Adaptive Caching for Faster Video Generation with Diffusion Transformers - Supplementary Material

Kumara Kahatapitiya[1,2*] Haozhe Liu[1] Sen He[1] Ding Liu[1] Menglin Jia[1]
Chenyang Zhang[1] Michael S. Ryoo[2] Tian Xie[1]
[1]Meta AI [2]Stony Brook University
adacache-dit.github.io

A.1. Video qualitative comparisons

We provide a webpage with video results to better highlight the qualitative comparisons, which we encourage the reader to view. It includes additional video generations, comparisons with prior work (*e.g.* PAB [7]) and ablations on Motion regularization, validating the better quality-latency trade-offs of the proposed method.

A.2. Design decisions

Motion-score and motion-gradient: We rely on two metrics in our Motion regularization: namely, motion-score (m_t) and motion-gradient (mg_t). As previously-discussed, motion-score can be unreliable particularly in early diffusion steps as it is estimated based on noisy-latents. For instance, in videos with higher motion content, our motion-score often starts small and gradually increases towards the end of diffusion process (see the two rightmost columns in Fig. A.1). In slow-moving videos, motion-score can start higher can converge to a smaller value (see the leftmost column in Fig. A.1). Simply put, we need a predictor of actual motion (*i.e.*, motion in latter steps $\approx$ motion in pixel space) early in the diffusion process for a proper caching regularization. Therefore, we compute a motion-gradient across diffusion-steps, which can act as such a reasonable predictor (orange bars in Fig. A.1). Together, motion-score and motion-gradient regularize the caching schedule, allocating computations based on the motion content of the video being generated.

Codebook of basis cache-rates: We devise our caching schedule based on a pre-defined codebook of *basis cache-rates*. It is a collection of cache-rates that is specific to a denoising schedule (*i.e.*, #steps), coupled with distance metric (c_t) thresholds for selection. Both basis cache-rates and thresholds are hyperparameters. Here, optimal thresholds may need to be tuned per video-DiT baseline, whereas the cache-rates can be adjusted depending on the required speedup (*e.g.* AdaCache-fast, AdaCache-slow). For instance, on Open-Sora [8] baseline, we use the codebook {`0.08:6, 0.16:5, 0.24:4, 0.32:3, 0.40:2, 1.00:1`} for AdaCache-fast in a 30-step denoising schedule, and the codebook {`0.03:12, 0.05:10, 0.07:8, 0.09:6, 0.11:4, 1.00:3`} in a 100-step schedule. For AdaCache-slow in a 30-step schedule, we use the codebook {`0.08:3, 0.16:2, 0.24:1 1.00:1`}. A specific cache-rate is selected if the distance metric is smaller than the corresponding threshold (and larger than any previous thresholds). Note that AdaCache-slow (30-step) above has smaller basis cache-rates (w/ same thresholds) compared to AdaCache-fast (30-step), corresponding to more-frequent feature re-computations, and hence, smaller speedups.

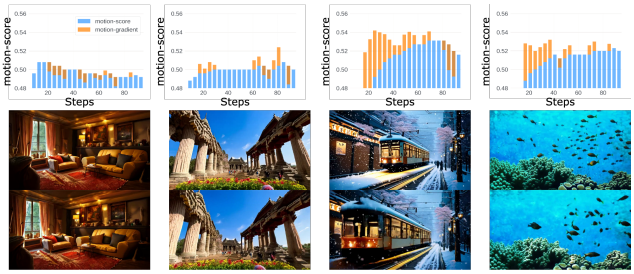

Figure A.1. **Change in motion-score and motion-gradient across steps:** We show the histograms of Motion Regularization metrics (namely, motion-score and motion-gradient) across diffusion steps. Here, motion-score is estimated as latent frame-differences, which correlates well with the perceived motion of a given video sequence. However, it can be unreliable in early denoising steps as such latent representations are noisy. To predict the actual motion (*i.e.*, motion in latter steps $\approx$ motion in pixel space) early, we rely on motion-gradient *across diffusion steps*. Together, motion-score and motion-gradient provide a reasonable regularization. Best-viewed with zoom-in.

