Adversarial Distribution Matching for Diffusion Distillation Towards Efficient Image and Video Synthesis

Supplementary Material

A. Adversarial Distribution Matching

During the ADM distillation process, the fake score estimator, generator, and discriminator are updated alternately. The Algorithm 1 below clarifies the training procedure. Our ablation experiments in Sec. 5.3 demonstrate that TTUR has minimal impact on the final performance. Therefore, in our experiments, we set TTUR to 1, meaning that the fake model and generator are updated at the same frequency.

Algorithm 1 ADM Training Procedure

```
**Input:** pretrained teacher model as real score estimator $\boldsymbol{F}_\phi$
**Output:** few-step generator $\boldsymbol{G}_\theta$ with schedule $\{t_0, t_1, ..., t_N\}$
**Initialize:** fake score estimator $\boldsymbol{f}_\psi \leftarrow \boldsymbol{F}_\phi$, generator $\boldsymbol{G}_\theta \leftarrow \boldsymbol{F}_\phi$,
  latent-space discriminator $\boldsymbol{D}_\tau \leftarrow \boldsymbol{F}_\phi$ with multiple trainable heads,
  generator iteration $genIter \leftarrow 0$, global iteration $globalIter \leftarrow 0$
**while** $genIter < maxIter$ **do**
    $globalIter \mathrel{+}= 1$

    $//\ update\ fake\ score\ estimator\ \boldsymbol{f}_\psi$
    sample pure noise $\boldsymbol{z} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$
    solve the PF-ODE w.r.t. $N$ steps in schedule $\boldsymbol{x}_0 \leftarrow \boldsymbol{G}_\theta(\boldsymbol{z}, \cdot)$
    sample new pure noise $\boldsymbol{z}_f$ and random timestep $t_f$
    update $\psi$ with $(\boldsymbol{x}_0, t_f, \boldsymbol{z}_f)$ and pretrain loss in Eq. (2) or Eq. (3)
    **if not** $(globalIter\ \%\ \text{TTUR}) == 0$ **then continue**

    $//\ update\ generator\ \boldsymbol{G}_\theta$
    sample pure noise $\hat{\boldsymbol{z}}$ and random index $n \in [1, N]$
    solve the PF-ODE w/o grad following $t_N \rightarrow t_{N-1} \rightarrow ... \rightarrow t_n$,
  i.e. $\hat{\boldsymbol{z}} = \hat{\boldsymbol{x}}_{t_N} \rightarrow \hat{\boldsymbol{x}}_{t_{N-1}} \rightarrow ... \rightarrow \hat{\boldsymbol{x}}_{t_n}$
    solve the PF-ODE w/ grad w.r.t. $t_0$, i.e. $\hat{\boldsymbol{x}}_0 = \boldsymbol{G}_\theta(\hat{\boldsymbol{x}}_{t_n}, t_n)$
    sample new pure noise $\boldsymbol{z}_g$ and random timestep $t \sim \mathcal{U}(0, T)$
    diffuse sample $\hat{\boldsymbol{x}}_0$ with $\boldsymbol{z}_g$ and Eq. (1), i.e. $\boldsymbol{x}_t = \boldsymbol{q}(\boldsymbol{x}_t | \hat{\boldsymbol{x}}_0)$
    solve the PF-ODE of $\boldsymbol{f}_\psi$ w.r.t. $(t - \Delta t)$ to obtain $\boldsymbol{x}^{\text{fake}}_{t-\Delta t}$
    solve the PF-ODE of $\boldsymbol{F}_\phi$ w.r.t. $(t - \Delta t)$ to obtain $\boldsymbol{x}^{\text{real}}_{t-\Delta t}$
    update $\theta$ with $(\boldsymbol{x}^{\text{fake}}_{t-\Delta t}, t - \Delta t)$ and Eq. (7)
    $genIter \mathrel{+}= 1$

    $//\ update\ discriminator\ \boldsymbol{D}_\tau$
    update $\tau$ with $(\boldsymbol{x}^{\text{fake}}_{t-\Delta t}, \boldsymbol{x}^{\text{real}}_{t-\Delta t}, t - \Delta t)$ and Eq. (8)
**end while**
```

B. Implementation Details

B.1. 2D Discriminator Design

In Fig. 6, we thoroughly illustrate the design of our discriminators and the difference between two training stages. For all the trainable heads appended to discriminator backbone for text-to-image experiments, we have a fixed 2D design following SDXL-Lightning [23], which consists of simple blocks of 4×4 2D convolution with a stride of 2, group normalization [76] with 32 groups, and SiLU activation [10, 54] layer. The difference is that we will append multiple heads at different layers of the network. Whether it is the output of UNet [57], DiT [49] or ViT [5], we uniformly reshape it into $[\boldsymbol{Batch}, \boldsymbol{Channel}, \boldsymbol{Height}, \boldsymbol{Width}]$ and then use it as the input to the discriminator head. For SDXL [56], we take the output of the last ResNet [9] of each block (including down-sampling, mid and up-sampling blocks), yielding a total of 7 discriminator heads. For SD3 series [6] models, we take the output of each DiT block, yielding 24 and 38 discriminator heads for SD3-Medium and SD3.5-Large, respectively. For SAM [19] and DINOv2 [48], we take the output of layers 3, 6, 9 and 12, yielding 4 discriminator heads.

	Training Iteration	GPU Number	Elapsed Time	GPU Hours	Micro BatchSize	Max Memory
DMD2	20K	64	60 hours	3840	2	-
DMDX	**8K+8K**	**32**	**70 hours**	**2240**	**4**	**39.6 GiB**
- ADP	8K	32	55 hours	1760	4	39.6 GiB
- ADM	8K	32	15 hours	480	4	24.1 GiB

Table 7. Comparisons on A100 GPU efficiency with DMD2. The elapsed time for ADP already includes collection of ODE pairs.

B.2. 3D Discriminator Design

Our 3D discriminator head for text-to-video latent diffusion models consists of simple blocks of 3×3×3 3D convolution with a stride of 1, 3×3 2D convolution with a stride of 2, group normalization with 32 groups and SiLU activation layer. This is similar to the design in 2D discriminator head except that we additionally insert several 3D convolution layers to extract time-dependent feature. The output of specific blocks within video DiT backbone are reshaped into $[\boldsymbol{Batch}, \boldsymbol{Channel}, \boldsymbol{Time}, \boldsymbol{Height}, \boldsymbol{Width}]$ and input to corresponding discriminator head. In practice, we extract features every 3 DiT blocks due to the computational effort of 3D convolution, yielding a total of 10 and 14 discrimiantor heads for 2B and 5B models, respectively.

B.3. GPU efficiency.

In Tab. 7, we present the training configurations and GPU consumption of our proposed method compared to DMD2. The table demonstrates that we actually achieve better performance over DMD2 with less GPU time and don't impose excessive demands on GPU memory. Although maintaining more networks during training process, our implementation attains manageable memory footprint with several optimizations detailed later.

