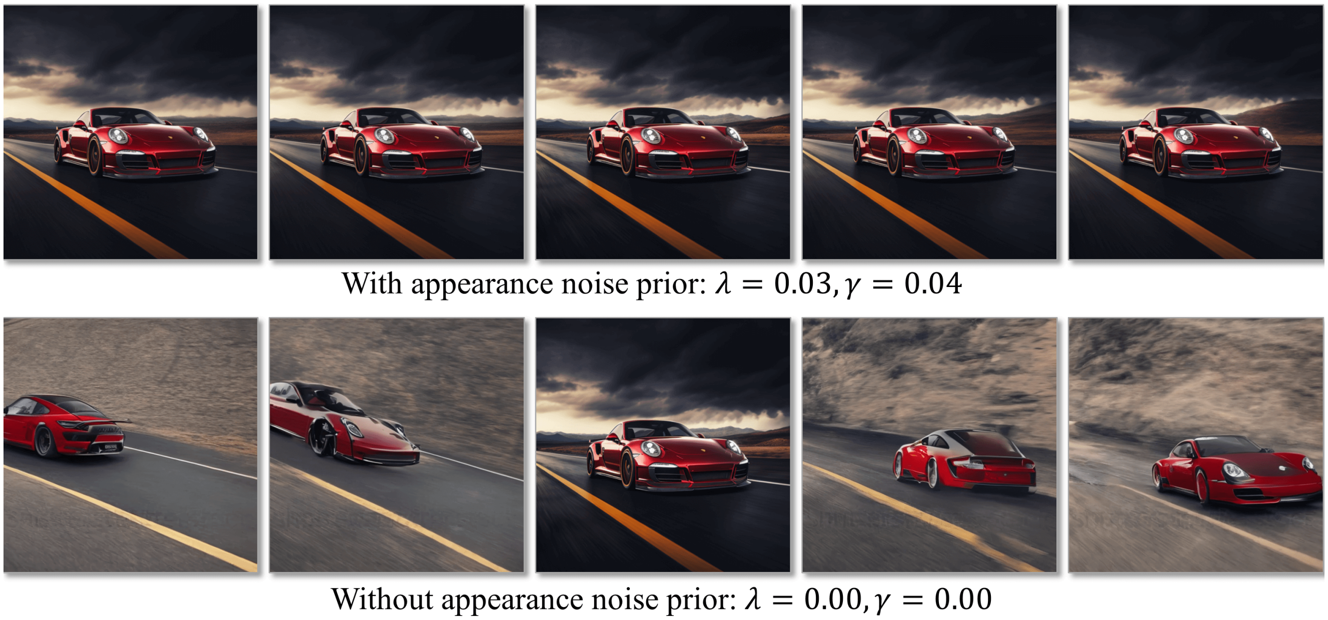
Hyperparameter	**Base Image&Text-to-Video Model**	**Temporal Interpolation Model**
Temporal Layers		
Architecture		
Input shape (C,N,H,W)	4,9,40,40	4,5,40,40
Model channels	320	
Channel multipliers	[1,2,4,4]	
Attention resolutions	[4,2,1]	
Head channels	64	
Positional encoding	Sinusoidal	
Temporal conv kernel size	3,1,1	
Temporal attention size	9,1,1	5,1,1
Image Conditioning		
Condition frame	$\boldsymbol{z}^c$	$\boldsymbol{z}^1, \boldsymbol{z}^N$
Extending into video	Repeat	Interpolate
Text Conditioning		
Embedding dimension	1024	-
CA resolutions	[4, 2, 1]	-
CA sequence length	77	-
Drop rate	0.1	1.0
Training		
# train steps	800K	40K
Learning rate	2×10^{-5}	2×10^{-5}
Batch size per GPU	4	4
# GPUs	4	4
GPU-type	A100-80GB	A100-80GB
Training data FPS	2	8, 30
Prediction mode	eps-pred	
Diffusion Setup		
Diffusion steps	1000	
Noise schedule	Linear	
β_0	0.00085	
β_T	0.0120	
Appearance noise prior λ	0.03	
Sampling Parameters		
Sampler	EulerEDM	
Steps	50	
$\gamma^\dagger$	0	
Text guidance scale	7.5	1.0
Appearance noise prior $(\lambda + \gamma)$	0.03+0.02	
Generation Time	12 s	7 s