Codebook calibration and generalization: To tune the codebook hyperparameters, we follow the steps below: *(i)* select a small calibration set of random video generation prompts (*e.g.* 16 prompts), *(ii)* observe the range in cache-metric distribution (*e.g.* L1 distance) across the denoising process of baseline, *(iii)* split this range uniformly to pair with a number of basis cache-rates (depending on a desired

*Work done at Meta

Method	FID ↓	sFID ↓	IS ↑	Precision ↑	Recall ↑	Latency (s)	Speedup
DIT-XL/2 [4]	2.30	4.56	276.56	0.83	0.58	16.15	1.00×
+ FORA [5] (Thres=3)	2.82	6.04	253.96	0.80	0.58	6.68	2.42×
+ FORA [5] (Thres=5)	4.97	9.15	222.97	0.76	0.59	4.80	3.36×
+ AdaCache	3.27	7.19	243.21	0.79	0.59	5.98	2.70×

Method	VBench ↑	Latency (s)	Speedup
CogVideoX-2B [6]	82.20	152.70	1.00×
+ FasterCache [2]	82.13	102.32	1.49×
+ AdaCache-fast	82.00	92.51	1.65×
+ AdaCache-slow	82.46	102.47	1.49×

Table A.1. **Additional quantitative results:** (Left) We evaluate AdaCache (w/o Motion Regularization) on DiT-XL/2 [4] baseline for class-conditioned image generation. (Right) We evaluate AdaCache on multi-modal DiT backbone CogVideoX [6] for video generation.

granularity), and finally *(iv)* set the basis cache-rates depending on the required quality-latency trade-off— which are user-defined and can be adjusted at inference without needing to tune anything else. We use the same process (w/ the same prompts) across all our baselines, achieving consistently-better quality-latency trade-offs compared to prior training-free DiT acceleration methods. Such a generalization is due to *(a)* the consistent range of feature similarities between subsequent steps, in a given DiT backbone (as shown in Fig. 2b), and *(b)* our normalized cache-metric.

A.3. Additional quantitative results

Here, we show the generalization of AdaCache with additional experiments on image-DiTs and multi-modal video DiTs. In Table A.1 (left), we implement AdaCache (w/o Motion Regularization) on top of an image generation baseline: DiT-XL/2 [4], and compare with the concurrent work FORA [5]. Conceptually, FORA is different from AdaCache, as it is a caching mechanism proposed purely for accelerating image-DiTs (not extended to video generation), and is not adaptive w.r.t. the input. We observe that AdaCache shows a competitive performance with FORA on all quantitative metrics. This shows that AdaCache (originally proposed for accelerating video generation) can also generalize to image generation pipelines. In Table A.1 (right), we implement AdaCache on top of a multi-modal diffusion transformer for video generation: CogVideoX [6], and compare with the concurrent work FasterCache [2]— a training-free inference acceleration method that is not content adaptive. Here, we generate 480p - 6s videos following the baseline, and evaluate on prompts from Open-Sora gallery. We observe that AdaCache shows a better quality-latency trade-off compared to FasterCache, validating that it can work with multi-modal DiTs.

