Mimir: Improving Video Diffusion Models for Precise Text Understanding

Supplementary Material

In the main paper, we provide a method diagram and textual description of `Mimir`. Here, we present the detailed pseudocode of the **Token Fuser** in `Mimir` in Algorithm 1 for direct reference. In the following sections, We introduce the data processing in Sec. B, the evaluation metric in Sec. C, the user study in Sec. D, and additional experimental results in Sec. E. We also introduce limitations and social impact of our work in Sec. F.5 and Sec. F.6 respectively. Training requires 64 A100 for about 5 days.

A. Implement Details

We select the widely used CogVideoX-5B [11] as our baseline and adopted the same implementation settings, such as v-prediction (*i.e.*, $\alpha_t\epsilon - \delta_t x$), zero SNR (*i.e.*, rescales betas to have zero terminal SNR), the DDIMScheduler with std_dev_t=0 without randomness, and standard loss function ($\left\|\epsilon - \epsilon_\theta\left(\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t\right)\right\|^2$).

B. Data Processing

We construct a collection of relatively high-quality video clips with text descriptions using a combination of video filtering and recaptioning models. As shown in Fig. 1, the collected data undergoes multiple filtration steps: Basic Filtration, Quality Filtration, Aesthetic Filtration, Watermark Filtration, which removes data that does not meet fundamental requirements. After these video-based filtration steps, captions are generated for the videos. The videos and their captions are then evaluated for consistency to ensure the caption accurately describes the video content. Following this process, approximately 500,000 single-shot clips remain, with each clip averaging about 10 seconds. These high-quality video clips are ultimately used for training Mimir. Next, we provide a detailed explanation of each stage of this pipeline.

Basic Filtration. At this stage, we focus on computing video metadata and filtering out invalid videos.

1. *Metadata Extraction:* Most of important video properties such as length, width, frame rate, frame count, and duration are obtained and saving using FFmpeg.
2. *Filtering Rules:*
   - Videos with fewer than 65 frames, a duration of less than 1s, or an aspect ratio (width / height) outside the range [1, 2] are excluded.
   - Videos with a motion score of 0, determined using optical flow, are excluded.

Quality Filtration. At this stage, we calculate basic quality indicators for the videos and remove those that do not meet the standards.

Algorithm 1 Token Fuser

```
# Inputs

# Text prompt provided by the user
text_prompt = "Input␣text␣prompt"
# Instructional input for fine-tuning
instruction_prompt = "Instruction␣description"

# 1. Encoding and Tokenization

# Obtain token embeddings from text encoder
e_theta = TextEncoder(text_prompt)
# Obtain token embeddings from decoder-only model
e_beta = DecoderModel(text_prompt)
# Obtain instruction token from decoder-only model
e_i = DecoderModel(instruction_prompt)

# 2. Token Fusion to Address Feature Distribution
    Gap
# Normalize e_beta and apply learnable scale
# Apply normalization
e_beta = Normalize(e_beta)
# Scale normalized features
e_beta = LearnableScale(e_beta)

# Apply Zero-Conv to e_beta and e_theta to maintain
    original semantic space
# Maintain semantic space for e_beta
e_beta = ZeroConv(e_beta)
# Maintain semantic space for e_theta
e_theta = e_theta + ZeroConv(e_theta)

# Sum modified tokens to form combined tokens
e = e_theta + e_beta # Shape: [n, 4096]

# 3. Stabilizing Divergent Semantic Features
# Initialize learnable tokens and add to instruction
    tokens
# Four learnable tokens, shape: [4, 4096]
e_l = InitializeLearnableTokens(count=4, dim=4096)
# Stabilize instruction features
e_s = e_i + e_l

# 4. Final Token Fusion and Video Generation
# Concatenate e_combined and stabilized tokens
e_final = Concatenate(e_combined, e_stabilized) #
    Shape: [n+4, 4096]

# Generate videos using the final fused tokens
generated_video = VideoGenerator(e_final)

# Output
return generated_video
```

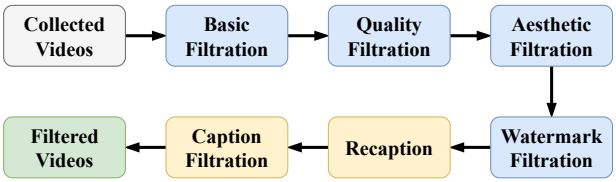


Figure 1. The pipeline for preparing data.

1. *Quality Metrics:* We use OpenCV to calculate the black area percentage, brightness, and black frame rate.
2. *Filtering Rules:*
   - Black area > 0.8, excluding.
   - Brightness < 0.2, excluding.
   - Black frame rate > 0.4, excluding.

Aesthetic Filtration. At this stage, we filter videos based on aesthetic-related operators.