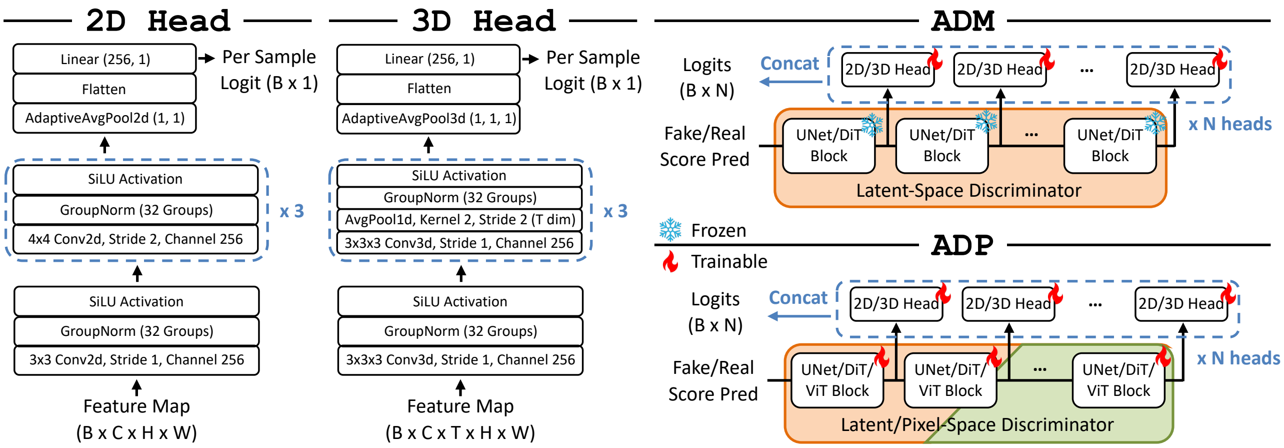


Figure 6. Illustration of our discriminator design and the difference between ADM and ADP.

B.4. Memory efficiency.

To reduce GPU memory footprint and improve efficiency, we utilize several acceleration techniques in our implementation including Fully Sharded Data Parallel (FSDP) [53], gradient checkpoint [4] and BF16 mixed precision [72]. For text-to-video models, we additionally integrate, Context Parallel (CP) [83] and Sequence Parallel (SP) [15] following common practice in MovieGen[46] to speed up training and inference. More importantly, a CPU offloading technique that has been built into Pytorch FSDP is essential for training multiple networks to save memory.

With CPU offloading enabled, each parameter along with the corresponding gradient and optimizer state can be offloaded from the GPU to CPU memory. In conjunction with gradient checkpointing, the GPU memory footprint in the forward and backward process is nearly the same as when there is only one single network, because the peak memory is now determined by the maximum activation of each block. This comes at the cost of increased CPU memory and longer time per iteration. While the CPU memory is usually sufficient and cheap, our more effective approaches require fewer iterations to achieve convergence and satisfactory results, and as Tab. 7 show that our DMDX takes less time than DMD2 on one-step SDXL distillation.

B.5. Hyperparameters.

For all models of the optimizer (including generator, fake model and discriminator in both text-to-image and text-to-video experiments), we use AdamW [29] optimizer without weight decay, with beta parameters (0.0, 0.99) to capture the changes in distribution more up-to-date. The learning rates of discriminator and fake model across all of our experiments are fixed at 5e-6 and 1e-6, respectively.

For SDXL, the learning rates for generator during ADP and ADM training are 1e-6 and 1e-7, respectively. As for multi-step ADM distillation, the learning rates for generator of SD3-Medium LoRA training and SD3.5-Large fully fine-tuning are given to 1e-6 and 1e-8, respectively. In case of text-to-video diffusion distillation, we set the same learning rate 1e-7 for different 8-step CogVideoX generators.

Among all the ADM experiments, the Classifier-Free Guidance (CFG) is required for real model as DMD does [85]. For SDXL, SD3-Medium, SD3.5-Large, and CogVideoX, the uniformly random sampling ranges for the CFG values are set to [6.0, 8.0], [6.0, 8.0], [3.0, 4.0], and [5.0, 7.0], respectively. The chosen ranges are based on the recommended CFG values from the original baseline's inference with some allowable variations. We observed that this setting is adequate for achieving satisfactory distilled performance without requiring extensive tuning.

The fake model training does not incorporate CFG and uses the same loss function as the standard pre-training of diffusion models, except that we didn't set any dropout. For noise-parameterized models, the prediction target is noise, while for velocity-parameterized models, it is velocity.

C. Main Results

C.1. Efficient Image Synthesis

Fig. 7 qualitatively compares our method with other state-of-the-art distillation techniques on SD3 [6] series models. The results demonstrate that our method is competitive to the original model in terms of color, detail, structure and image-text alignment, while outperforming other methods including TSCD, PCM [69], Flash [3] and LADD [60].

C.2. Efficient Video Synthesis

Tabs. 8 and 9 present the details of VBench [14] results on the base model and few-step generators of CogVideoX [83]. In Figs. 11 to 16, we present several cases for qualitative comparisons between our CogVideoX [83] generators and baseline model. The results show that our 8-step genera-

Method	**Step**	**NFE**	**Final Score**	**Quality Score**	**Semantic Score**	Subject Consistency	Background Consistency	Temporal Flickering	Motion Smoothness	Dynamic Degree	Aesthetic Quality	Imaging Quality
ADM	8	8	78.58	80.82	69.62	96.72	96.55	97.01	98.14	48.61	57.80	65.28
+Longer Training ×2	8	8	80.76	**83.03**	71.69	96.58	96.71	98.12	97.68	73.33	57.90	65.72
ADM w/ CFG	8	16	79.86	80.93	75.56	96.16	96.96	96.86	97.69	54.44	59.78	63.18
+Longer Training ×2	8	16	**81.79**	83.00	**76.94**	96.83	96.90	98.51	98.07	63.05	61.03	64.62
CogVideoX-2b	100	200	80.03	80.80	76.97	92.53	95.22	97.79	97.00	69.44	60.38	60.69
ADM	8	8	**82.06**	**83.22**	**77.42**	96.42	96.87	96.96	97.69	68.88	61.17	69.01
ADM w/ CFG	8	16	80.98	82.16	76.25	96.15	96.59	95.99	98.57	56.66	61.01	68.68
CogVideoX-5b	100	200	81.22	81.78	78.98	92.52	96.68	98.34	96.97	70.55	61.67	61.88

Table 8. VBench [14] detailed results on **overall scores** and separate score for each quality dimension.

Method	**Step**	**NFE**	Object Class	Multiple Objects	Human Action	Color	Spatial Relationship	Scene	Appearance Style	Temporal Style	Overall Consistency
ADM	8	8	83.97	47.19	87.40	77.79	62.93	42.64	24.16	22.35	25.27
+Longer Training ×2	8	8	87.84	56.53	85.00	80.28	69.52	44.33	23.15	22.60	25.11
ADM w/ CFG	8	16	89.55	64.78	92.60	82.31	62.61	52.73	24.31	24.46	26.12
+Longer Training ×2	8	16	91.67	71.58	92.20	82.01	71.79	50.26	23.54	24.54	26.30
CogVideoX-2b	100	200	80.01	67.23	98.60	89.98	49.05	68.60	24.04	25.37	25.68
ADM	8	8	92.94	65.89	95.80	84.97	72.92	56.06	22.63	23.64	26.17
ADM w/ CFG	8	16	89.41	69.89	97.00	71.35	81.26	53.90	21.48	23.79	25.92
CogVideoX-5b	100	200	87.64	67.34	99.60	83.93	68.24	56.35	25.16	25.82	27.79

Table 9. VBench [14] detailed results on separate score for each semantic dimension.