With appearance noise prior: $\lambda = 0.03, \gamma = 0.02$

Without appearance noise prior: $\lambda = 0.00, \gamma = 0.00$

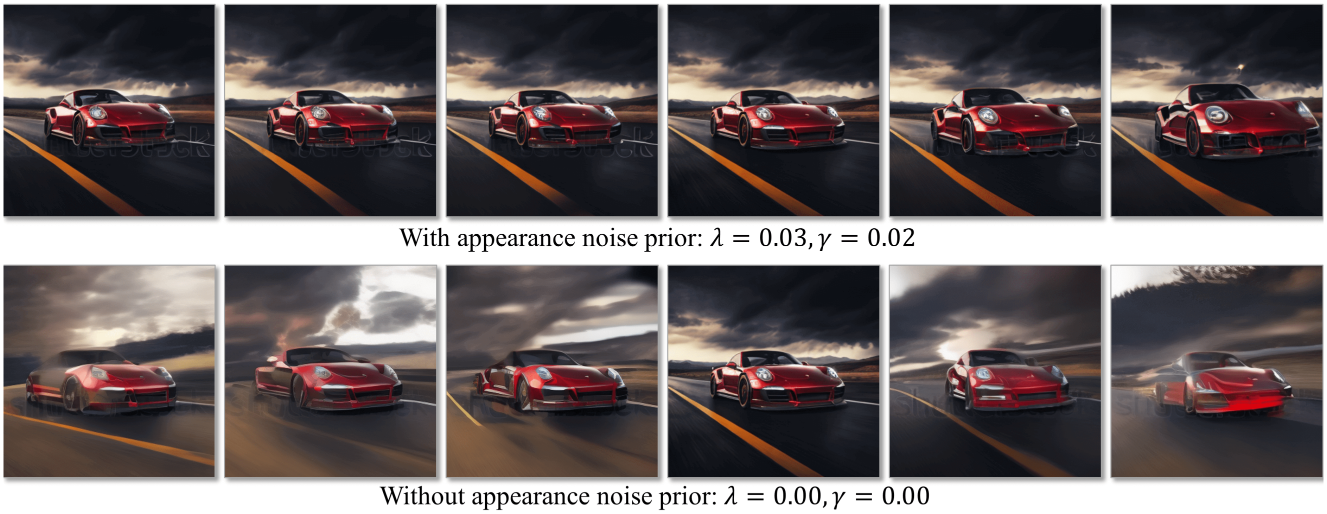
Figure 3. The appearance noise prior is very useful for the model in maintaining the appearance of beautiful images. The prompts for the three videos, arranged from top to bottom, are as follows: 'Red Porsche running on the road, high resolution, 8k', 'Disney animation style, One frosty day, when snow blanketed everything like a white quilt, a little girl named Zosia was coming home from school. With gloves keeping her hands warm and a cozy jacket, she walked along the path', and 'Persian cat on a beautiful Polish woman.'



With appearance noise prior: $\lambda = 0.03, \gamma = 0.04$

Without appearance noise prior: $\lambda = 0.00, \gamma = 0.00$

Figure 4. The appearance noise prior enables the model to produce reasonable videos when the inference resolution (512x512) differs from the resolution (320x320) used during training.



With appearance noise prior: $\lambda = 0.03, \gamma = 0.02$

Without appearance noise prior: $\lambda = 0.00, \gamma = 0.00$

Figure 5. The appearance noise prior enables the model to produce reasonable videos with sample steps of 5, especially for simple motion. This efficiency allows a video to be created in just 1.3 seconds.