1. *Aesthetic Metrics:* We use the aesthetic predictor [*] to calculate aesthetic score and OCR coverage.
2. *Filtering Rules:*
   - Aesthetic score < 4.0, excluding.
   - OCR coverage > 0.1, excluding.

Watermark Filtration. At this stage, videos containing watermarks are excluded. Each video is analyzed using QWen2-VL-7B [1] to detect the presence of watermarks. Videos flagged as "*containing watermarks*" are excluded.

Re-Caption. At this stage, we use CogVim2 [3, 8] to generate captions, which produces semantic and detailed descriptions of visual contents in videos.

Caption Filtration. Due to hallucinations in large language models, not all output captions are immediately usable. To address this, we employ human designed rule-based methods and text quality metrics to clean the captions.

1. *Text Quality Metrics:*
   - N-gram [†] repetition rates
   - Semantic alignment between the video and the generated caption using CLIP Score.
2. *Filtering Rules:*
   - 2-gram repetition > 0.056, excluding.
   - 5-gram repetition > 0.047, excluding.
   - 10-gram repetition > 0.045, excluding.
   - Semantic consistency (CLIP score) < 0.25, excluding.

This pipeline ensures the collection of high-quality video clips with accurate captions, which are suitable for training.

C. Evaluation Metric

We employ several evaluation metrics in VBench [5] to quantitatively assess our results, including *Background Consistency*, *Aesthetic Quality*, *Imaging Quality*, *Object Class*, *Multiple Objects*, *Color Consistency*, *Spatial Relationship*, and *Temporal Style*. The detailed metrics are introduced as follows:

- *Background Consistency*. This metric evaluates the temporal consistency of background scenes by calculating the similarity of CLIP [7] features across frames.
- *Aesthetic Quality*. This assesses the artistic and aesthetic value perceived by humans for each video frame using the LAION aesthetic predictor. It reflects qualities such as layout, color richness and harmony, photo-realism, naturalness, and overall artistic quality across frames.
- *Imaging Quality*. This measures distortions (e.g., over-exposure, noise, blur) present in generated frames. It is evaluated using the MUSIQ [6] image quality predictor trained on the SPAQ [2] dataset.
- *Object Class*. This metric is computed using GRiT [10] to measure the success rate of generating the specific object classes described in the text prompt.
- *Multiple Objects*. This evaluates the success rate of generating all the objects specified in the text prompt within each video frame. Beyond generating a single object, it assesses the model's ability to compose multiple objects from different classes in the same frame, which is an essential aspect of video generation.
- *Color Consistency*. This measures whether the synthesized object colors align with the text prompt. It uses GRiT [10] for color captioning and compares the results against the expected color.
- *Spatial Relationship*. This metric evaluates whether the spatial relationships in the generated video follow those specified by the text prompt. It focuses on four primary types of spatial relationships and performs rule-based evaluation similar to [4].
- *Temporal Style*. This assesses the consistency of temporal style by using ViCLIP [9] to calculate the similarity between video features and temporal features.

D. User Study

To obtain genuine feedback reflective of practical applications, the 10 participants in our user study experiment come from diverse academic backgrounds. Since many of them do not major in computer vision, we provide detailed explanations for each question to assist their judgments.

- Instruction Following: Determine which video aligns more closely with the prompt, evaluate whether the main content is adequately presented in the video, and assess the accuracy and completeness of the prompt.
- Physics Simulation: Determine which video aligns more closely with real-world physical laws, including object motion, transformations, and other dynamics.
- Visual Quality: Determine which video has a more harmonious overall visual composition and showcases finer details more exquisitely.

E. Additional Experimental Results

E.1. Short / Long Prompt

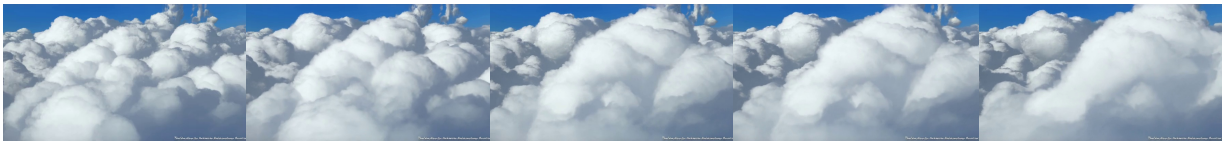
To investigate the performance differences of `Mimir` when inputting short and coarse prompts versus long and fine prompts, we randomly sampled 4 prompts from the VBench dataset. Additionally, VBench provides enhanced versions of these 4 prompts through a large language model. We

[*] https://github.com/christophschuhmann/improved-aesthetic-predictor
[†] https://github.com/EurekaLabsAI/ngram

Short & Coarse Prompt: A raccoon dressed in suit playing the trumpet, stage background.

