CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer

Zhuoyi Yang* **Jiayan Teng*** **Wendi Zheng** **Ming Ding** **Shiyu Huang**
Jiazheng Xu **Yuanming Yang** **Xiaohan Zhang** **Xiaotao Gu** **Guanyu Feng** **Da Yin**
Wenyi Hong **Weihan Wang** **Yean Cheng** **Yuxuan Zhang** **Ting Liu** **Bin Xu**
Yuxiao Dong **Jie Tang**

Zhipu AI Tsinghua University

Abstract

We introduce CogVideoX, a large-scale diffusion transformer model designed for generating videos based on text prompts. To efficently model video data, we propose to levearge a 3D Variational Autoencoder (VAE) to compresses videos along both spatial and temporal dimensions. To improve the text-video alignment, we propose an expert transformer with the expert adaptive LayerNorm to facilitate the deep fusion between the two modalities. By employing a progressive training technique, CogVideoX is adept at producing coherent, long-duration videos characterized by significant motion. In addition, we develop an effectively text-video data processing pipeline that includes various data preprocessing strategies and a video captioning method. It significantly helps enhance the performance of CogVideoX, improving both generation quality and semantic alignment. Results show that CogVideoX demonstrates state-of-the-art performance across both multiple machine metrics and human evaluations. The model weight of CogVideoX-2B is publicly available at `https://github.com/THUDM/CogVideo`.

1 Introduction

In recent years, diffusion models have made groundbreaking advancements in multimodal generation, such as image, video, speech and 3D generation. Among these, video generation is a rapidly evolving field and being extensively explored. Given the successful experiences with Large Language Models (LLMs), comprehensive scaling up of data volume, training iterations, and model size consistently enhances model performance. Additionally, there is more mature scaling experience with transformers compared to UNet. And DiT (Peebles & Xie, 2023) has shown that transformers can effectively replace UNet as the backbone of diffusion models. Thus, transformer is a better choice for video generation. However, long-term consistent video generation remains a significant challenge.

The first challenge is that constructing a web-scale video data pipeline is considerably more difficult than for textual data. Video data is extremely diverse in distribution, quality varies greatly, and simple rule-based filtering is often insufficient for effective data selection. Consequently, processing video data is both time-consuming and highly complex. There are numerous meaningless unrealistic videos, such as poor-quality edits and computer screen recordings. And many videos are difficult to watch normally, such as those with excessively shaky cameras. These types of data are harmful to the generative model's ability to learn genuine dynamic information. They need to be meticulously processed and filtered out to ensure the quality of the training dataset.

*Equal contributions. Core contributors: Zhuoyi, Jiayan, Wendi, Ming, and Shiyu.
{yangzy22,tengjy20}@mails.tsinghua.edu.cn, {yuxiaod,jietang}@tsinghua.edu.cn

Additionally, most video data available online lacks accurate textual descriptions, significantly limiting the model's ability to grasp precise semantic understanding. To address this issue, we trained a video understanding model capable of accurately describing video content. We use it to generate new textual descriptions for all video data. To advance the field of video generation, we have decided to open-source this description model.

The high training cost is another significant challenge. If the video is unfolded into a one-dimensional sequence in the pixel space, the length would be extraordinarily long. To keep the computational cost within a feasible range, we trained a 3D VAE that compresses the video along both spatial and temporal dimensions. Additionally, unlike previous video models that use a 2D VAE to encode each frame separately, 3D VAE ensures continuity among frames so that the generated videos do not flicker.

Moreover, to improve the alignment between videos and texts, we propose an expert transformer to facilitate the interaction between the two modalities. Then, to ensure the consistency of video generation and to capture large-scale motions, it is necessary to comprehensively model the video along both temporal and spatial dimensions. Therefore, we opt for 3D full attention, as detailed in Section 2.2.

2 The CogVideoX Architecture

In the section, we present the CogVideoX model. Figure 1 illustrates the overall architecture. Given a pair of video and text input, we design a **3D causal VAE** to compress the video into the latent space, and the latents are then patchified and unfolded into a long sequence denoted as z_{vision}. Simultaneously, we encode the textual input into text embeddings z_{text} using T5 (Raffel et al., 2020). Subsequently, z_{text} and z_{vision} are concatenated along the sequence dimension. The concatenated embeddings are then fed into a stack of **expert transformer** blocks. Finally, the model outputs are unpatchified to restore the original latent shape, which is then decoded using 3D causal VAE decoder to reconstruct the video. We illustrate the technical design of the 3D causal VAE and expert transfomer in detail.

2.1 3D Causal VAE

Videos encompass not only spatial information but also substantial temporal information, usually resulting orders of magnitude more data volumes than images. To tackle the computational challenge of modeling video data, we propose to implement a video compression module based on 3D Variational Autoencoders (3D VAEs) (Yu et al., 2023). The idea is to incorporate three-dimentional convolutions to compress videos both spatially and temporally. This can help achieve a higher compression ratio with largely improved quality and continuity of video reconstruction when compared to previous image VAEs (Rombach et al., 2022; Esser et al., 2021).

Figure 2 (a) shows the structure of the proposed 3D VAE. It comprises an encoder, a decoder and a latent space regularizer. The Gaussian latent space is constrained by a Kullback-Leibler (KL) regularizer. The encoder and decoder consist of four symmetrically arranged stages, respectively performing $2\times$ downsampling and upsampling by the interleaving of resnet block stacked stages. The first two rounds of downsampling and upsampling involve both the spatial and temporal dimensions, while the last round only samples spatially. This enables the 3D VAE to achieve a $4\times$ compression in the temporal dimension and an 8×8 compression in the spatial dimension. In total, this achieves a $4\times8\times8$ compression from pixels to the latents.

We adopt the temporally causal convolution (Yu et al., 2023), which places all the paddings at the beginning of the convolution space, as shown in Figure 2 (b). This ensures the future information not to influence the present or past predictions. Given that processing videos with a large number of frames introduces excessive GPU memory usage, we apply the context parallel technique at the temporal dimension for 3D convolution to distribute computation among multiple devices. As illustrated by Figure 2 (b), due to the causal nature of the convolution, each rank simply sends a segment of length $k-1$ to the next rank, where k indicates the temporal kernel size. This results in relatively low communication overhead.