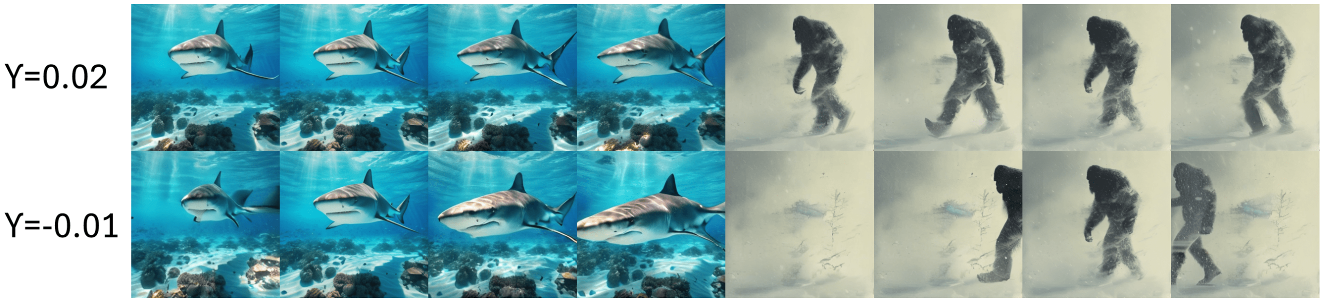
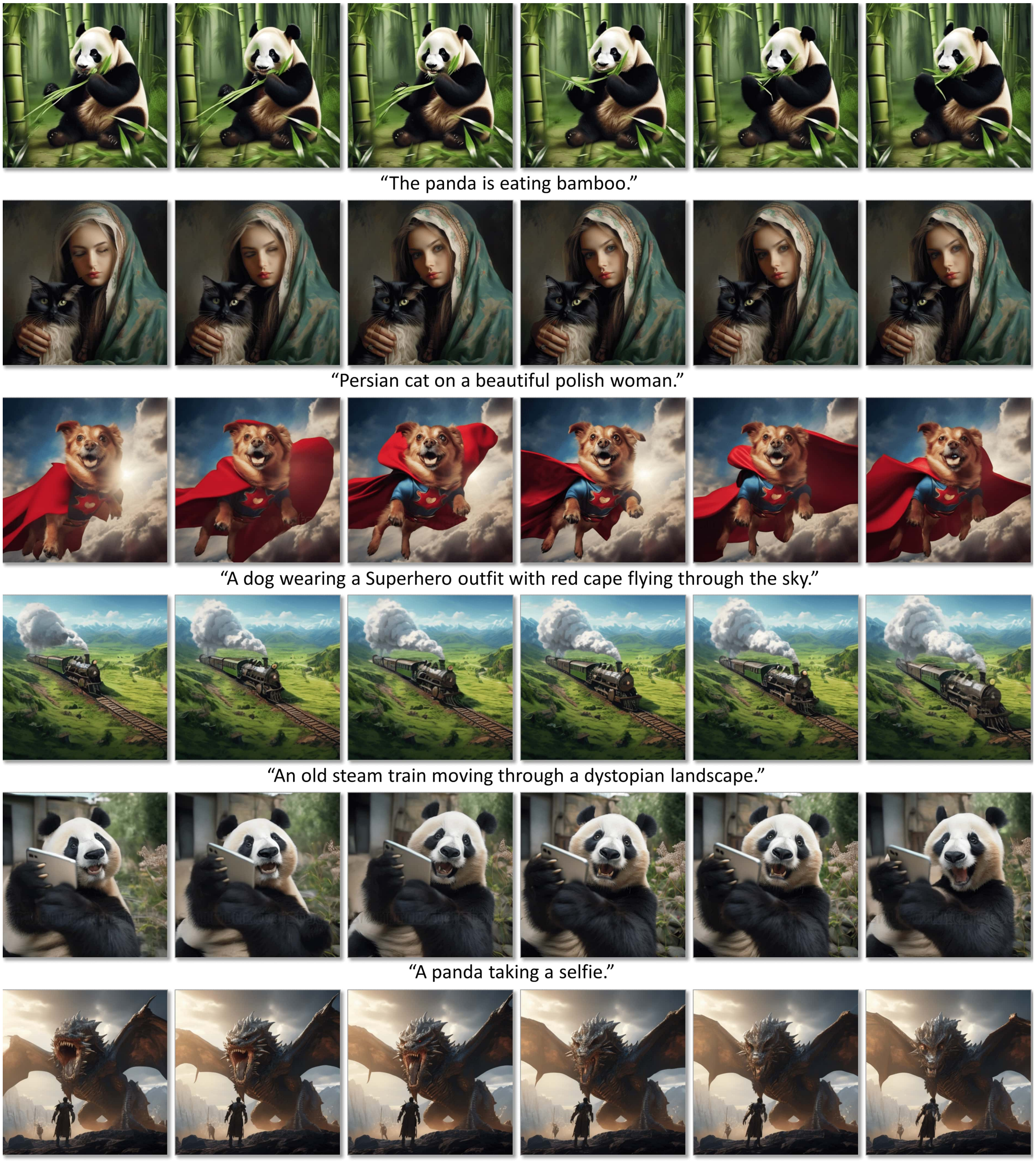