A.4. Additional qualitative results

In Fig. A.2, we provide additional qualitative results, comparing AdaCache and AdaCache (w/ MoReg) with a baseline Open-Sora [8]. Here, we consider 480p - 2s video generations at 30-steps, based on a few VBench [1] prompts. Both versions with and without motion regularization achieve comparable speedups (2.10× and 2.24×, respectively). Yet, MoReg helps stabilize the generation quality— especially towards the end-of-sequence in long-videos— by allocating computations proportional to the amount of motion. The generations with motion regularization also follow the corresponding baseline generations more-faithfully. In Fig. A.3, we present additional qualitative comparisons with prior-art at a comparable inference speedup. Here, we consider 720p - 2s video generations at 100-steps, based on a few Sora [3] prompts. Our comparison includes PAB [7]: another training-free video-DiT acceleration method. AdaCache consistently shows a better generation quality at a 2.61× speedup, compared to PAB, even at a 1.66× speedup. This behavior is also observed in Fig. 5, as the generation quality of PAB degrades quickly at faster speeds.

A.5. Limitations

Despite the strong quality-latency tradeoffs demonstrated by AdaCache, here we discuss a few limitations of its current implementation: *(i)* As we do not rely on any re-training (or, finetuning) of the baseline model (which gives considerable compute savings and data acquisition costs), any limitations that are present in the corresponding baseline may transfer to the AdaCache variant of the same model. It is important that we raise caution about this to the user. *(ii)* In the current setup, the hyperparameters related to the caching schedule (*e.g.* basis cache-rates, cache metric thresholds) are set based on heuristics and empirical validation on a small set of video prompts. Although these generalize well as we observe in our experiments, they may require some tuning when adopting to different baseline models or denoising schedules. *(iii)* Finally, as our computational graph is adaptive, it may be less-suited in custom hardware architectures that rely on fixed (*i.e.*, static) computational graphs for running model inference (*e.g.* custom chips for on-device inference). AdaCache variant with a fixed caching schedule (tuned with a pre-defined calibration dataset) will work better in such scenarios.

A.6. Ethics Statement

This paper introduces a generic training-free inference acceleration mechanism for video diffusion transformers. The merits of the proposed method are evaluated on publicly-available open-source video-DiTs without being tied to a specific model or any commercial application. Consequently, the potential negative impacts of our method align with those of other video generation models and it pose no unique risk that requires special consideration.