During actual implementation, we first train a 3D VAE on lower resolutions and fewer frames to save computation. We observe that the encoding of larger resolution generalizes naturally, while extending the number of frames to be encoded does not work as seamlessly. Therefore, we conduct a two-stage

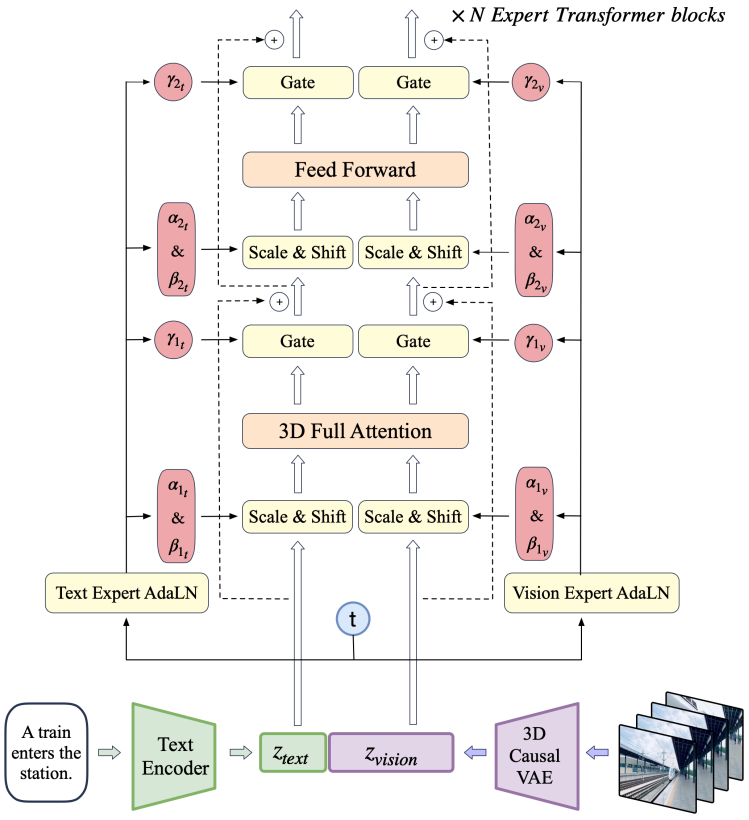


Figure 1: **The overall architecture of CogVideoX.**

training process by first training on short videos and finetuning by context parallel on long videos. Both stages of training utilize a weighted combination of the L2 loss, LPIPS (Zhang et al., 2018) perceptual loss, and GAN loss from a 3D discriminator.

2.2 Expert Transformer

We introduce the design choices in Transformer for CogVideoX, including the patching, positional embedding, and attention strategies for handling the text-video data effectively and efficiently.

Patchify. The 3D causal VAE encodes a video latent vector of shape $T \times H \times W \times C$, where T represents the number of frames, H and W represent the height and width of each frame, C represents the channel number, respectively. The video latents are then patchified along the spatial dimensions, generating sequence z_{vision} of length $T \cdot \frac{H}{p} \cdot \frac{W}{p}$. Note that, we do not patchify along the temporal dimension in order to enable joint training of images and videos.

3D-RoPE. Rotary Position Embedding (RoPE) (Su et al., 2024) is a relative positional encoding that has been demonstrated to capture inter-token relationships effectively in LLMs, particularly excelling in modeling long sequences. To adapt to video data, we extend the original RoPE to 3D-RoPE. Each latent in the video tensor can be represented by a 3D coordinate (x, y, t). We independently apply 1D-RoPE to each dimension of the coordinates, each occupying $3/8$, $3/8$, and $2/8$ of the hidden states' channel. The resultsing encoding are then concatenated along the channel dimension to obtain the final 3D-RoPE encoding.

We empirically examine the use of RoPE. Figure3a shows the comparision between 3D RoPE and sinusoidal absolute position encoding. We can observe that the loss curve using 3D RoPE converges significantly faster than that with sinusoidal encoding. We further compare the use of 3D RoPE alone against the combination of 3D RoPE and learnable absolute position embedding. Figure 3b indicates

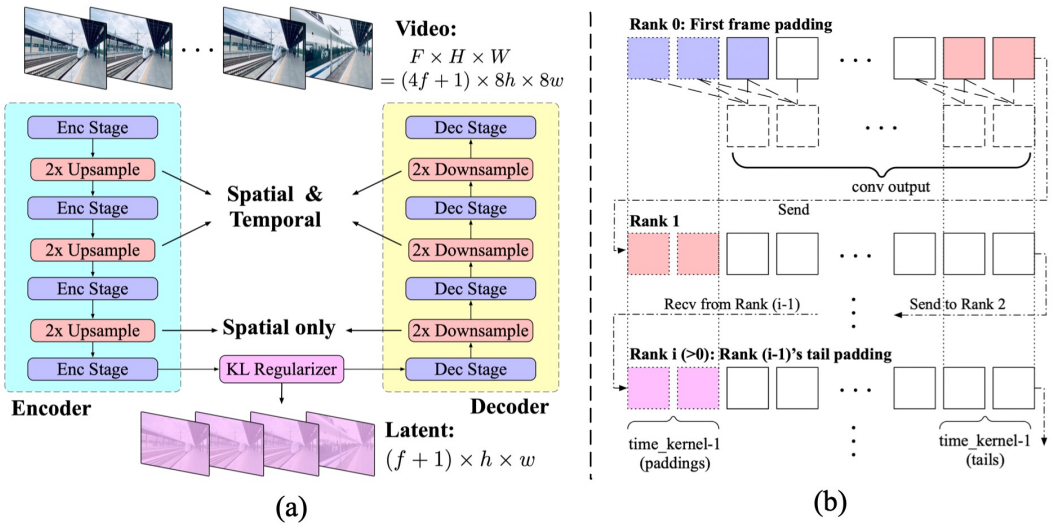
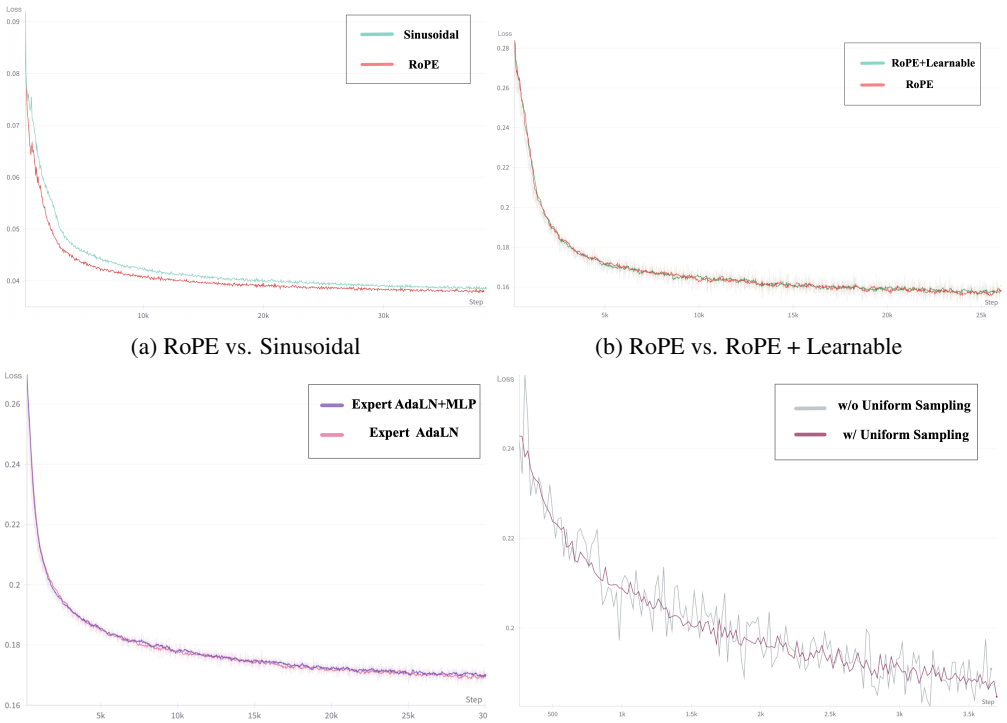