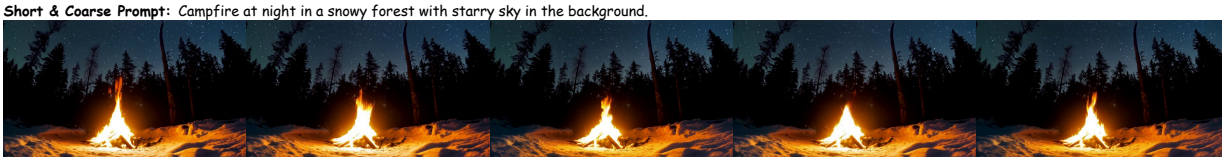
Long & Fine Prompt: A dapper raccoon, dressed in a perfectly tailored black suit with a crisp white shirt and a red bow tie, stands center stage under a spotlight. The stage background is adorned with rich, velvet curtains in deep burgundy, creating an elegant ambiance. The raccoon, holding a gleaming golden trumpet, begins to play, its tiny paws expertly pressing the valves. The raccoon's eyes are closed, lost in the music, as the sound of the trumpet fills the air. The stage lights cast a warm glow, highlighting the raccoon's expressive face and the polished brass of the trumpet, creating a captivating and whimsical performance.

Short & Coarse Prompt: An animated painting of fluffy white clouds moving in sky.



Long & Fine Prompt: A mesmerizing animated painting depicts fluffy white clouds drifting gracefully across a vibrant blue sky. The scene begins with a close-up of the clouds, their soft edges and varying shades of white creating a sense of depth and texture. As the camera pans out, the sky's rich blue hues become more prominent, contrasting beautifully with the clouds. The clouds move slowly and fluidly, their shapes constantly shifting and morphing, evoking a sense of calm and tranquility. Occasionally, a gentle breeze causes the clouds to stretch and elongate, adding a dynamic element to the serene atmosphere. The overall effect is a captivating blend of art and animation, bringing the sky to life in a soothing and visually stunning display.

Short & Coarse Prompt: Campfire at night in a snowy forest with starry sky in the background.



Long & Fine Prompt: In a serene, snow-covered forest, a crackling campfire casts a warm, golden glow, illuminating the surrounding trees and creating a cozy haven amidst the cold. The night sky above is a breathtaking tapestry of countless stars, twinkling brightly against the deep, velvety blackness. Snowflakes gently fall, adding a touch of magic to the scene. The firelight dances on the snow, creating a mesmerizing interplay of light and shadow. The air is crisp and still, with only the soft crackle of the fire and the occasional rustle of the trees breaking the silence. The scene exudes tranquility and wonder, capturing the essence of a peaceful winter night under the stars.

Short & Coarse Prompt: Motion colour drop in water, ink swirling in water, colourful ink in water, abstraction fancy dream cloud of ink.

Long & Fine Prompt: Vibrant swirls of ink cascade into crystal-clear water, creating an ethereal dance of colors. Rich blues, fiery reds, and lush greens intertwine, forming intricate patterns that resemble a dreamlike cloud. The ink moves gracefully, expanding and contracting, as if alive, creating mesmerizing abstract shapes. Each droplet bursts into a myriad of hues, blending seamlessly into one another, evoking a sense of fluid motion and boundless creativity. The scene is a hypnotic display of color and movement, capturing the essence of a fanciful dreamscape where imagination knows no bounds.

Figure 2. The comparison between results with short & course prompts and long & fine prompts.

Figure 3. More examples in terms of color rendering.

input both versions into `Mimir` and generated corresponding videos. As shown in Fig. 2, leveraging the reasoning ability of the decoder-only LLM, even with short and coarse prompts, `Mimir` can generate results as detailed as those produced with long and fine prompts. This demonstrates that `Mimir`'s token fuser effectively expands the semantic, leading to precise text understanding capabilities.

E.2. More Interesting Prompts

E.2.1. Spatial Semantic Understanding

Color Rendering. As shown in Fig. 3, our method demonstrates the ability to accurately understand the color