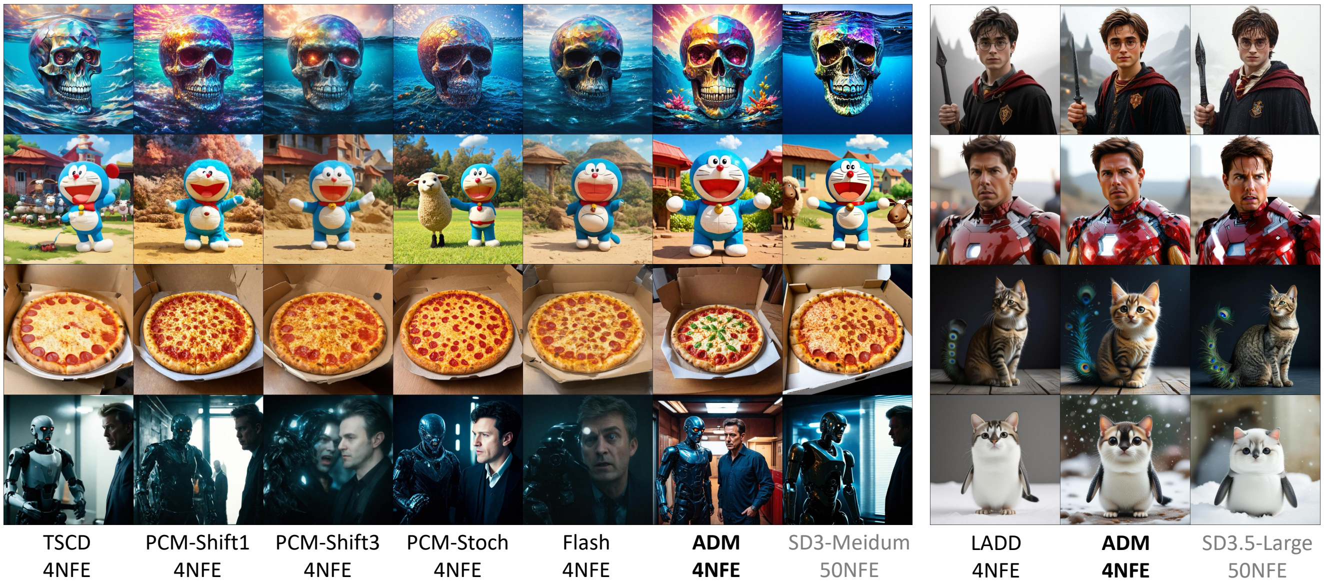


Figure 7. Qualitative results on LoRA fine-tuning SD3-Medium and fully tine-tuning SD3.5-Large.

tors are generally semantically comparable to the original model, even with semantic enhancements on some cases, e.g., the change of light in Fig. 11 and the movement of the sheep in Fig. 14. While in terms of imaging quality, generators with CFG are generally more detailed and have more delicate textures than those without CFG. The deficiencies in detail are reflected in, for example, the slightly rough hand and the incorrect number of fingers in Fig. 15, whereas the one with CFG is much more natural. As well as the generator without CFG is also much higher in color contrast, which visually looks sometimes too vibrant to be sufficiently realistic. These demonstrate the importance of CFG for text-to-video models, which might not be fully reflected by quantitative metrics.

C.3. Ablation Studies

As for ablation on adversarial distillation shown in Fig. 8, the two main problems with other baseline settings are

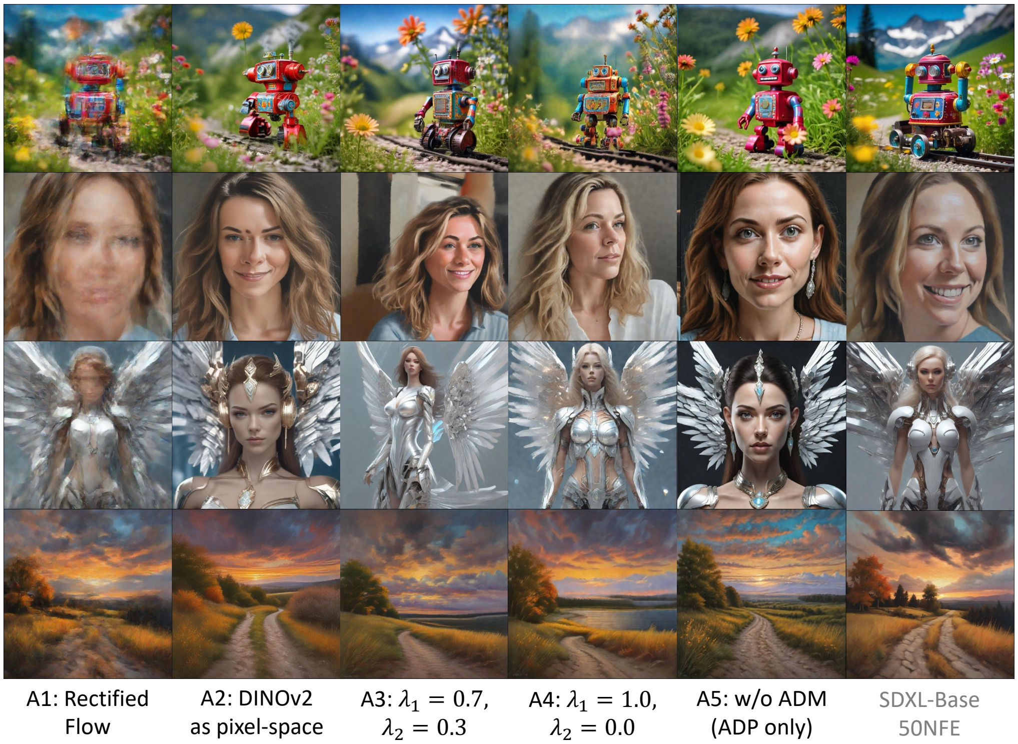

A1: Rectified Flow
A2: DINOv2 as pixel-space
A3: $\lambda_1 = 0.7$, $\lambda_2 = 0.3$
A4: $\lambda_1 = 1.0$, $\lambda_2 = 0.0$
A5: w/o ADM (ADP only)
SDXL-Base 50NFE

Figure 8. Qualitative comparisons for ablation studies on adversarial distillation.