Figure 2: (a) The structure of the 3D VAE in CogVideoX. It comprises an encoder, a decoder and a latent space regularizer, achieving a 4×8×8 compression from pixels to the latents. (b) The context parallel implementation on the temporally causal convolution.

that the loss curves of both methods converge almost identically. Therefore, we choose to use 3D RoPE alone for simplicity.



(a) RoPE vs. Sinusoidal

(b) RoPE vs. RoPE + Learnable

(c) Expert Ada. LN vs. Expert Ada. LN + MLP

(d) Uniform Sampling vs. No Uniform Sampling

Figure 3: Training loss curve of different ablations.

Expert Transformer Block. We concatenate the embeddings of both text and video at the input stage to better align visual and semantic information. However, the feature spaces of these two modalities differ significantly, and their embeddings may even have different numerical scales. To

better process them within the same sequence, we employ the Expert Adaptive Layernorm to handle each modality independently. As shown in Figure 1, following DiT (Peebles & Xie, 2023), we use the timestep t of the diffusion process as the input to the modulation module. Then, the Vision Expert Adaptive Layernorm (Vison Expert AdaLN) and Text Expert Adaptive Layernorm (Text Expert AdaLN) apply this modulation mechanism to the vision hidden states and text hidden states, respectively. This strategy promotes the alignment of feature spaces across two modalities while minimizing additional parameters.

To verify the adoption of Expert Adaptive Layernorm, we experiment with different ways of incorporating experts: expert LayerNorm and MLP, and expert Layernorm only. Our experiments find that adding expert MLP does not effectively accelerate the model's convergence (Cf. Figure 3c). To reduce the model parameters, we only choose to use Expert Adaptive Layernorm.

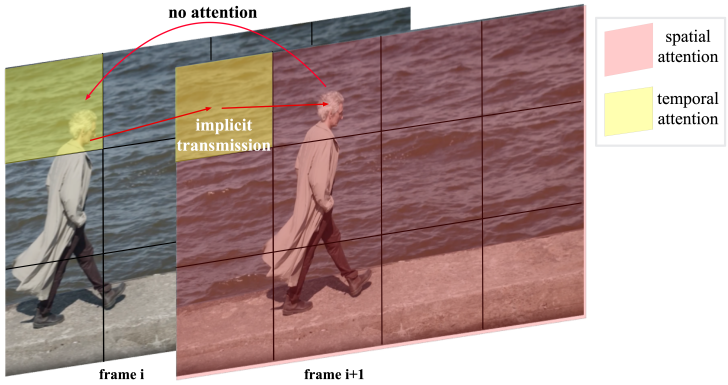


Figure 4: The separated spatial and temporal attention makes it challenging to handle the large motion between adjacent frames. In the figure, the head of the person in frame $i+1$ cannot directly attend to the head in frame i. Instead, visual information can only be implicitly transmitted through other background patches. This can lead to inconsistency issues in the generated videos.

3D Full Attention. Previous works (Singer et al., 2022; Guo et al., 2023) often employ separated spatial and temporal attention to reduce computational complexity and facilitate fine-tuning from text-to-image models. However, as illustrated in Figure 4, this separated attention approach requires extensive implicit transmission of visual information, significantly increasing the learning complexity and making it challenging to maintain the consistency of large-movement objects. Considering the great success of long-context training in LLMs (AI@Meta, 2024; Bai et al., 2024; Xiong et al., 2023) and the efficiency of FlashAttention (Dao et al., 2022), we propose a 3D text-video hybrid attention mechanism. This mechanism not only achieves better results but can also be easily adapted to various parallel acceleration methods.

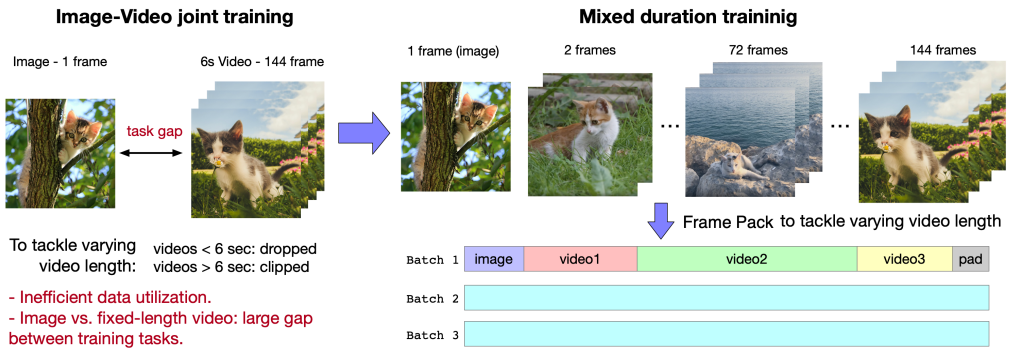


Figure 5: The diagram of mixed-duration training and Frame Pack. To fully utilize the data and enhance the model's generalization capability, we train with videos of different durations within the same batch.