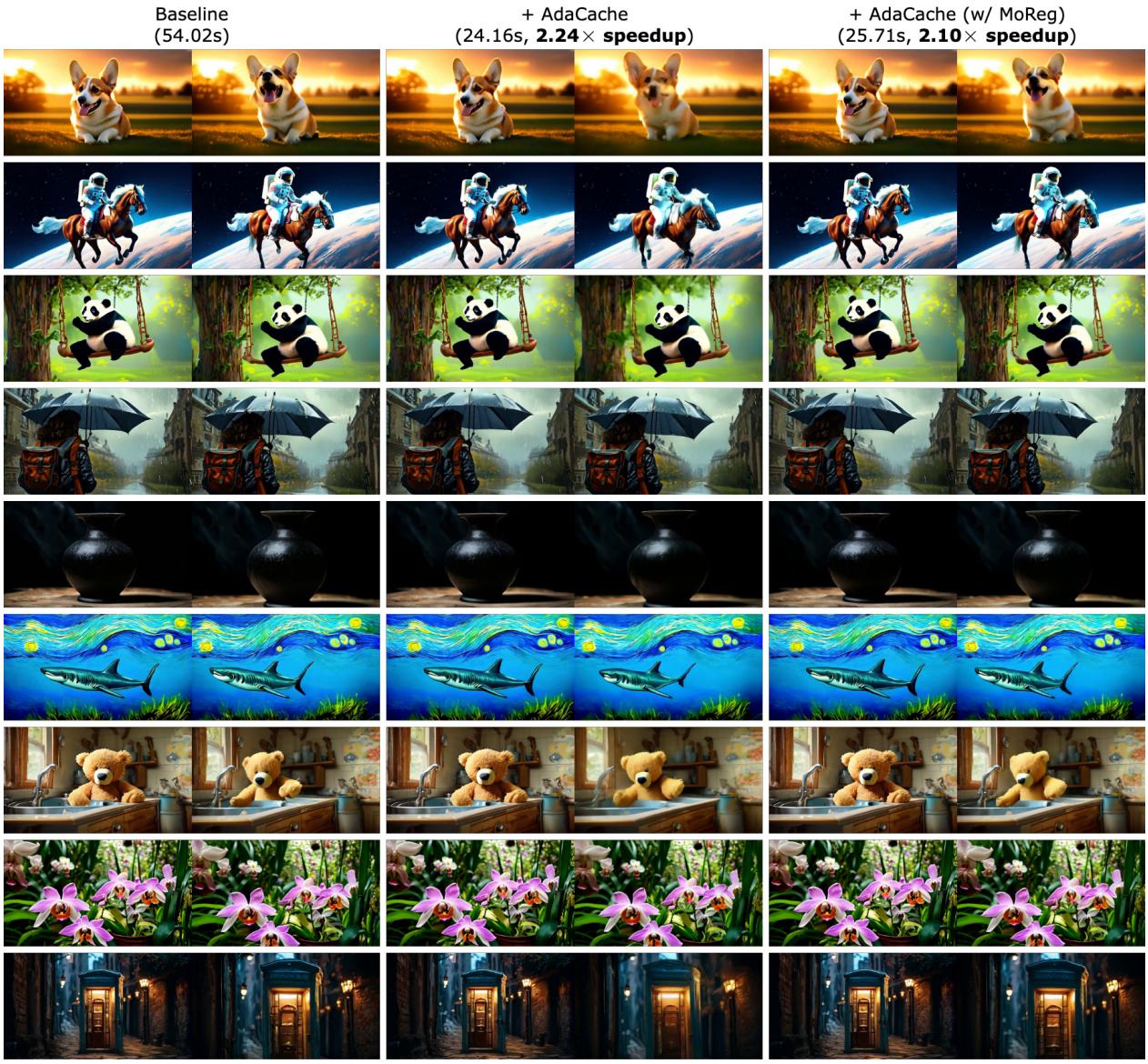

Figure A.2. **Additional qualitative results on our Motion Regularization:** We show a qualitative comparison of AdaCache and AdaCache (w/ MoReg), applied on top of Open-Sora [8] baseline. Here, we consider generation of 480p - 2s clips at 30-steps. Despite giving a 2.24× speedup, AdaCache can also introduce some inconsistencies over time. Our Motion Regularization helps avoid most of them by allocating computations proportional to the amount of motion (still giving a 2.10× speedup). Best-viewed with zoom-in.

A.7. Reproducibility Statement

This paper considers open-source video DiTs (w/ publicly-available code and pretrained-weights) in all presented experiments. As it relies on zero-shot (*i.e.*, training-free) inference acceleration, it requires no updates to pre-trained weights. All quantitative evaluations and generated videos correspond to benchmark prompts that are publicly-available. The paper details all required steps to reproduce the proposed contributions and the code is also released to the public, supporting further research on efficient video generation.

A.8. Text prompts used in qualitative examples

In this subsection, we provide all the prompts used to generate the qualitative results shown in the paper. They consist of prompts from multiple sources including Open-Sora [8] gallery, VBench [1] benchmark and Sora [3], all of which are publicly-available.

Text prompts corresponding to the video generations in Fig. 1:

- `A Japanese tram glides through the snowy streets of a city, its sleek design cutting through the falling snowflakes`