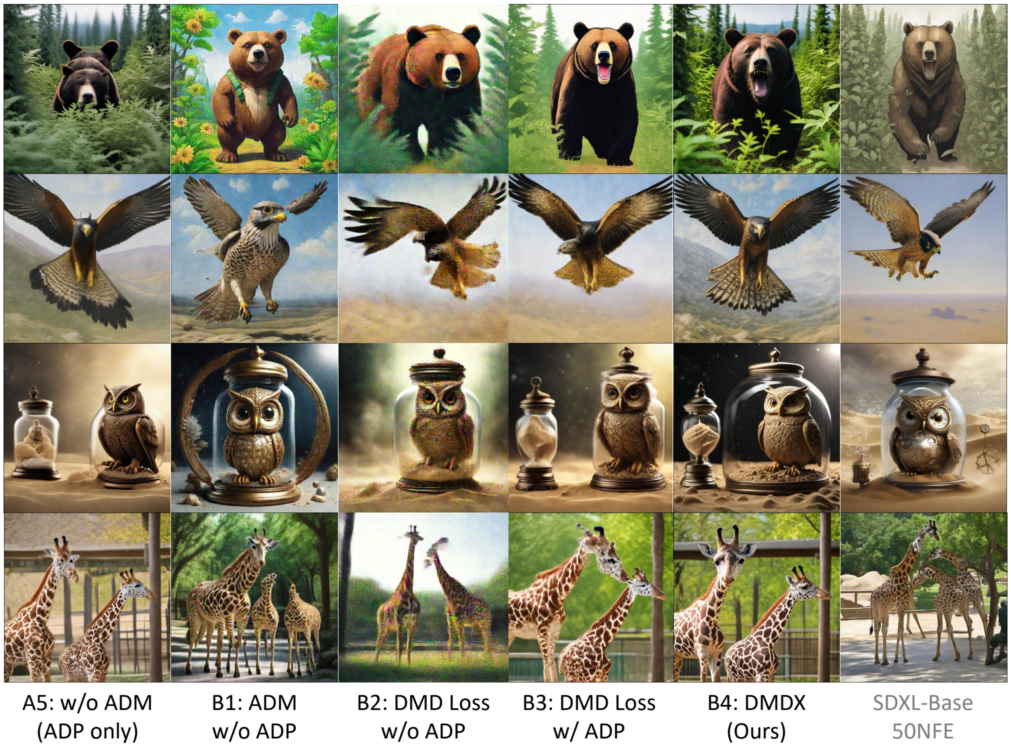
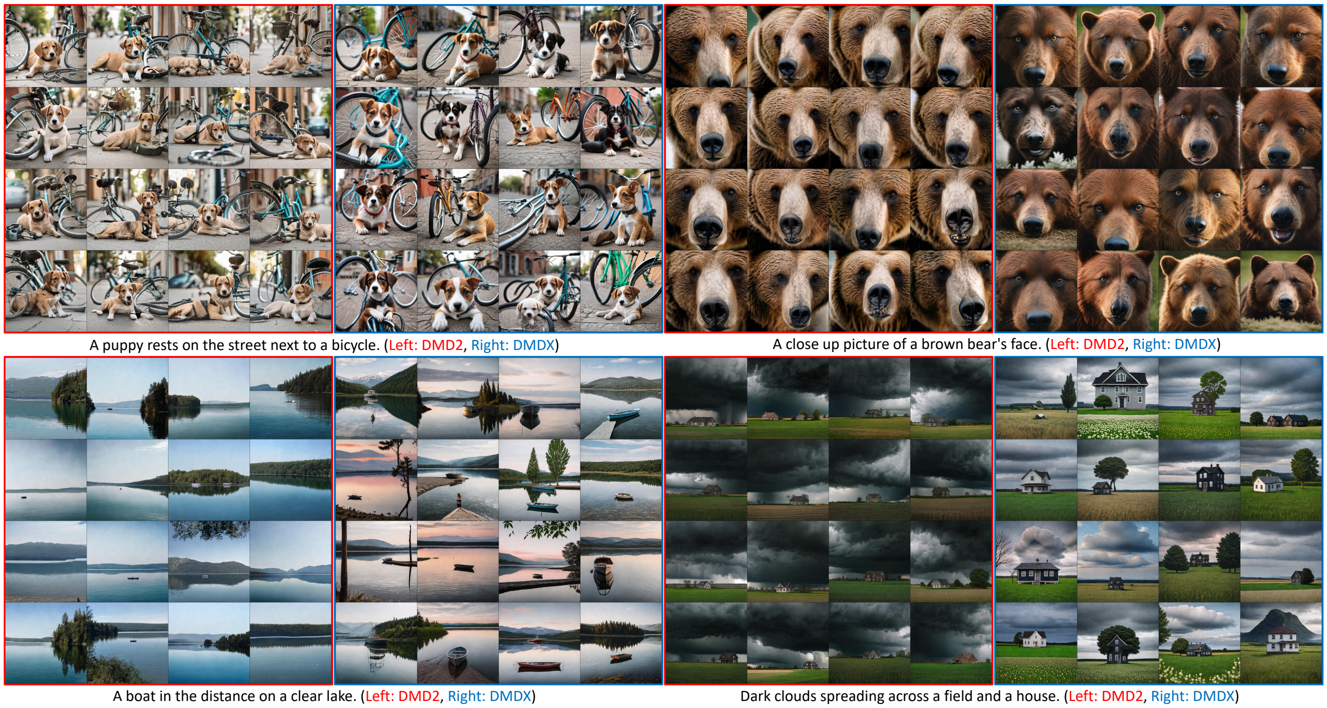


Figure 9. Qualitative comparisons for ablation studies on score distillation.


A puppy rests on the street next to a bicycle. (Left: DMD2, Right: DMDX)

A close up picture of a brown bear's face. (Left: DMD2, Right: DMDX)

A boat in the distance on a clear lake. (Left: DMD2, Right: DMDX)

Dark clouds spreading across a field and a house. (Left: DMD2, Right: DMDX)

Figure 10. Qualitative diversity comparisons with DMD2.

structure and blurriness. When using MSE loss for a single reflow process as in Rectified Flow [27], it is obvious that it is struggling to generate a structurally visible image. And switching the SAM [19] model to DINOv2 [48], we can clearly see the structural collapse of both the robot and the face in the figure, which is unexpected and may be caused by the fact that its input resolution is only 518px, and the images we generate are all 1024px need to be resized before they can be input. Another possible explanation is that the prior knowledge used by SAM for instance segmentation is richer than that provided by DINOv2 for discriminative self-supervised learning, which facilitates the generation of local fine-grained details. The structural problems encountered when increasing the weight of pixel-space λ_2 are similar, while decreasing its weight causes a very noticeable blurring that is clearly visible in the figure, so we suggest setting $\lambda_1 = 0.85, \lambda_2 = 0.15$ is a reasonable configuration.

In Fig. 9, we provide qualitative comparisons for ablation studies on score distillation. Compared to the baseline without ADM (ADP only), we can see that the ADM distillation indeed serves as a fine-tuning process to refine the generator in terms of both color, detail and the most notable structure. Although standalone ADM can also produce efficient generator, the noise artifact within 1-step generations as similarly observed by [23, 85] still exists, and with our ADP this issue can be addressed well. Notably, the visualization results demonstrate that employing the DMD loss without ADP integration induces substantially severe noise artifacts. Compared to using ADM alone, its qualitative disadvantage is much more pronounced than the gap observed in the quantitative results. With ADP, the DMD loss generates relatively good results, yet it remains inferior to ADM in terms of visual fidelity and structural integrity. This indicates that its distribution matching capability is weaker than that of ADM, which is consistent with our analysis in the quantitative results of Sec. 5.3.

Additionally, we showcase additional randomly curated multi-seed samples in Fig. 10 compared with DMD2, clearly demonstrating that our images exhibit richer variations in texture, color, brightness, contrast and structural composition.

D. Broader Impact

Considering that many current methods leverage generated data from foundation models as assistance [44], our acceleration approach for diffusion models can substantially expedite this process, thereby benefiting numerous downstream tasks such as recognition [77], detection [42], retrieval [31, 41], domain adaptation [32, 62], etc. Alternatively, we can train LoRA to acquire an acceleration plugin, enhancing the efficiency of customized vertical models for image [33] or video [80] generation.

E. Prompt List

Below we list the text prompts used for the generated content shown in this paper (from top to bottom, from left to right). Note that since models like SDXL-Base [56] only use CLIP [52] as a text encoder, which only supports a maximum of 77 tokens, the response and text-image alignment may be insufficient for some long prompts and its limited capacity in understanding.