3 Training CogVideoX

We mix images and videos during training, treating each image as a single-frame video. Additionally, we employ progressive training from the resolution perspective. For the diffusion setting, we adopt v-prediction (Salimans & Ho, 2022) and zero SNR (Lin et al., 2024), following the noise schedule used in LDM (Rombach et al., 2022). During diffusion training for timestep sampling, we also employ an explicit uniform timestep sampling method, which benefits training stability.

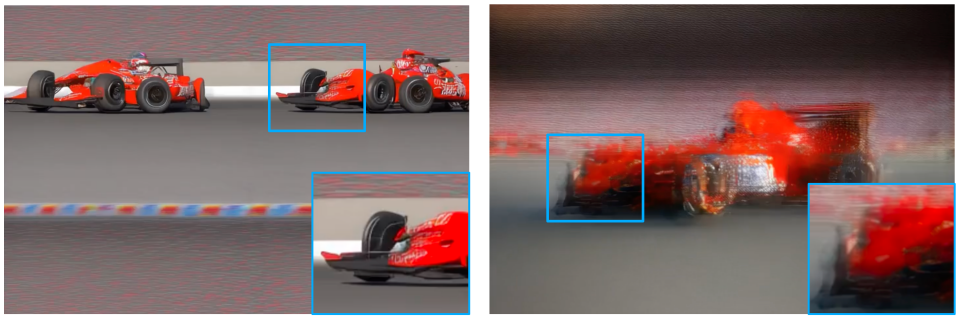
3.1 Frame Pack

Previous video training methods often involve joint training of images and videos with fixed number of frames (Singer et al., 2022; Blattmann et al., 2023). However, this approach usually leads to two issues: First, there is a significant gap between the two input types using bidirectional attention, with images having one frame while videos having dozens of frames. We observe that models trained this way tend to diverge into two generative modes based on the token count and to not have good generalization. Second, to train with a fixed duration, we have to discard short videos and truncate long videos, which prevents full utilization of the videos of varying number of frames.

To address these issues, we choose mixed-duration training, which means training videos of different lengths together. However, inconsistent data shapes within the batch make training difficult. Inspired by Patch'n Pack (Dehghani et al., 2024), we place videos of different lengths into the same batch to ensure consistent shapes within each batch, a method we refer to as *Frame Pack*. The process is illustrated in 5.

3.2 Resolution Progressive Training

The training pipeline of CogVideoX is divided into three stages: low-resolution training, high-resolution training, and high-quality video fine-tuning. Similar to images, videos from Internet usually include a significant amount of low-resolution ones. Progressive training can effectively utilize videos of various resolutions. Moreover, training at low resolution initially can equip the model with coarse-grained modeling capabilities, followed by high-resolution training to enhance its ability to capture fine details. Compared to direct high-resolution training, staged training can also help reduce the overall training time.



RoPE Extrapolation RoPE Interpolation

Figure 6: The comparison between the initial generation states of extrapolation and interpolation when increasing the resolution with RoPE encoding. Extrapolation tends to generate multiple small, clear, and repetitive images, while interpolation generates a blurry large image.

Extrapolation of Position Code. When adapting low-resolution position encoding to high-resolution, we consider two different methods: interpolation and extrapolation. We show the effects of two methods in Figure 6. Interpolation tens to preserve global information more effectively, whereas the extrapolation better retains local details. Given that RoPE is a relative position encoding, We chose the extrapolation to maintain the relative position between pixels.

High-Quality Fine-Tuning. Since the filtered pre-training data still contains a certain proportion of dirty data, such as subtitles, watermarks, and low-bitrate videos, we selected a subset of higher

quality video data, accounting for 20% of the total dataset, for fine-tuning in the final stage. This step effectively removed generated subtitles and watermarks and slightly improved the visual quality. However, we also observed a slight degradation in the model's semantic ability.

3.3 Explicit Uniform Sampling

Ho et al. (2020) defines the training objective of diffusion as

$$L_{\text{simple}}(\theta) := \mathbf{E}_{t,x_0,\epsilon}\left\|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\right\|^2, \tag{1}$$

where t is uniformly distributed between 1 and T. The common practice is for each rank in the data parallel group to uniformly sample a value between 1 and T, which is in theory equivalent to Equation 1. However, in practice, the results obtained from such random sampling are often not sufficiently uniform, and since the magnitude of the diffusion loss is related to the timesteps, this can lead to significant fluctuations in the loss. Thus, we propose to use *Explicit Uniform Sampling* to divide the range from 1 to T into n intervals, where n is the number of ranks. Each rank then uniformly samples within its respective interval. This method ensures a more uniform distribution of timesteps. As shown in Figure 3d, the loss curve from training with Explicit Uniform Sampling is noticeably more stable.

3.4 Data

We construct a collection of relatively high-quality video clips with text descriptions through video filters and recaption models.

Video Filtering. Video generation models need to learn the dynamic information of the world, but unfiltered video data is of highly noisy distribution, primarily due to two reasons: First, videos are human-created, and artificial editing may distort the authentic dynamic information; Second, the quality of videos can significantly drop due to issues during filming, such as camera shakes and substandard equipment.

In addition to the intrinsic quality of the videos, we also consider how well the video data supports model training. Videos with minimal dynamic information or lacking connectivity in dynamic aspects are considered detrimental. Consequently, we have developed a set of negative labels, which include:

- **Editing**: Videos that have undergone obvious artificial processing, such as re-editing and special effects, causing degradation of the visual integrity.
- **Lack of Motion Connectivity**: Video segments with image transitions lacking motion connectivity, commonly seen in videos artificially spliced or edited from images.
- **Low Quality**: Poorly shot videos with unclear visuals or excessive camera shake.
- **Lecture Type**: Videos focusing primarily on a person continuously talking with minimal effective motion, such as educational content, lectures, and live-streamed discussions.
- **Text Dominated**: Videos containing a substantial amount of visible text or primarily focusing on textual content.
- **Noisy Screenshots**: Noisy videos recorded from phone or computer screens.

We sample 20,000 video data samples and label the presence of negative tags in each of them. By using these annotations, we train several filters based on video-llama (Zhang et al., 2023b) to screen out low-quality video data.

In addition, we calculate the optical flow scores and image aesthetic scores of all training videos and dynamically adjust the threshold ranges during training to ensure the fluency and aesthetic quality of the generated videos.

Video Caption. Typically, most video data does not come with corresponding descriptive text, so it is necessary to convert the video data into textual descriptions to provide the essential training data for text-to-video models. Currently, there are some video caption datasets available, such as Panda70M (Chen et al., 2024b), COCO Caption (Lin et al., 2014), and WebVid Bain et al. (2021).