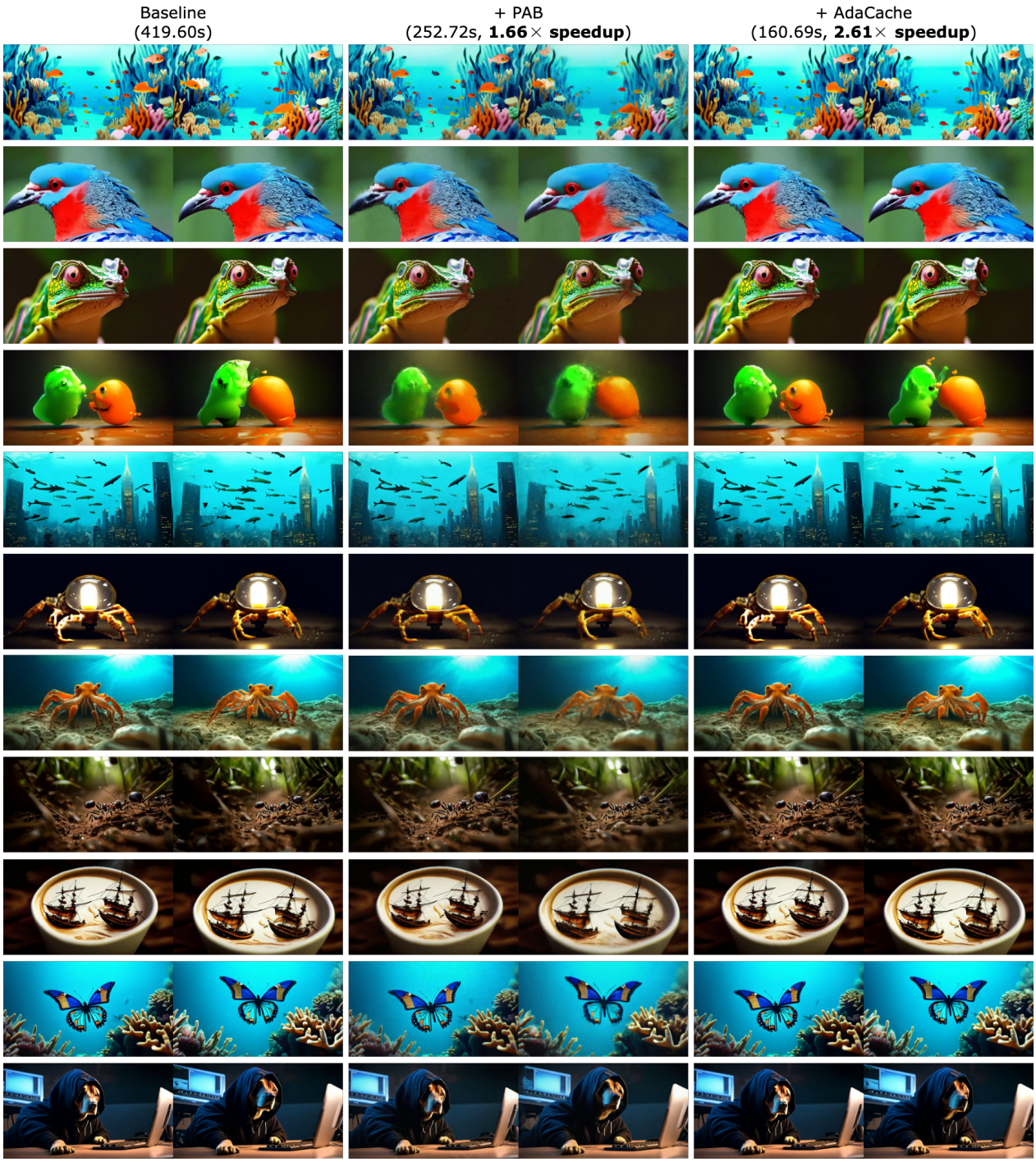

Figure A.3. **Additional qualitative comparisons with prior-art:** We show qualitative comparisons with prior-art on baseline Open-Sora [8] (720p - 2s at 100-steps). Here, we evaluate against prior *training-free* inference acceleration method PAB [7] at a comparable speedup. AdaCache consistently shows a better generation quality. Best-viewed with zoom-in.

```
with grace.  The tram's illuminated
windows cast a warm glow onto the snowy
surroundings, creating a cozy atmosphere
inside.  Snowflakes dance in the air,
swirling around the tram as it moves
along its tracks.  Outside, the city is
blanketed in a layer of snow, transforming
familiar streets into a winter wonderland.
Cherry blossom trees, now bare, stand
quietly along the tram tracks, their
```

branches dusted with snow. People hurry along the sidewalks, bundled up against the cold, while the tram's bell rings softly, announcing its arrival at each stop.