We use the following prompts for Fig. 5:

- A beautiful woman facing to the camera, smiling confidently, colorful long hair, diamond necklace, deep red lip, medium shot, highly detailed, realistic, masterpiece.
- An owl perches quietly on a twisted branch deep within an ancient forest. Its sharp yellow eyes are keen and watchful.
- A young badger delicately sniffing a yellow rose, with a lion lurking in the background.
- A pickup truck going up a mountain switchback.

We use the following prompts for Fig. 7:

- A photograph of a giant diamond skull in the ocean, featuring vibrant colors and detailed textures.
- A still of Doraemon from ”Shaun the Sheep” by Aardman Animation.
- A pizza is displayed inside a pizza box.
- movie still of a man and a robot in a moment of horror, movie still, cinematic composition, cinematic light, by edgar wright and david lynch
- harry potter as a skyrim character
- film still of Tom Cruise as Ironman in the Avengers
- A beautiful award winning picture of a cute cat in front of a dark background. The cat is a cat-peacock hybrid and has a peacock tail and short peacock feathers on the body. fluffy, extremely detailed, stunning, high quality, atmospheric lighting
- a cute animal that's a penguin cat hybrid

We use the following prompts for Fig. 8:

- A colorful tin toy robot runs a steam engine on a path near a beautiful flower meadow in the Swiss Alps with a mountain panorama in the background, captured in a long shot with motion blur and depth of field.
- A portrait painting of Leighann Vail.
- A photo of a mechanical angel woman with crystal wings, in the sci-fi style of Stefan Kostic, created by Stanley Lau and Artgerm.
- A painting depicting a foothpath at Indian summer with an epic evening sky at sunset and low thunder clouds.

We use the following prompts for Fig. 9:

- A bear walks through a group of bushes with a plant in its mouth.
- A falcon in flight, depicted in a highly detailed painting by Ilya Repin, Phil Hale, and Kent Williams.
- A steampunk pocketwatch owl is trapped inside a glass jar buried in sand, surrounded by an hourglass and swirling mist.
- Some giraffes are walking around the zoo exhibit.

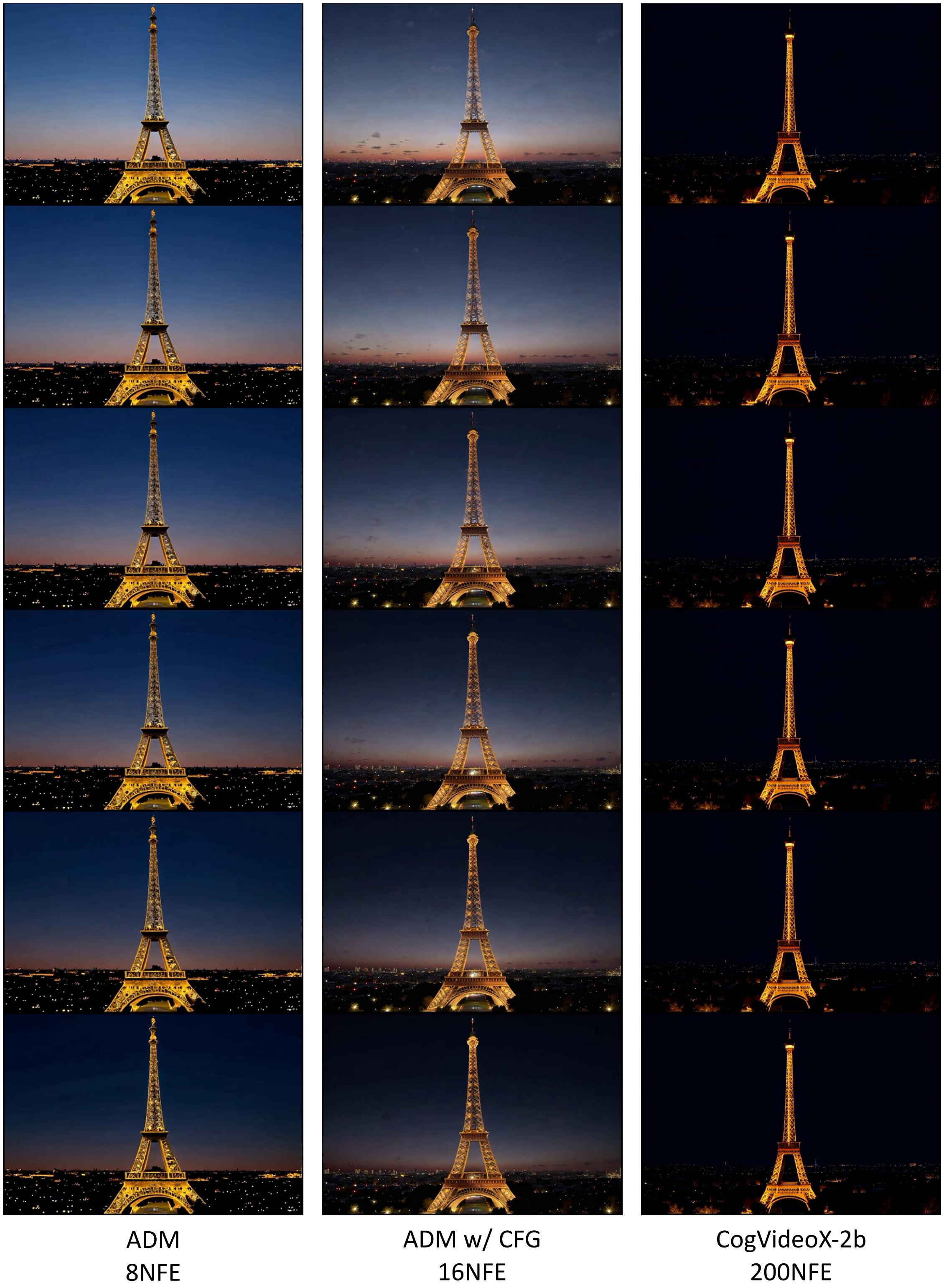


Figure 11. Qualitative comparisons on CogVideoX-2b generators. The random seed has been fixed. **Prompt:** A time-lapse sequence captures the transformation of the iconic Eiffel Tower fromdaylight into the evening. The tower, standing tall and majestic in its originalgolden hue, gradually transitions into a silhouette against the twilight sky. Asthe sun sets, the city lights begin to flicker on, casting a warm glow over theParisian landscape. The tower's intricate iron lattice structure becomes more defined,its shadow lengthening across the Champ de Mars. The background includes the SeineRiver and the Parisian rooftops, adding depth and context to the scene. As darknessfalls, the Eiffel Tower is illuminated by its own lights, turning into a beaconof Paris, shimmering against the starry backdrop.