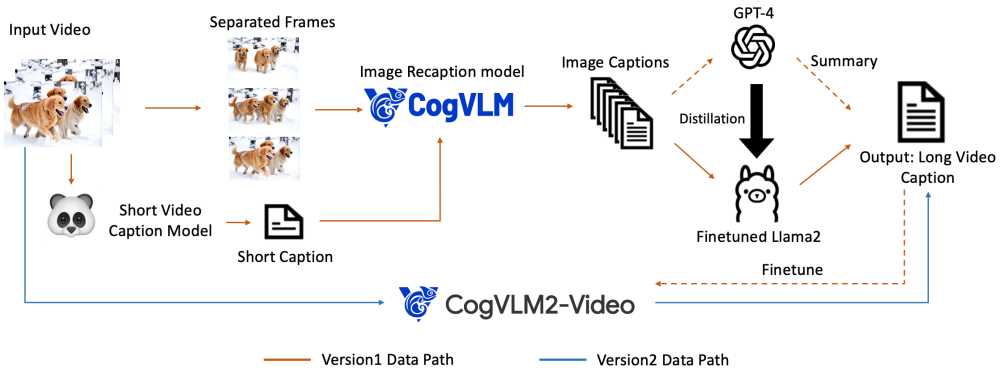


Figure 7: The pipeline for dense video caption data generation. In this pipeline, we generate short video captions with the Panda70M model, extract frames to create dense image captions, and use GPT-4 to summarize these into final video captions. To accelerate this process, we fine-tuned a Llama 2 model with the GPT-4 summaries.

However, the captions in these datasets are usually very short and fail to describe the video's content comprehensively.

To generate high-quality video caption data, we establish a *Dense Video Caption Data Generation* pipeline, as detailed in Figure 7. The idea is to generate video captions from image captions.

First, we use the Panda70M video captioning model (Chen et al., 2024b) to generate short captions for the videos. Then, we employ the image recaptioning model CogVLM (Wang et al., 2023a) used in Stable Diffusion 3 (Esser et al., 2024) and CogView3 (Zheng et al., 2024a) to create dense image captions for each of the frames within a video. Subsequently, we use GPT-4 to summarize all the image captions to produce the final video caption. To accelerate the generation from image captions to video captions, we fine-tune a Llama2 model (Touvron et al., 2023) using the summary data generated by GPT-4 (Achiam et al., 2023), enabling large-scale video caption data generation. Additional details regarding the video caption data generation process can be found in Appendix C.

The pipeline above generates the caption data that is used to trained the CogVideoX model introduced in this report. To further accelerate video recaptioning, we also fine-tune an end-to-end video understanding model CogVLM2-Caption, based on the CogVLM2-Video[1] and Llama3 (AI@Meta, 2024), by using the dense caption data generated from the aforementioned pipeline. The video caption data generated by CogVLM2-Caption is used to train the next generation of CogVideoX. Examples of video captions generated by this end-to-end CogVLM2-Caption model are shown in Appendix D. In Appendix E, we also present some examples of video generation where a video is first input into CogVLM2-Caption to generate captions, which are then used as input for CogVideoX to generate new videos, effectively achieving video-to-video generation.

4 Empirical Evaluation

In this section, we present the performance of CogVideoX through two primary methods: *automated metric evaluation* and *human assessment*. We train several CogVideoX models with different parameter sizes. The following evaluation defaults to using our largest model, which can also be accessed via the website `https://ChatGLM.cn` or via APIs at `https://bigmodel.cn`.

To facilitate the development of text-to-video generation, we open-source the model weight of CogVideoX-2B at `https://github.com/THUDM/CogVideo`.

[1]The CogVLM2-Video model weight is openly available at `https://github.com/THUDM/CogVLM2`.

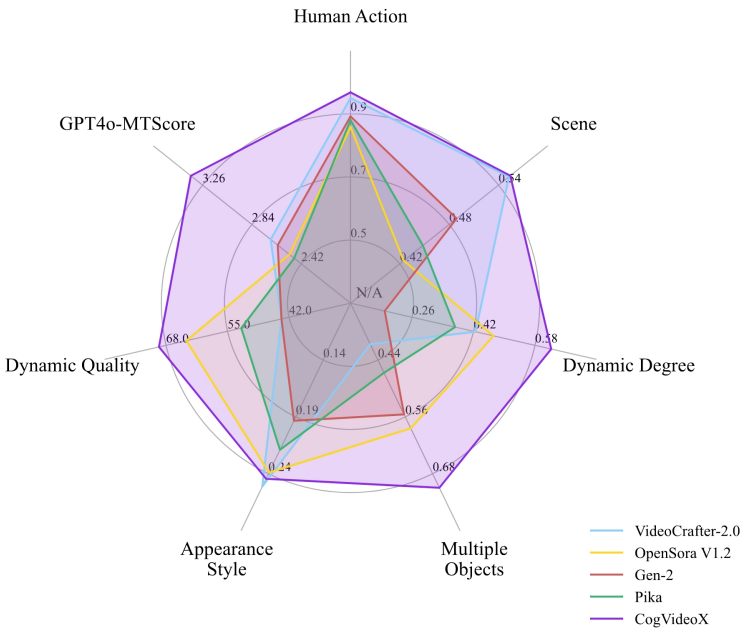


Figure 8: The radar chart comparing the performance of different models. CogVideoX represents the largest one.

4.1 Automated Metric Evaluation

Baselines. We choose openly-accessible top-performing text-to-video models as baselines, including T2V-Turbo (Li et al., 2024), AnimateDiff (Guo et al., 2023), VideoCrafter2 (Chen et al., 2024a), OpenSora (Zheng et al., 2024b), Show-1 (Zhang et al., 2023a), Gen-2 (runway, 2023a), Pika (runway, 2023b), and LaVie-2 (Wang et al., 2023b).

Evaluation Metrics. To evaluate the text-to-video generation, we employed several metrics from VBench (Huang et al., 2024): *Human Action*, *Scene*, *Dynamic Degree*, *Multiple Objects*, and *Appearance Style*. VBench is a suite of tools designed to automatically assess the quality of generated videos. We have selected certain metrics from VBench, excluding others that do not align with our evaluation needs. For example, the color metric, intended to measure the presence of objects corresponding to specific colors across frames in the generated video, assesses the model's quality by calculating the probability. However, this metric may mislead video generation models that exhibit greater variation, thus it is not to include it in our evaluation.