- a picturesque scene of a tranquil beach at dawn. the sky is painted in soft pastel hues of pink and orange, reflecting on the calm, crystal-clear water. gentle waves lap against the sandy shore, where a lone seashell lies near the water's edge. the horizon is dotted with distant, low-lying clouds, adding depth to the serene atmosphere. the overall mood of the video is peaceful and meditative, with no text or additional objects present. the focus is on the natural beauty and calmness of the beach, captured in a steady, wide shot.
- a bustling night market scene with vibrant stalls on either side selling food and various goods. the camera follows a person walking through the crowded, narrow alley. string lights hang overhead, casting a warm, festive glow. people of all ages are talking, browsing, and eating, creating an atmosphere full of lively energy. occasional close-ups capture the details of freshly cooked dishes and colorful merchandise. the video is dynamic with a mixture of wide shots and close-ups, capturing the essence of the night market without any text or sound.
- a dynamic aerial shot showcasing various landscapes. the sequence begins with a sweeping view over a dense, green forest, transitioning smoothly to reveal a winding river cutting through a valley. next, the camera rises to capture a panoramic view of a mountain range, the peaks dusted with snow. the shot shifts to a coastal scene, where waves crash against rugged cliffs under a partly cloudy sky. finally, the aerial view ends over a bustling cityscape, with skyscrapers and streets filled with motion and life. the video does not contain any text or additional overlays.
- a cozy living room scene with a christmas tree in the corner adorned with colorful ornaments and twinkling lights. a fireplace with a gentle flame is situated across from a plush red sofa, which has a few wrapped presents placed beside it. a window to the left reveals a snowy landscape outside, enhancing the festive atmosphere. the camera slowly pans from the window to the fireplace, capturing the warmth and tranquility of the room. the soft glow from the tree lights and the fire illuminates the room, casting a comforting ambiance. there are no people or text in the video, focusing purely on the holiday decor and cozy setting.

Text prompts corresponding to new video generations in Fig. 2:

- a breathtaking aerial view of a river meandering through a lush green landscape. the river, appearing as a dark ribbon, cuts through the verdant fields and hills, reflecting the soft light of the pinkish-orange sky. the sky, painted in hues of pink and orange, suggests the time of day to be either sunrise or sunset. the landscape is dotted with trees and bushes, adding to the natural beauty of the scene. the perspective of the video is from above, providing a bird's eye view of the river and the surrounding landscape. the colors , the river, the landscape, and the sky all come together to create a serene and picturesque scene.
- a nighttime scene in a bustling city filled with neon lights and futuristic architecture. the streets are crowded with people, some dressed in high-tech attire and others in casual cyberpunk fashion. holographic advertisements and signs illuminate the area in vibrant colors, casting a glow on the buildings and streets. futuristic vehicles and motorcycles are speeding by, adding to the city's dynamic atmosphere. in the background, towering skyscrapers with intricate designs stretch into the night sky. the scene is filled with energy, capturing the essence of a cyberpunk world.
- a close-up shot of a vibrant coral reef underwater. various colorful fish swim leisurely around the corals, creating a lively scene. the lighting is natural and slightly subdued, emphasizing the deep-sea

environment. soft waves ripple across the view, occasionally bringing small bubbles into the frame. the background fades into a darker blue, suggesting deeper waters beyond. there are no texts or human-made objects visible in the video.