Figure 6. Larger γ (0.02) yields smaller motion and vice versa. Smaller γ (-0.01) produces larger motion.



“The panda is eating bamboo.”

“Persian cat on a beautiful polish woman.”

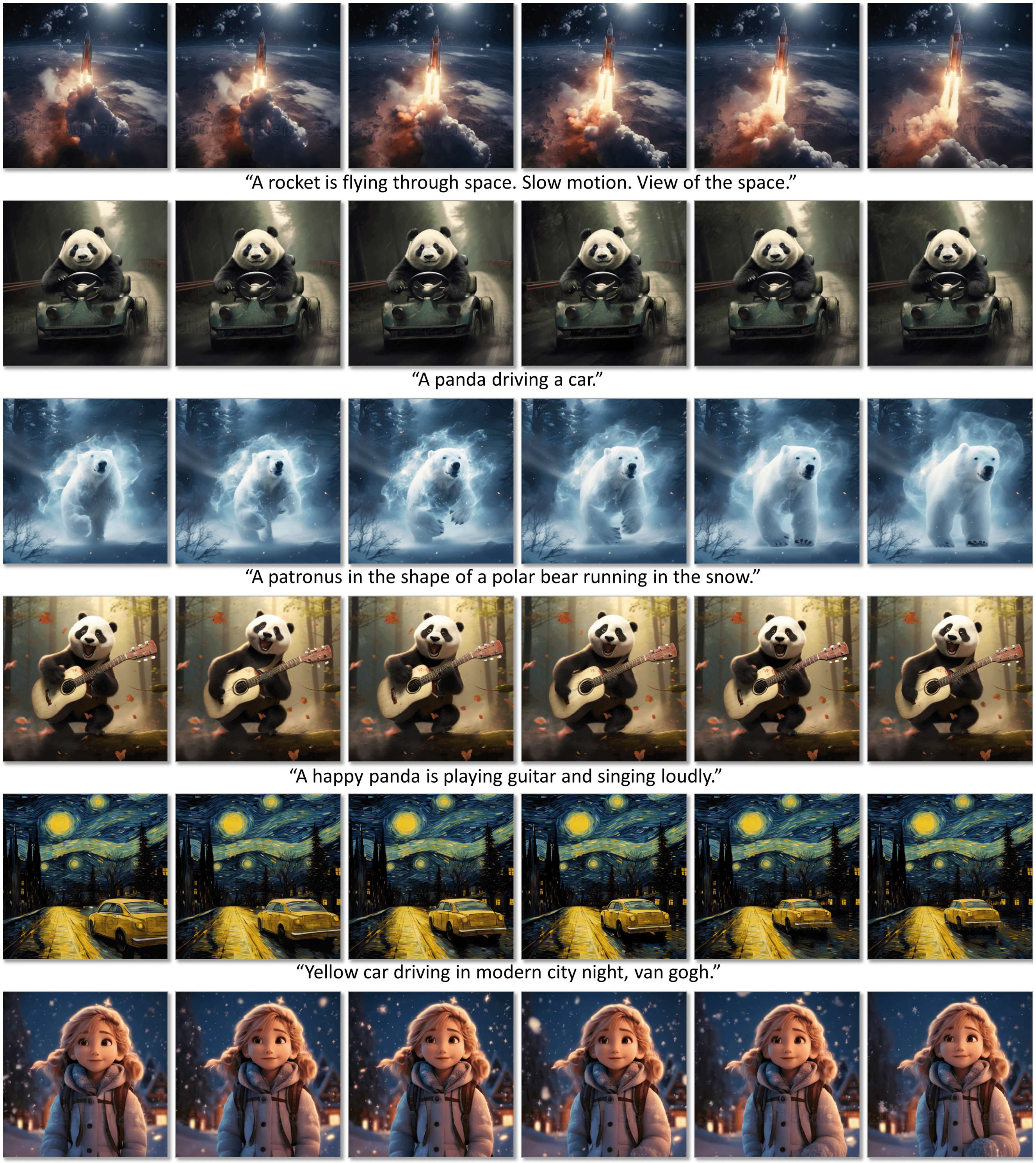
“A dog wearing a Superhero outfit with red cape flying through the sky.”