For longer-generated videos, some models might produce videos with minimal changes between frames to obtain higher scores, but these videos lack rich content. Therefore, a metric for evaluating the dynamism of the video becomes more important. To address this, we employ two video evaluation tools: *Dynamic Quality* from Devil (Liao et al., 2024) and *GPT4o-MTScore* from ChronoMagic (Yuan et al., 2024), which focus more on the dynamic characteristics of videos. Dynamic Quality is defined by the integration of various quality metrics with dynamic scores. This approach mitigates biases arising from negative correlations between video dynamics and video quality, leading to a more thorough assessment of video quality. ChronoMagic, for instance, introduces GPT4o-MTScore, a metric designed to measure the metamorphic amplitude of time-lapse videos, such as those depicting physical, biological, and meteorological changes. This metric is obtained by extracting frames from the generated videos at regular intervals and using GPT-4o (OpenAI, 2024) to score the degree of change, providing a fine-grained assessment of video dynamism. This method ensures a more accurate evaluation of the content's variability over time, countering the potential bias of static frame sequences in scoring.

Results. Table 1 provides the performance comparison of CogVideoX and other models. CogVideoX achieves the best performance in five out of the seven metrics and shows competitive results in the remaining two metrics. These results demonstrate that the model not only excels in

Table 1: Evaluation results.

Models	Human Action	Scene	Dynamic Degree	Multiple Objects	Appearance Style	Dynamic Quality	GPT4o-MT Score
T2V-Turbo	95.2	**55.58**	49.17	54.65	24.42	–	–
AnimateDiff	92.6	50.19	40.83	36.88	22.42	–	2.62
VideoCrafter-2.0	95.0	55.29	42.50	40.66	**25.13**	43.6	2.68
OpenSora V1.2	85.8	42.47	47.22	58.41	23.89	63.7	2.52
Show-1	95.6	47.03	44.44	45.47	23.06	57.7	–
Gen-2	89.2	48.91	18.89	55.47	19.34	43.6	2.62
Pika	88.0	44.80	37.22	46.69	21.89	52.1	2.48
LaVie-2	96.4	49.59	31.11	64.88	25.09	–	2.46
CogVideoX	**96.8**	55.44	**62.22**	**70.95**	24.44	**69.5**	**3.36**

video generation quality but also outperforms previous models in handling various complex dynamic scenes. In addition, Figure 8 presents a radar chart that better illustrates the performance advantages of CogVideoX.

4.2 Human Evaluation

In addition to automated scoring mechanisms, a comparative analysis between the Kling (Team, 2024) and CogVideoX is conducted using a manual scoring system. One hundred meticulously crafted prompts are used for human evaluators, characterized by their broad distribution, clear articulation, and well-defined conceptual scope. We randomize videos for blind evalution. A panel of evaluators is instructed to assign scores for each detail on a scale from zero to one, with the overall total score rated on a scale from zero to five, where higher scores reflect better video quality. Reasons for any score deductions are also carefully documented. The results shown in Table 2 indicate that CogVideoX wins the human preference over Kling across all aspects. More details about human evaluation are shown in F .

Table 2: Human evaluation between CogVideoX and Kling.

Model	Sensory Quality	Instruction Following	Physics Simulation	Cover Quality	Total Score
Kling	0.638	0.367	0.561	0.668	2.17
CogVideoX	**0.722**	**0.495**	**0.667**	**0.712**	**2.74**

5 Conclusion

In this paper, we present CogVideoX, a state-of-the-art text-to-video diffusion model. It leverages a 3D VAE and an Expert Transformer architecture to generate coherent long-duration videos with significant motion. By implementing a comprehensive data processing pipeline and a video re-captioning method, we significantly improve the quality and semantic alignment of the generated videos. Our progressive training techniques, including mixed-duration training and resolution progressive training, further enhance the model's performance and stability. Our ongoing efforts focus on refining the CogVideoX's ability to capture complex dynamics and ensure even higher quality in video generation. We are also exploring the scaling laws of video generation models and aim to train larger and more powerful models to generate longer and higher-quality videos, pushing the boundaries of what is achievable in text-to-video generation.

Acknowledgments

We would like to thank all the data annotators, infra-operating staff, collaborators, and partners as well as everyone at Zhipu AI and Tsinghua University not explicitly mentioned in the report who have provided support, feedback, and contributed to the CogVideoX.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

AI@Meta. Llama 3 model card. 2024. URL `https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md`.

Yushi Bai, Xin Lv, Jiajie Zhang, Yuze He, Ji Qi, Lei Hou, Jie Tang, Yuxiao Dong, and Juanzi Li. Longalign: A recipe for long context alignment of large language models. *arXiv preprint arXiv:2401.18058*, 2024.

Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1728–1738, 2021.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023.

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models, 2024a.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13320–13331, 2024b.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359, 2022.

Mostafa Dehghani, Basil Mustafa, Josip Djolonga, Jonathan Heek, Matthias Minderer, Mathilde Caron, Andreas Steiner, Joan Puigcerver, Robert Geirhos, Ibrahim M Alabdulmohsin, et al. Patch n'pack: Navit, a vision transformer for any aspect ratio and resolution. *Advances in Neural Information Processing Systems*, 36, 2024.

Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12873–12883, 2021.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

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.

Jiachen Li, Weixi Feng, Tsu-Jui Fu, Xinyi Wang, Sugato Basu, Wenhu Chen, and William Yang Wang. T2v-turbo: Breaking the quality bottleneck of video consistency model with mixed reward feedback. *arXiv preprint arXiv:2405.18750*, 2024.

Mingxiang Liao, Hannan Lu, Xinyu Zhang, Fang Wan, Tianyu Wang, Yuzhong Zhao, Wangmeng Zuo, Qixiang Ye, and Jingdong Wang. Evaluation of text-to-video generation models: A dynamics perspective, 2024. URL `https://arxiv.org/abs/2407.01094`.

Shanchuan Lin, Bingchen Liu, Jiashi Li, and Xiao Yang. Common diffusion noise schedules and sample steps are flawed. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 5404–5411, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

OpenAI. Gpt-4o. 2024.

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

runway. Gen-2. 2023a. URL `https://runwayml.com/ai-tools/gen-2-text-to-video`.

runway. Pika beta. 2023b. URL `https://pika.art/home`.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Kuaishou AI Team. Kling. 2024. URL `https://kling.kuaishou.com/en`.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023a.

Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *arXiv preprint arXiv:2309.15103*, 2023b.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, et al. Effective long-context scaling of foundation models. *arXiv preprint arXiv:2309.16039*, 2023.