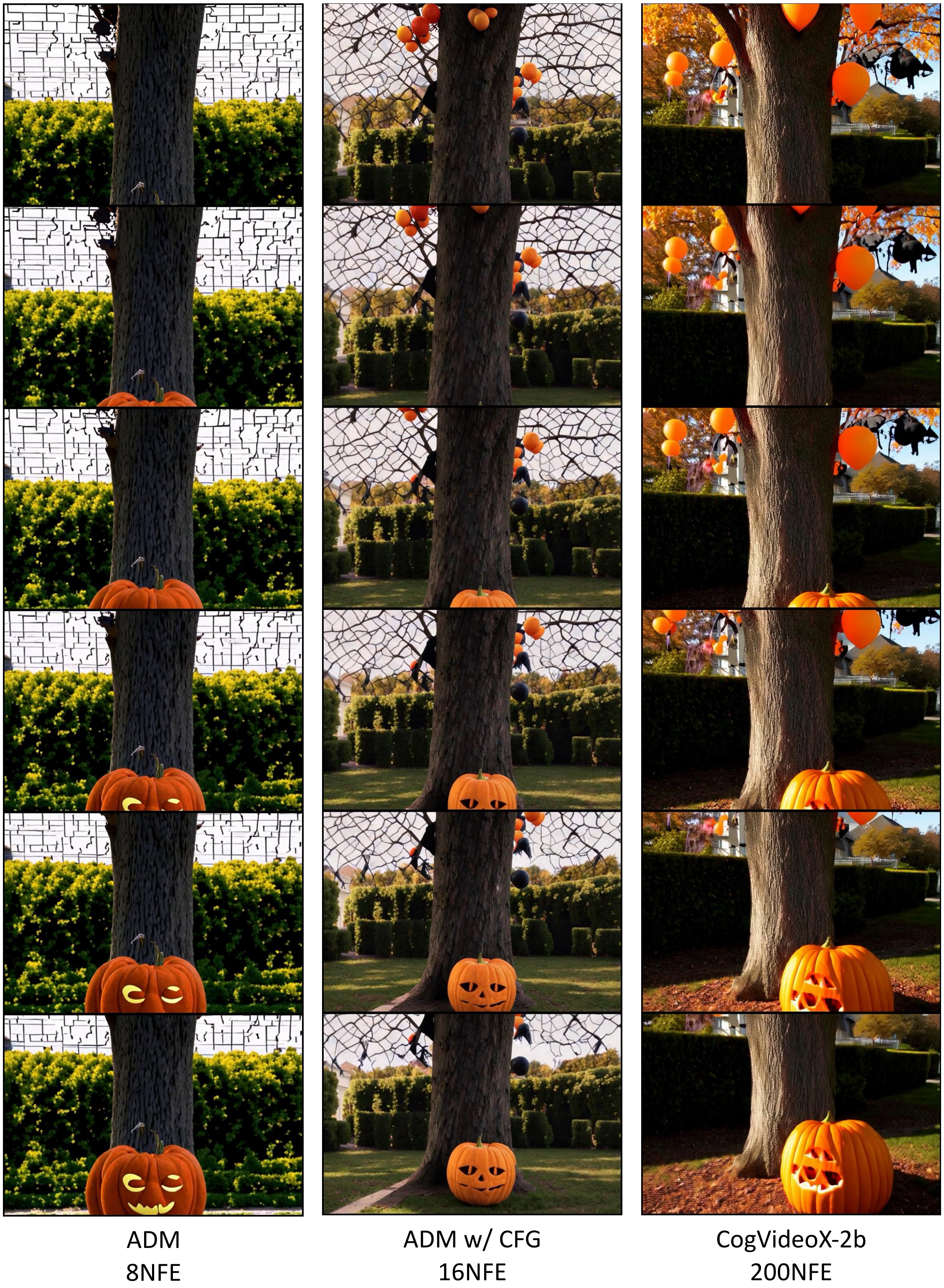


Figure 12. Qualitative comparisons on CogVideoX-2b generators. The random seed has been fixed. **Prompt:** A vibrant oak tree, adorned with festive Halloween decorations, stands tall in asuburban backyard. The trunk is thick and sturdy, supporting a variety of decorations.Hanging from its branches are luminous orange and black balloons, spooky spiderwebs,and fluttering ghosts. A large, carved pumpkin sits at the base, its intricate faceaglow with a warm, welcoming light. The scene is set against a backdrop of neatlytrimmed hedges and a path leading up to a quaint house, all bathed in the soft glowof autumn sunlight.