“An old steam train moving through a dystopian landscape.”

“A panda taking a selfie.”

“Big realistic dragons dark dragons in a huge natural landscape, cinematic, mists, huge and scaly, octane render, goldheart warrior in armor with cloak in foreground, sideview. flying creatures.”

Figure 7. Examples from MicroCinema demonstrate that our model is capable of generating exquisite imagery and crisp motion. Benefiting from a divide-and-conquer strategy, the model, though trained on the WebVid-10M dataset, can leverage given images to produce videos in various styles.



"A rocket is flying through space. Slow motion. View of the space."

"A panda driving a car."

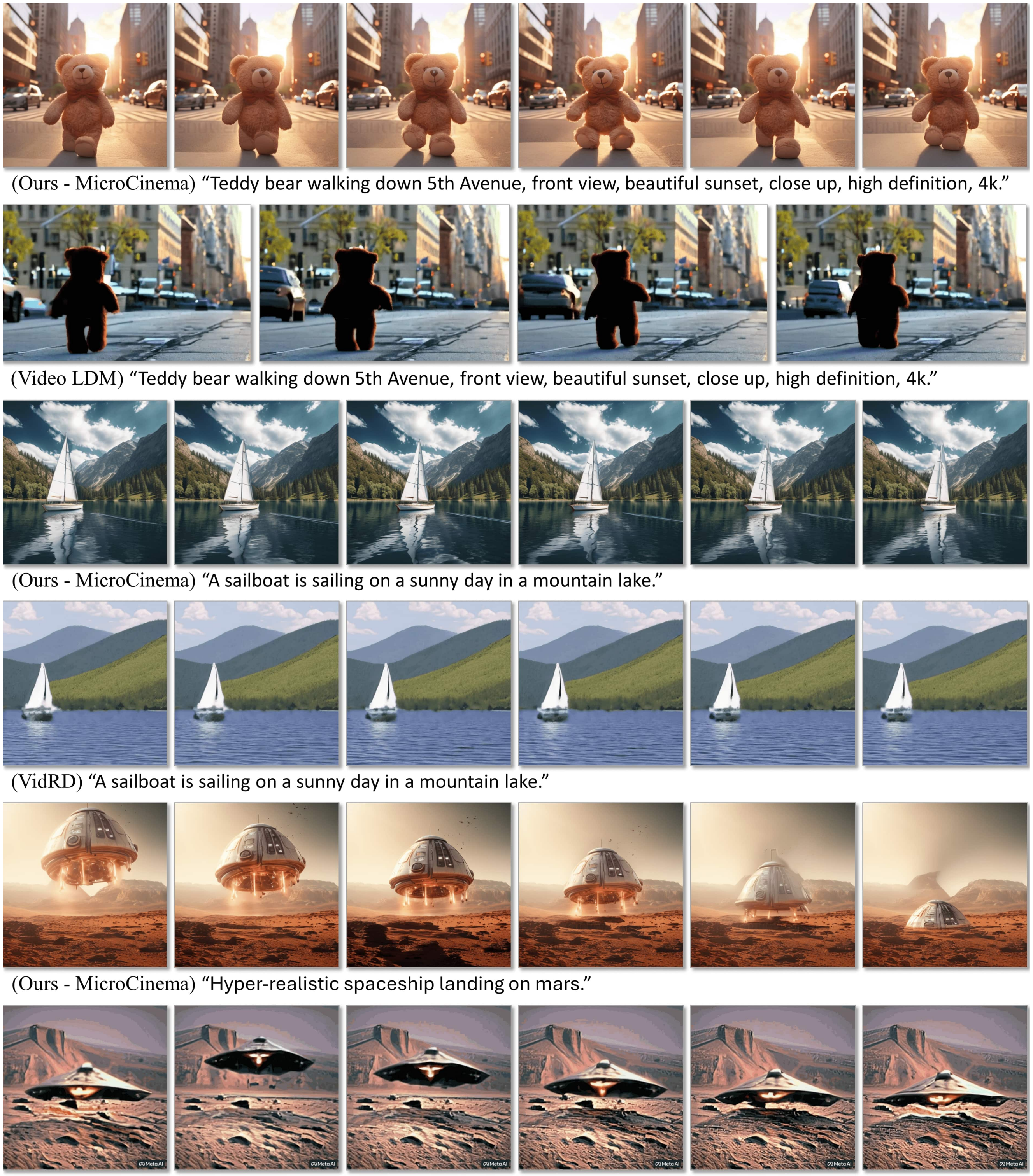
"A patronus in the shape of a polar bear running in the snow."