Lijun Yu, José Lezama, Nitesh B Gundavarapu, Luca Versari, Kihyuk Sohn, David Minnen, Yong Cheng, Agrim Gupta, Xiuye Gu, Alexander G Hauptmann, et al. Language model beats diffusion–tokenizer is key to visual generation. *arXiv preprint arXiv:2310.05737*, 2023.

Shenghai Yuan, Jinfa Huang, Yongqi Xu, Yaoyang Liu, Shaofeng Zhang, Yujun Shi, Ruijie Zhu, Xinhua Cheng, Jiebo Luo, and Li Yuan. Chronomagic-bench: A benchmark for metamorphic evaluation of text-to-time-lapse video generation. *arXiv preprint arXiv:2406.18522*, 2024.

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

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023b.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

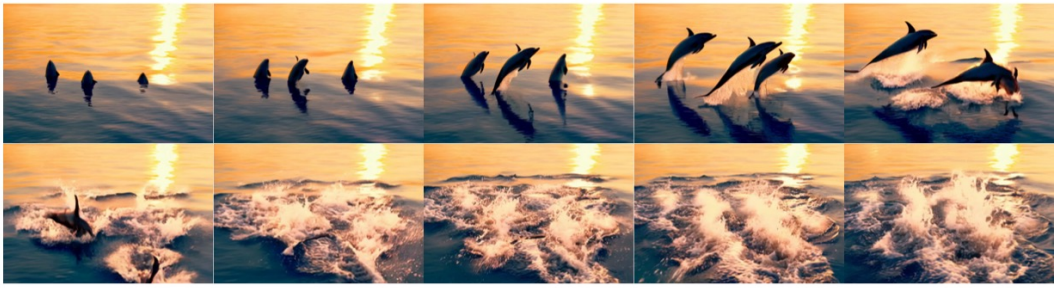
Wendi Zheng, Jiayan Teng, Zhuoyi Yang, Weihan Wang, Jidong Chen, Xiaotao Gu, Yuxiao Dong, Ming Ding, and Jie Tang. Cogview3: Finer and faster text-to-image generation via relay diffusion. *arXiv preprint arXiv:2403.05121*, 2024a.

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, March 2024b. URL `https://github.com/hpcaitech/Open-Sora`.

Text Prompt: A few golden retrievers playing in the snow

Text Prompt: A big stone on the mountain, suddenly, a bolt of lightning in the clear sky struck the stone, a Monkey King monkey dressed in battle robes jumped out of the stone cracks, the stone scattered all around, and then in the sky erupted a strong energy fluctuation, tugging at the air to force people, the film wind, the camera advances

Text Prompt: Three dolphins leap out of the ocean at sunset, then splash into the water



Text Prompt: The camera rotates around a stack of vintage televisions that show a variety of programs - 1950s sci-fi movies, horror movies, news, stills, 1970s sitcoms, etc. - set in a large gallery at the New York Museum.

Text Prompt: Mushroom turns into a bear

Figure 9: Text to video showcases. The displayed prompt will be upsampled before being fed into the model. The generated videos contain large motion and can produce various video styles.

Text Prompt: Push upward at a low angle, slowly look up, an evil dragon suddenly appears on the iceberg, and then the dragon spots you and rushes towards you. Hollywood movie style

Text Prompt: An old-fashioned automobile drives through the streets of the Republic. While driving right in the middle of it, bombs suddenly fall from the sky, the car is blown up, the people in the car are blown up, the screen shakes, the movie winds up

Text Prompt: A man running in the snow

Text Prompt: The camera follows behind a white vintage SUV with a black roof rack as it speeds up a steep dirt road surrounded by pine trees on a steep mountain slope, dust kicks up 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 are redwoods, with patches of greenery scattered throughout. The car is seen from the rear following the curve with ease, making it seem as if it is on a rugged drive through the rugged terrain. The dirt road itself is surrounded by steep hills and mountains, with a clear blue sky above with wispy clouds.

Text Prompt: In the cafe by the window, a man in a suit sits at the table and slowly raises his coffee to sip it, his eyes looking out the window, the street is full of traffic, the man is in deep thought.

Figure 10: Text to video showcases.

A Image To Video Model

We finetune an image-to-video model from the text-to-video model. Drawing from the (Blattmann et al., 2023), we add an image as an additional condition alongside the text. The image is passed through 3D VAE and concatenated with the noised input in the channel dimension. Similar to super-resolution tasks, there is a significant distribution gap between training and inference(the first frame of videos vs. real-world images). To enhance the model's robustness, we add large noise to the image condition during training. Some examples are shown in 11, 12. CogVideoX can handle different styles of image input.

Figure 11: Image to video showcases. The displayed prompt will be upsampled before being fed into the model.

Text Prompt: An elephant slowly walks out of a cloud of fog, the fog shatters and flows

Text Prompt: A cat walks through water, breaking the surface, splashing water, light particles randomly fly, flowers and leaves sway

Text Prompt: A girl lowers her head and rubs her face against a puppy, the puppy looks up at the girl

Text Prompt: A woman presses a camera shutter, her hair flying

Text Prompt: A puppy closes its eyes, opens its mouth, and turns its head to bark

Figure 12: Image to video showcases.

B Caption Upsampler

To ensure that text input distribution during inference is as close as possible to the distribution during training, similar to (Betker et al., 2023), we use a large language model to upsample the user's input during inference, making it more detailed and precise. Finetuned LLM can generate better prompts than zero/few-shot.

For image-to-video, we use the vision language model to upsample the prompt, such as GPT4V, CogVLMWang et al. (2023a).

Zero-shot prompt for Text Upsampler

```
You are part of a team of bots that create videos. You work
with an assistant bot that will draw anything you say in
square brackets. For example, outputting \" a beautiful
morning in the woods with the sun peaking through the
trees \" will trigger your partner bot to output a video
of a forest morning, as described. You will be prompted
by people looking to create detailed, amazing videos.
The way to accomplish this is to take their short prompts
and make them extremely detailed and descriptive.
There are a few rules to follow :
You will only ever output a single video description
per user request.
When modifications are requested, you should not simply
make the description longer. You should refactor the
entire description to integrate the suggestions.
```

C Dense Video Caption Data Generation

In the pipeline for generating video captions, we extract one frame every two seconds for image captioning. Ultimately, we collected 50,000 data points to fine-tune the summary model. Below is the prompt we used for summarization with GPT-4:

Prompt for GPT-4 Summary

```
We extracted several frames from this video and described
each frame using an image understanding model,  stored
in the dictionary variable ‘image_captions: Dict[str: str]‘.
In ‘image_captions‘,  the key is the second at which the image
appears in the  video,  and the value is a detailed description
of the image at that moment. Please describe the content of
this video  in as much detail as possible,  based on the
information  provided by ‘image_captions‘,  including
the objects, scenery, animals, characters, and camera
movements within the video. \n image_captions={new_captions}\n
You should output your summary directly,  and not mention
variables like ‘image_captions‘ in your response.
Do not include ‘\\n’ and the word ’video’ in your response.
Do not use introductory phrases such as: \"The video
presents\", \"The video depicts\", \"This video showcases\",
\"The video captures\" and so on.\n Please start the
description with the video content directly, such as \"A man
first sits in a chair, then stands up and walks to the
kitchen....\"\n Do not use phrases like: \"as the video
progressed\" and \"Throughout the video\".\n Please describe
```

```
the content of the video and the changes that occur, in
chronological order.\n Please keep the description of this
video within 100 English words.
```

D Video Caption Example

Below we present some examples to compare the performance of the Panda-70M video captioning model and our CogVLM2-Caption model:

Caption Generated by Panda-70M

```
A close up of a white dragon with blue eyes.
```

Caption Generated by CogVLM2-Caption

```
A meticulously crafted white dragon with a serene expression
and piercing blue eyes is showcased in various close-ups, its
scales detailed with golden highlights. Initially, the dragon's
horns are curved back, but as time passes, they change to
spiraling gold, adding elegance. The creature's demeanor shifts
from tranquil majesty to a more formidable presence, with its
mouth agape revealing sharp teeth at times. Its expressions
alternate between calmness and fierce determination,
maintaining an air of mythical serenity throughout. The
consistent neutral background focuses attention on the dragon's
textured features and expressive eyes.
```

Caption Generated by Panda-70M

```
A crab is walking on the beach with a light bulb on its back.
```

Caption Generated by CogVLM2-Caption

```
A hermit crab with a dark, glossy shell and reddish-brown legs
is seen carrying an illuminated light bulb on its back across
the sandy terrain of a beach at night. The scene transitions
from a soft glow to a more pronounced illumination as the crab
moves, highlighting its delicate limbs against the smooth sand
```

and tranquil sea backdrop. This surreal tableau blends natural beauty with human ingenuity, creating a serene yet whimsical atmosphere that emphasizes the crab's unique adaptation and the contrast between nature and technology in this quiet nocturnal setting.

Caption Generated by Panda-70M

A young black man is sitting on a cloud and reading a book with a blue sky in the background.

Caption Generated by CogVLM2-Caption

A young Black man with an afro hairstyle and a neatly trimmed beard is seen sitting cross-legged on fluffy white clouds, deeply engrossed in reading a book with a red cover. He wears a plain white T-shirt and dark pants against a vivid blue sky dotted with cumulus clouds. Throughout the scenes, his expression remains one of deep concentration and peaceful contemplation, highlighting a moment of intellectual pursuit amidst nature's grandeur. The imagery suggests a serene atmosphere that emphasizes solitude and introspection, with no other people or objects around him.

E Video to Video via CogVideoX and CogVLM2-Caption

In this section, we present several examples of video-to-video generation using CogVideoX and CogVLM2-Caption. Specifically, we first input the original video into CogVLM2-Caption to obtain the video's caption, and then feed this caption into the CogVideoX model to generate a new video. From the examples below, it can be seen that our pipeline achieves a high degree of fidelity to the original video:

Source Video

Caption Generated by CogVLM2-Caption

A man dressed as a wizard, with a blue robe adorned with white stars and a matching pointed hat, stands in a dark cave. He is engaged in casting spells from an open book, surrounded by mystical flames that illuminate the scene. Throughout the sequence, his right hand is raised to channel bright lightning bolts towards unseen targets, while his left hand holds the book firmly. His expression remains focused and serious, suggesting deep concentration on his magical endeavors. The atmosphere of mystery and ancient magic is enhanced by the surrounding darkness and the vivid display of light and shadow.

Video Generated by CogVideoX

Source Video

Caption Generated by CogVLM2-Caption

A picturesque evening descends on a cliffside village, showcasing whitewashed buildings with blue domes that glow against the darkening sky. The Aegean Sea mirrors this celestial hue, creating a serene tableau devoid of people and vehicles. As time passes, the scene remains tranquil, illuminated by golden lights from within homes and lit pathways weaving between structures. A solitary windmill stands out, symbolizing local culture amidst the peaceful setting. The absence of visible human activity emphasizes the stillness and beauty of the coastal hamlet, inviting contemplation in its embrace.

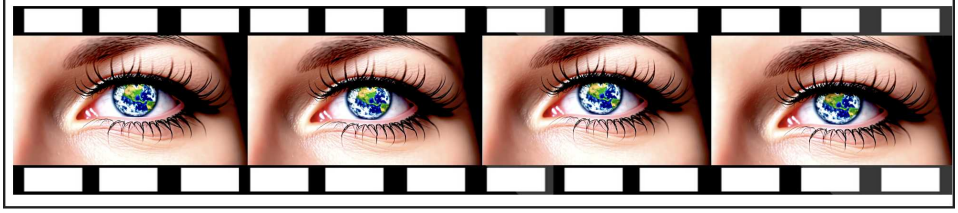
Video Generated by CogVideoX

Source Video

Caption Generated by CogVLM2-Caption

A woman's eye, in sharp focus and detailed with a bold black eyeliner, reflects the Earth. The vivid colors of blue oceans and green continents stand out against her clear iris, symbolizing a deep connection between humanity and our planet. Her expression remains neutral throughout, emphasizing introspection or awareness. As time passes, the reflection subtly shifts to include parts of Africa and Europe, suggesting a global perspective. The contrast between her dark eyelashes and light skin accentuates the visual metaphor for unity and interconnectedness, while her gaze suggests contemplation on environmental issues or a profound sense of responsibility towards the world.

Video Generated by CogVideoX



F Human Evaluation Details

Sensory Quality: This part focuses mainly on the perceptual quality of videos, including subject consistency, frame continuity, and stability.

Instruction Following: This part focuses on whether the generated video aligns with the prompt, including the accuracy of the subject, quantity, elements, and details.

Physics Simulation: This part focuses on whether the model can adhere to the objective law of the physical world, such as the lighting effect, interactions between different objects, and the realism of fluid dynamics.

Cover Quality: This part mainly focuses on metrics that can be assessed from single-frame images, including aesthetic quality, clarity, and fidelity.