- a neon-lit cityscape at night, featuring towering skyscrapers and crowded streets. the streets are bustling with people wearing futuristic attire, and vehicles hover above in organized traffic lanes. holographic advertisements are projected onto buildings, illuminating the scene with vivid colors. a light rain adds a reflective sheen to the ground, enhancing the cyberpunk atmosphere. the camera pans slowly through the scene, capturing the energy and technological advancements of the city. the video does not contain any text or additional objects.
- a breathtaking view of a mountainous landscape at sunset. the sky is painted with hues of orange and pink, casting a warm glow over the scene. the mountains, bathed in the soft light, rise majestically in the background, their peaks reaching towards the sky. in the foreground, a woman is seated on a rocky outcrop, her body relaxed as she takes in the vie w. she is dressed in a black dress and boots, her attire contrasting with the natural surroundings. her position on the rock provides a vantage point over a river that meanders through the valley below. the river, a ribbon of blue, winds its way through the landscape, adding a dynamic element to the scene. the woman's gaze is directed towards the river, suggesting a sense of contemplation or admiration for the beauty of nature. the video is taken from a high angle, looking down on the woman and the landscape. this perspective enhances the sense of depth and scale in the image, emphasizing the vastness of the mountains and the river.
- an animated scene featuring a young girl with short black hair and a bow tie, seated at a wooden desk in a warmly lit room. natural light filters through a window, illuminating the girl's wide eyes and open mouth, conveying a sense of surprise or shock. she is dressed in a blue shirt with a white collar and dark vest. the room's inviting atmosphere is complemented by wooden furniture and a framed picture on the wall. the animation style is reminiscent of japanese anime, characterized by vibrant colors and expressive character designs.
- a realistic 3d rendering of a female character with curly blonde hair and blue eyes. she is wearing a black tank top and has a neutral expression while facing the camera directly. the background is a plain blue sky, and the scene is devoid of any other objects or text. the character is detailed, with realistic textures and lighting, suitable for a video game or high-quality animation. there is no movement or additional action in the video. the focus is entirely on the character's appearance and realistic rendering.

Text prompts corresponding to new video generations in Fig. 6 :

- A cozy living room, surrounded by soft cushions and warm lighting. Describe the scene in vivid detail, capturing the feeling of comfort and relaxation.
- a breathtaking aerial view of a misty mountain landscape at sunrise. the sun is just beginning to peek over the horizon, casting a warm glow on the scene. the mountains, blanketed in a layer of fog, rise majestically in the background. the mist is so dense that it obscures the peaks of the mountains, adding a sense of mystery to the scene. in the foregro und, a river winds its way through the landscape, its path marked by the dense fog. the river appears calm, its surface undisturbed by the early morning chill. the colors in the video are predominantly cool, with the blue of the sky and the green of the trees contrasting with the warm orange of the sunrise. the video is taken from a high vantage point, p roviding a bird's eye view of the landscape. this perspective allows for a comprehensive view of the mountains and the river, as well as the fog that envelops them. the video doe s not contain any text or human activity,

focusing solely on the natural beauty of the landscape. the relative positions of the objects suggest a vast, untouched wilderness.

- a 3d rendering of a female character with curly blonde hair and striking blue eyes. she is wearing a black tank top and is standing in front of a fiery backdrop. the character is looking off to the side with a serious expression on her face. the background features a fiery orange and red color scheme, suggesting a volcanic or fiery environment. the lighting in the scene is dramatic, with the character's face illuminated by a soft light that contrasts with the intense colors of the background. there are no texts or other objects in the image. the style of the image is realistic with a high level of detail, indicative of a high-quality 3d rendering.