"A happy panda is playing guitar and singing loudly."

"Yellow car driving in modern city night, van gogh."

"Disney animation style, one frosty day, when snow blanketed everything like a white quilt, a little girl named Zosia was coming home from school. With gloves keeping her hands warm and a cozy jacket, she walked along the path."

Figure 8. More examples of MicroCinema.



(Ours - MicroCinema) “Teddy bear walking down 5th Avenue, front view, beautiful sunset, close up, high definition, 4k.”

(Video LDM) “Teddy bear walking down 5th Avenue, front view, beautiful sunset, close up, high definition, 4k.”

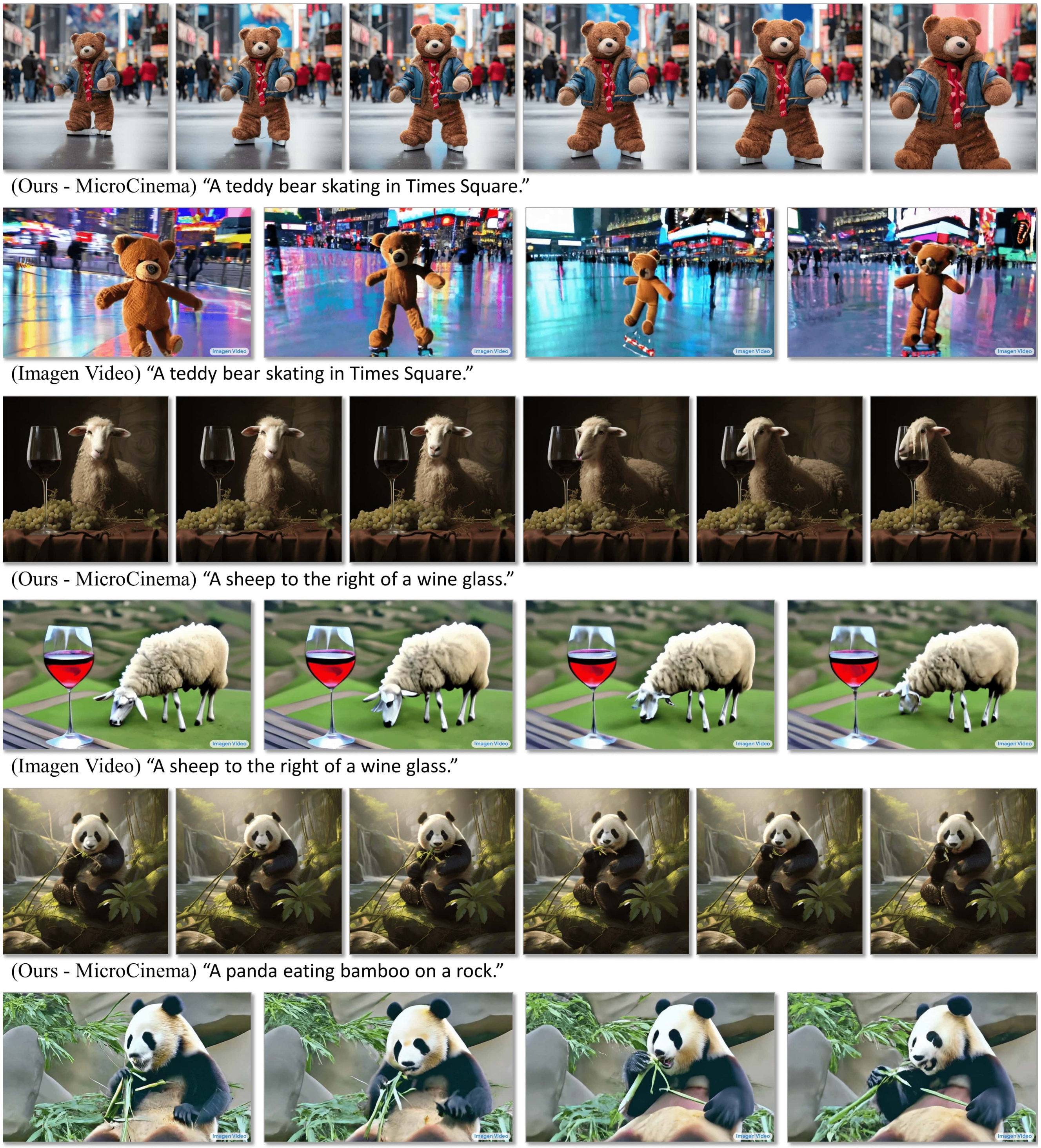
(Ours - MicroCinema) “A sailboat is sailing on a sunny day in a mountain lake.”