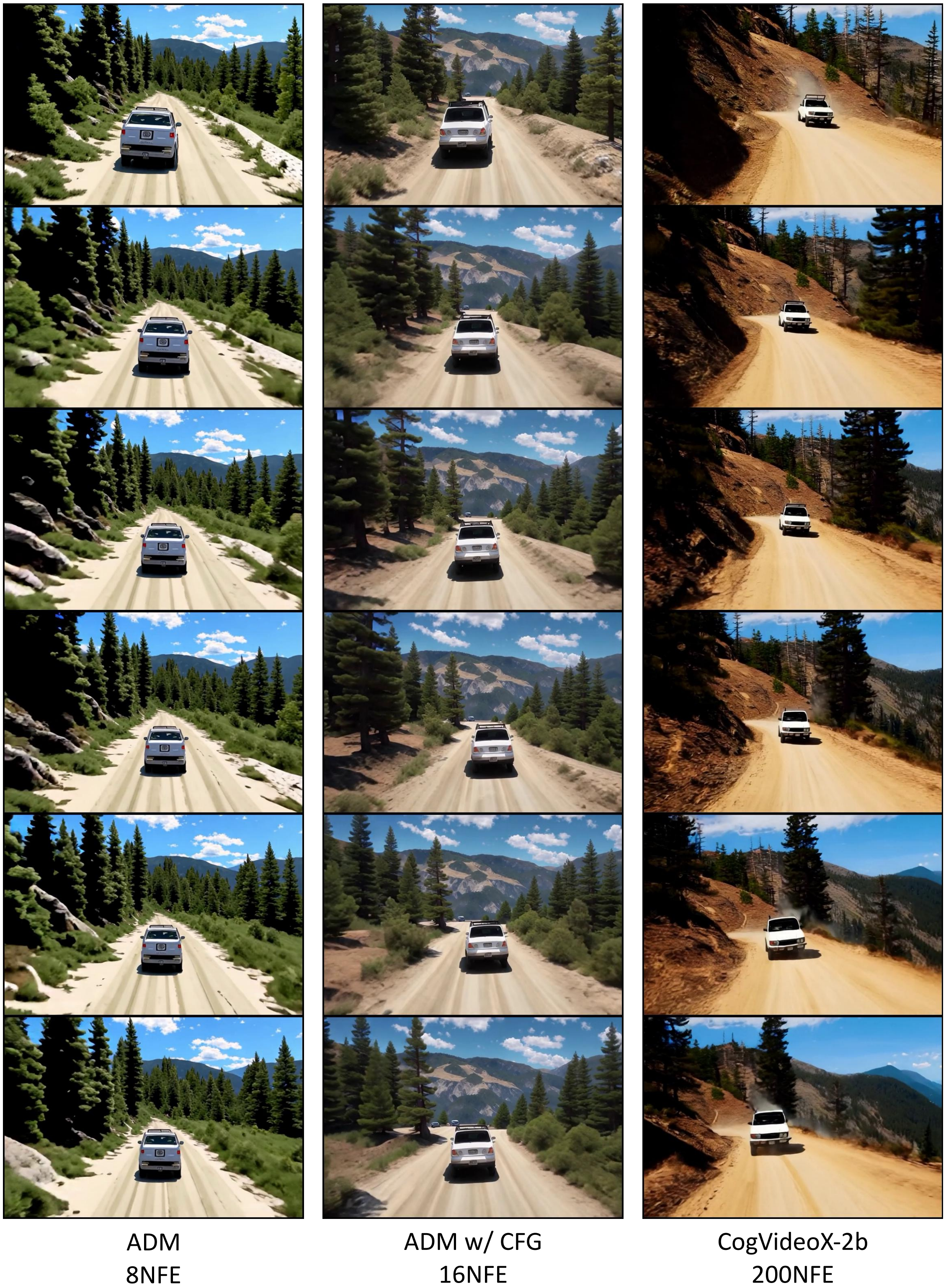


Figure 13. Qualitative comparisons on CogVideoX-2b generators. The random seed has been fixed. **Prompt:** The camera follows behind a white vintage SUV with a black roof rack as it speedsup a steep dirt road surrounded by pine trees on a steep mountain slope, dust kicksup from it's tires, the sunlight shines on the SUV as it speeds along the dirt road,casting a warm glow over the scene. The dirt road curves gently into the distance,with no other cars or vehicles in sight. The trees on either side of the road areredwoods, with patches of greenery scattered throughout. The car is seen from therear following the curve with ease, making it seem as if it is on a rugged drivethrough the rugged terrain. The dirt road itself is surrounded by steep hills andmountains, with a clear blue sky above with wispy clouds.

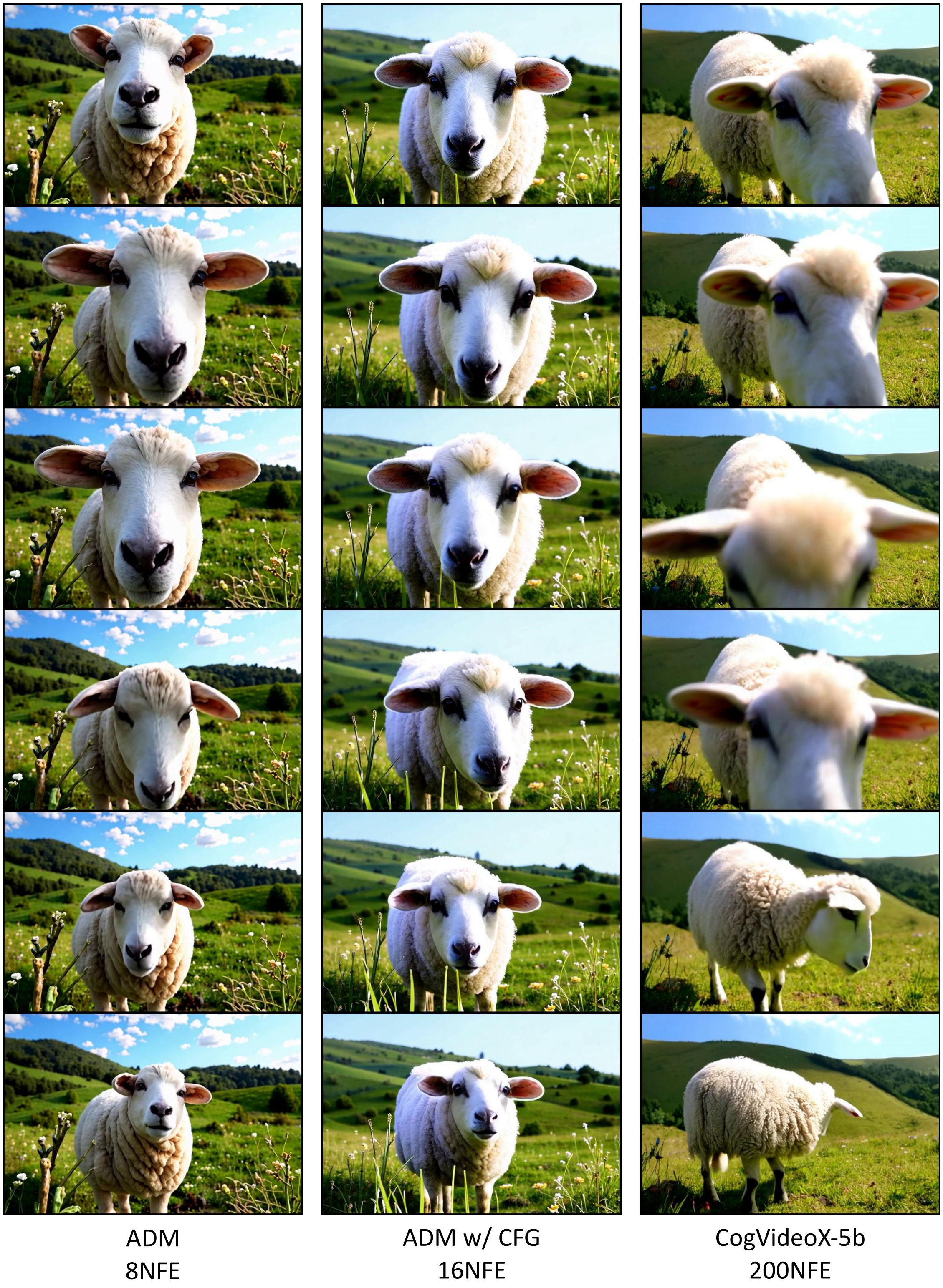


Figure 14. Qualitative comparisons on CogVideoX-5b generators. The random seed has been fixed. **Prompt:** A fluffy, white sheep stands in a lush, green meadow, its wool glistening under the warm afternoon sun. The scene transitions to a close-up of the sheep's gentle face, its big, curious eyes and soft, twitching ears capturing attention. The background features rolling hills dotted with wildflowers and a clear blue sky. The sheep then grazes peacefully, its movements slow and deliberate, as a gentle breeze rustles the grass. Finally, the sheep looks up, framed by the picturesque landscape, embodying tranquility and the simple beauty of nature.