Text prompts corresponding to new video generations in Fig. 8:

- a scenic shot of a historical landmark. the landmark is an ancient temple with tall stone columns and intricate carvings. the surrounding area is lush with greenery and vibrant flowers. the sky above is clear and blue, with the sun casting a warm glow over the scene. tourists can be seen walking around, taking pictures and admiring the architecture. there is no text or additional objects in the video.
- a vibrant cyberpunk street scene at night. neon signs and holographic advertisements illuminate the narrow street, casting colorful reflections on the rain-slicked pavement. various characters, dressed in futuristic attire, move along the sidewalks while robotic street vendors sell their wares. towering skyscrapers with glowing windows dominate the background, creating a sense of depth. the camera takes a wide-angle perspective, capturing the bustling and lively atmosphere of the cyberpunk cityscape. there are no texts or other objects outside of the described scene.

Text prompts corresponding to new video generations in Fig. A.2:

- A cute happy Corgi playing in park, sunset, surrealism style
- An astronaut is riding a horse in the space in a photorealistic style.
- A panda playing on a swing set
- a backpack and an umbrella
- a black vase
- a shark is swimming in the ocean, Van Gogh style
- A teddy bear washing the dishes
- A tranquil tableau of a peaceful orchid garden showcased a variety of delicate blooms
- A tranquil tableau of the phone booth was tucked away in a quiet alley

Text prompts corresponding to new video generations in Fig. A.3:

- A gorgeously rendered papercraft world of a coral reef, rife with colorful fish and sea creatures.
- This close-up shot of a Victoria crowned pigeon showcases its striking blue plumage and red chest. Its crest is made of delicate, lacy feathers, while its eye is a striking red color. The bird's head is tilted slightly to the side, giving the impression of it looking regal and majestic. The background is blurred, drawing attention to the bird's striking appearance.
- This close-up shot of a chameleon showcases its striking color changing capabilities. The background is blurred, drawing attention to the animal's striking appearance.
- a green blob and an orange blob are in love and dancing together
- New York City submerged like Atlantis. Fish, whales, sea turtles and sharks swim through the streets of New York.
- nighttime footage of a hermit crab using an incandescent lightbulb as its shell
- A large orange octopus is seen resting on the bottom of the ocean floor, blending in with the sandy and rocky terrain. Its tentacles are spread out around its body, and its eyes are closed. The octopus is unaware of a king crab that is crawling towards it from behind a rock, its claws raised and ready to attack. The crab

`is brown and spiny, with long legs and antennae. The scene is captured from a wide angle, showing the vastness and depth of the ocean. The water is clear and blue, with rays of sunlight filtering through. The shot is sharp and crisp, with a high dynamic range. The octopus and the crab are in focus, while the background is slightly blurred, creating a depth of field effect.`

- `A low to the ground camera closely following ants in the jungle down into the ground into their world.`
- `Photorealistic closeup video of two pirate ships battling each other as they sail inside a cup of coffee.`
- `a photorealistic video of a butterfly that can swim navigating underwater through a beautiful coral reef`
- `A computer hacker labrador retreiver wearing a black hooded sweatshirt sitting in front of the computer with the glare of the screen emanating on the dog's face as he types very quickly.`

References

[1] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2, 3

[2] Zhengyao Lv, Chenyang Si, Junhao Song, Zhenyu Yang, Yu Qiao, Ziwei Liu, and Kwan-Yee K Wong. Fastercache: Training-free video diffusion model acceleration with high quality. *arXiv preprint arXiv:2410.19355*, 2024. 2

[3] Inc. OpenAI. Video generation models as world simulators. 2024. *URL https://openai. com/research/video-generation-models-as-world-simulators*, 3, 2024. 2, 3

[4] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 2

[5] Pratheba Selvaraju, Tianyu Ding, Tianyi Chen, Ilya Zharkov, and Luming Liang. Fora: Fast-forward caching in diffusion transformer acceleration. *arXiv preprint arXiv:2407.01425*, 2024. 2

[6] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 2

[7] Xuanlei Zhao, Xiaolong Jin, Kai Wang, and Yang You. Real-time video generation with pyramid attention broadcast. *arXiv preprint arXiv:2408.12588*, 2024. 1, 2, 4

[8] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, 2024. 1, 2, 3, 4