(VidRD) “A sailboat is sailing on a sunny day in a mountain lake.”

(Ours - MicroCinema) “Hyper-realistic spaceship landing on mars.”

(Make-A-Video) “Hyper-realistic spaceship landing on mars.”

Figure 9. We compare our method, MicroCinema, with Video LDM, VidRD, and Make-A-Video. Our videos exhibit clearer imagery and more distinct motion compared to Video LDM. In relation to VidRD, our image quality is significantly superior. When compared with Make-A-Video, our method demonstrates better image quality and text consistency.



(Ours - MicroCinema) “A teddy bear skating in Times Square.”

(Imagen Video) “A teddy bear skating in Times Square.”

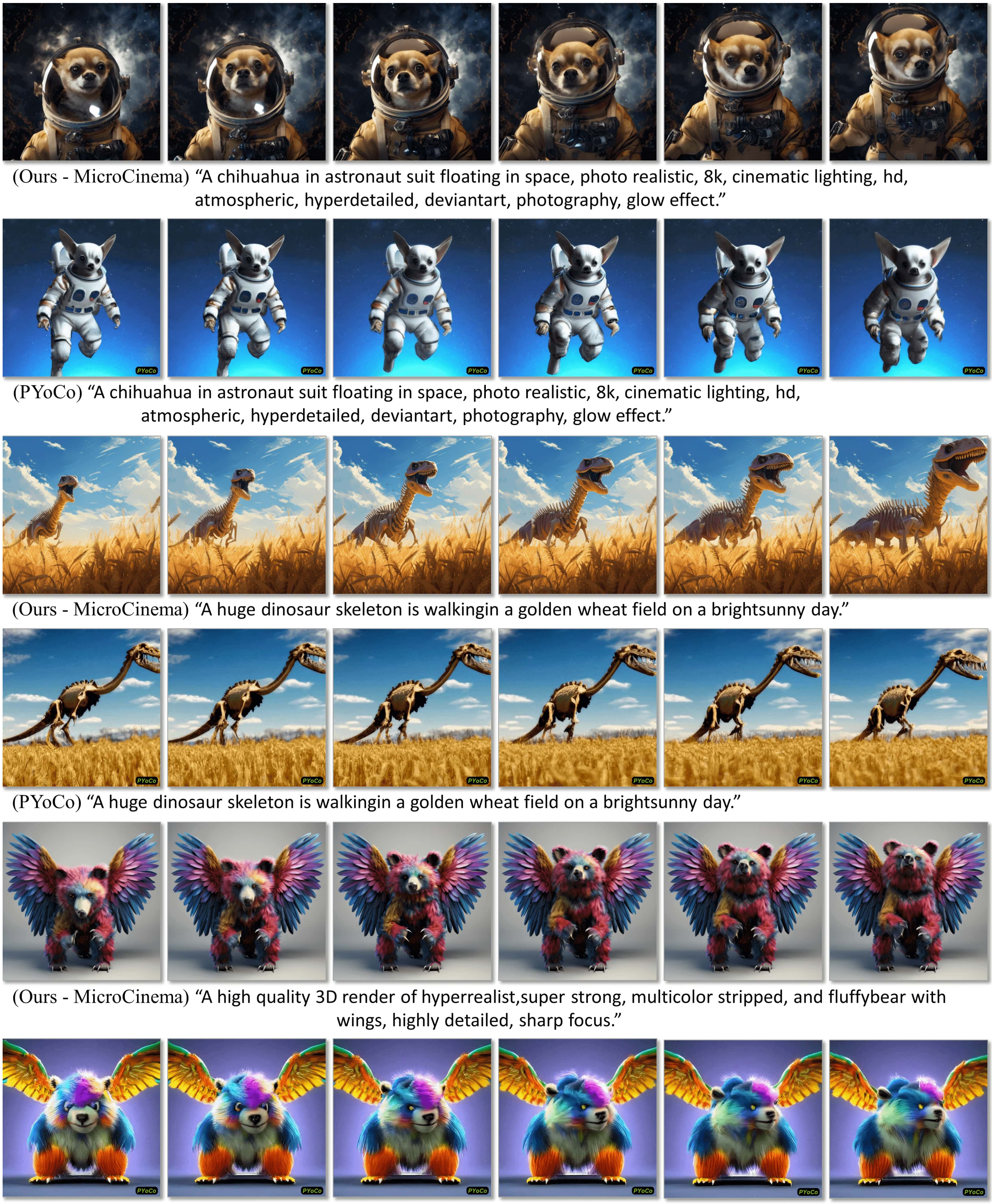
(Ours - MicroCinema) “A sheep to the right of a wine glass.”

(Imagen Video) “A sheep to the right of a wine glass.”

(Ours - MicroCinema) “A panda eating bamboo on a rock.”

(Imagen Video) “A panda eating bamboo on a rock.”

Figure 10. We compare our method, MicroCinema, with Imagen Video. Our approach demonstrates superior temporal consistency, as evidenced (first row), compared to the Imagen Video method (second row). Furthermore, our method exhibits more exquisite image details, highlighted (third and fifth rows), while the videos produced by Imagen Video lack such detail (fourth and sixth rows).



(Ours - MicroCinema) "A chihuahua in astronaut suit floating in space, photo realistic, 8k, cinematic lighting, hd, atmospheric, hyperdetailed, deviantart, photography, glow effect."

(PYoCo) "A chihuahua in astronaut suit floating in space, photo realistic, 8k, cinematic lighting, hd, atmospheric, hyperdetailed, deviantart, photography, glow effect."

(Ours - MicroCinema) "A huge dinosaur skeleton is walkingin a golden wheat field on a brightsunny day."

(PYoCo) "A huge dinosaur skeleton is walkingin a golden wheat field on a brightsunny day."

(Ours - MicroCinema) "A high quality 3D render of hyperrealist,super strong, multicolor stripped, and fluffybear with wings, highly detailed, sharp focus."

(PYoCo) "A high quality 3D render of hyperrealist,super strong, multicolor stripped, and fluffybear with wings, highly detailed, sharp focus."

Figure 11. We compare our method, MicroCinema, with PYoCo. Relative to PYoCo, MicroCinema excels in generating more intricate and visually appealing videos from complex descriptions. Our method showcases more pronounced motion and finer image details.