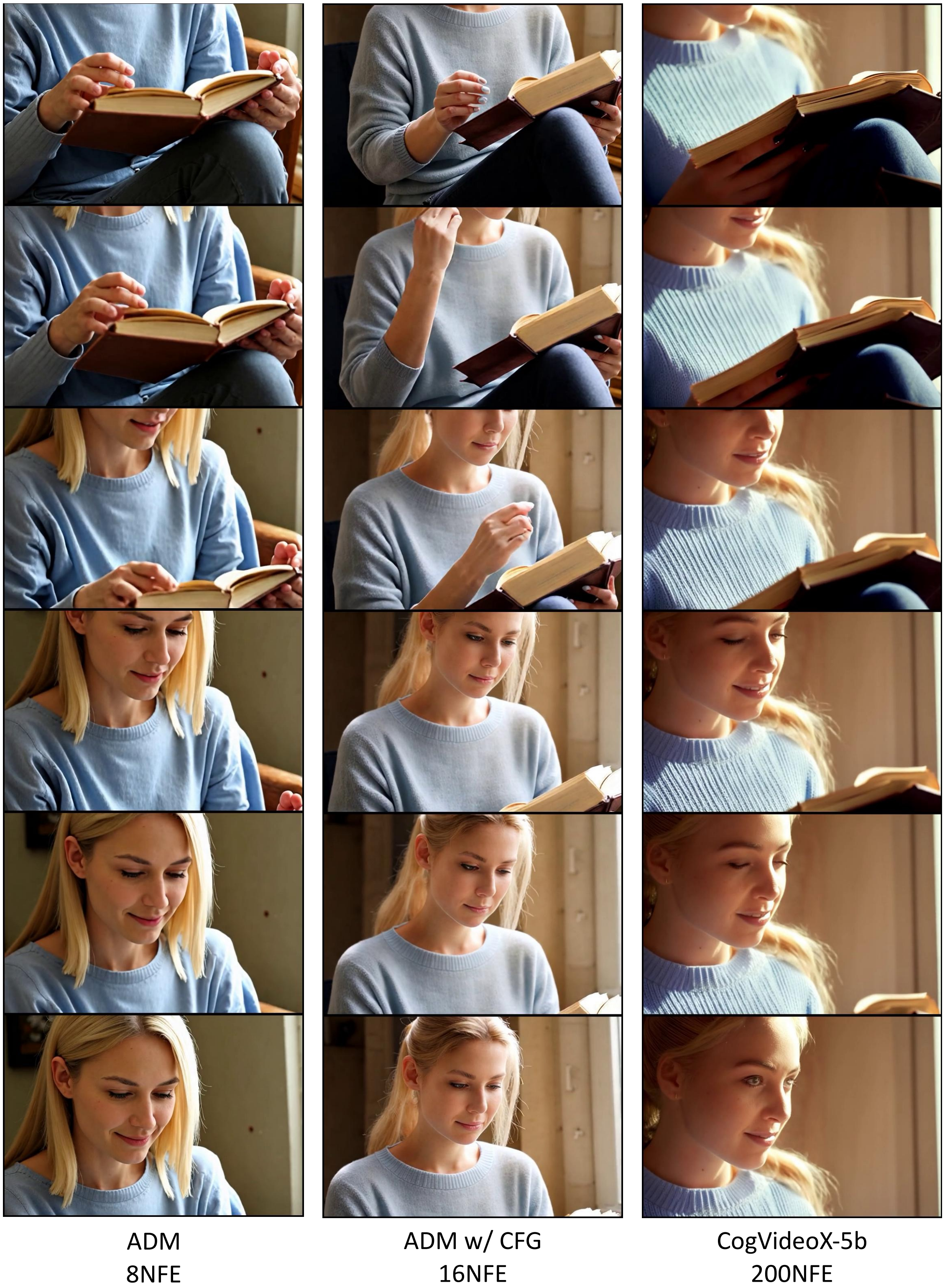


Figure 15. Qualitative comparisons on CogVideoX-5b generators. The random seed has been fixed. **Prompt:** Gwen Stacy, with her signature blonde hair tied back in a ponytail, sits in a cozy, sunlit room, engrossed in a thick, leather-bound book. She wears a casual yet stylish outfit: a light blue sweater, dark jeans, and black ankle boots. The camera starts at her hands, delicately turning a page, revealing her neatly painted nails. As the camera tilts up, it captures her focused expression, her eyes scanning the text with curiosity and intensity. The warm sunlight filters through a nearby window, casting a soft glow on her face, highlighting her serene and studious demeanor. The scene ends with a close-up of her thoughtful smile, suggesting a moment of discovery or reflection.

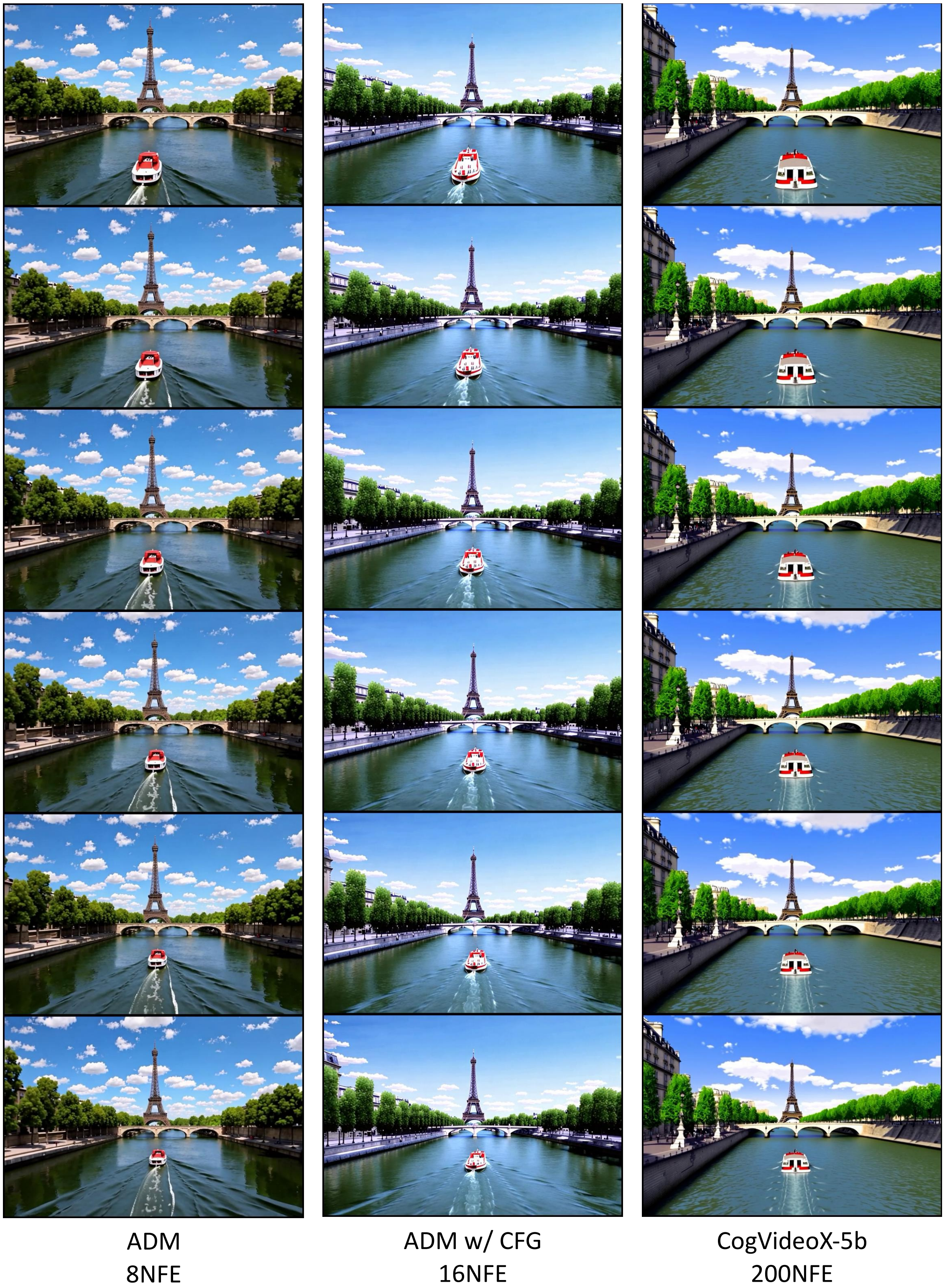


Figure 16. Qualitative comparisons on CogVideoX-5b generators. The random seed has been fixed. **Prompt:** A charming boat with a red and white hull sails leisurely along the serene Seine River, its gentle wake creating ripples in the water. The iconic Eiffel Tower stands majestically in the background, framed by a clear blue sky and fluffy white clouds. As the camera zooms out, the scene expands to reveal lush green trees lining the riverbanks, quaint Parisian buildings with their classic architecture, and pedestrians strolling along the cobblestone pathways. The boat continues its tranquil journey, passing under elegant stone bridges adorned with ornate lampposts, capturing the essence of a peaceful day in Paris.