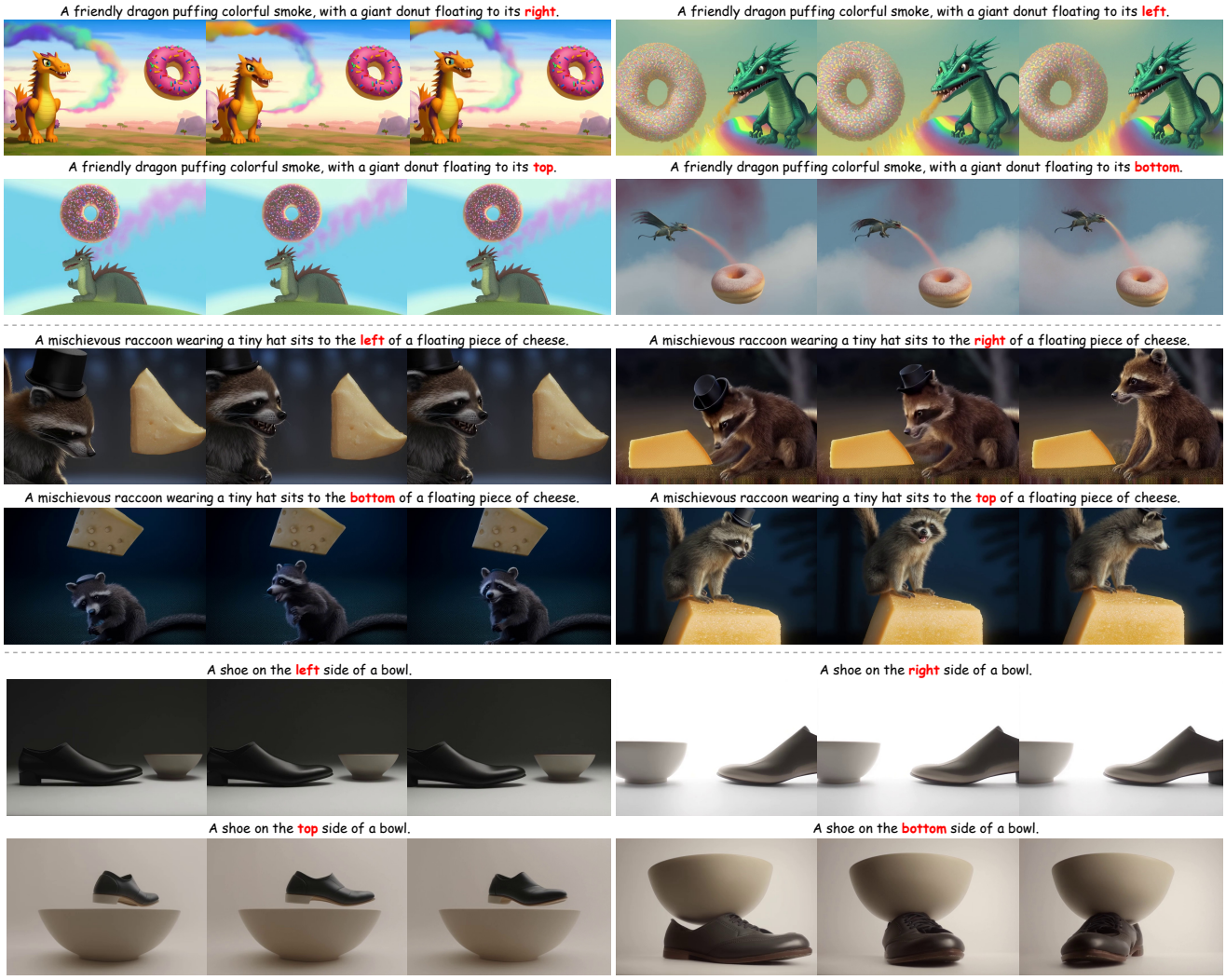


Figure 4. More examples in terms of absolute & relative position.

specifications in the prompt for different objects and generates videos containing objects with the correct colors. It highlights the effectiveness of our token fuser in ensuring semantic alignment between the input prompt and the generated video. By accurately capturing and representing color details, `Mimir` delivers coherent results, even in cases where multiple objects with distinct colors are specified.

Absolute & Relative Position. As shown in Fig. 4, our method effectively understands the spatial relationships (*i.e.*, the absolute & relative position) specified in the prompt, such as "*top*", "*below*", "*left*", and "*right*" and generates videos where objects are positioned correctly according to these relationships. By accurately representing spatial arrangements, `Mimir` ensures that the generated videos meet the semantic requirements of complex prompts involving positional relationships between objects.

Counting. As shown in Fig. 5, `Mimir` demonstrates a strong ability to understand counting. For example, if the prompt specifies a certain number of objects, `Mimir` accurately interprets this information and generates videos containing the correct quantity. By successfully handling quantity-specific prompts, `Mimir` proves its reliability in scenarios where precise numeric understanding is critical for video generation tasks.

E.2.2. Temporal Semantic Understanding

Sequential Actions. This involves capturing the sequence of actions performed by an object, such as a cat looking up, then down, or following a more complex pattern like up, down, and up again. It requires precise temporal understanding to maintain the correct order of actions. As shown in Fig. 6, `Mimir` precisely interprets and reproduces these action sequences.

Illumination Harmonization. It means light changes in the environment, such as dawn transitioning to sunrise and then to sunset. As shown in Fig. 7, `Mimir` precisely generates these gradual scene changes, ensuring the illumination

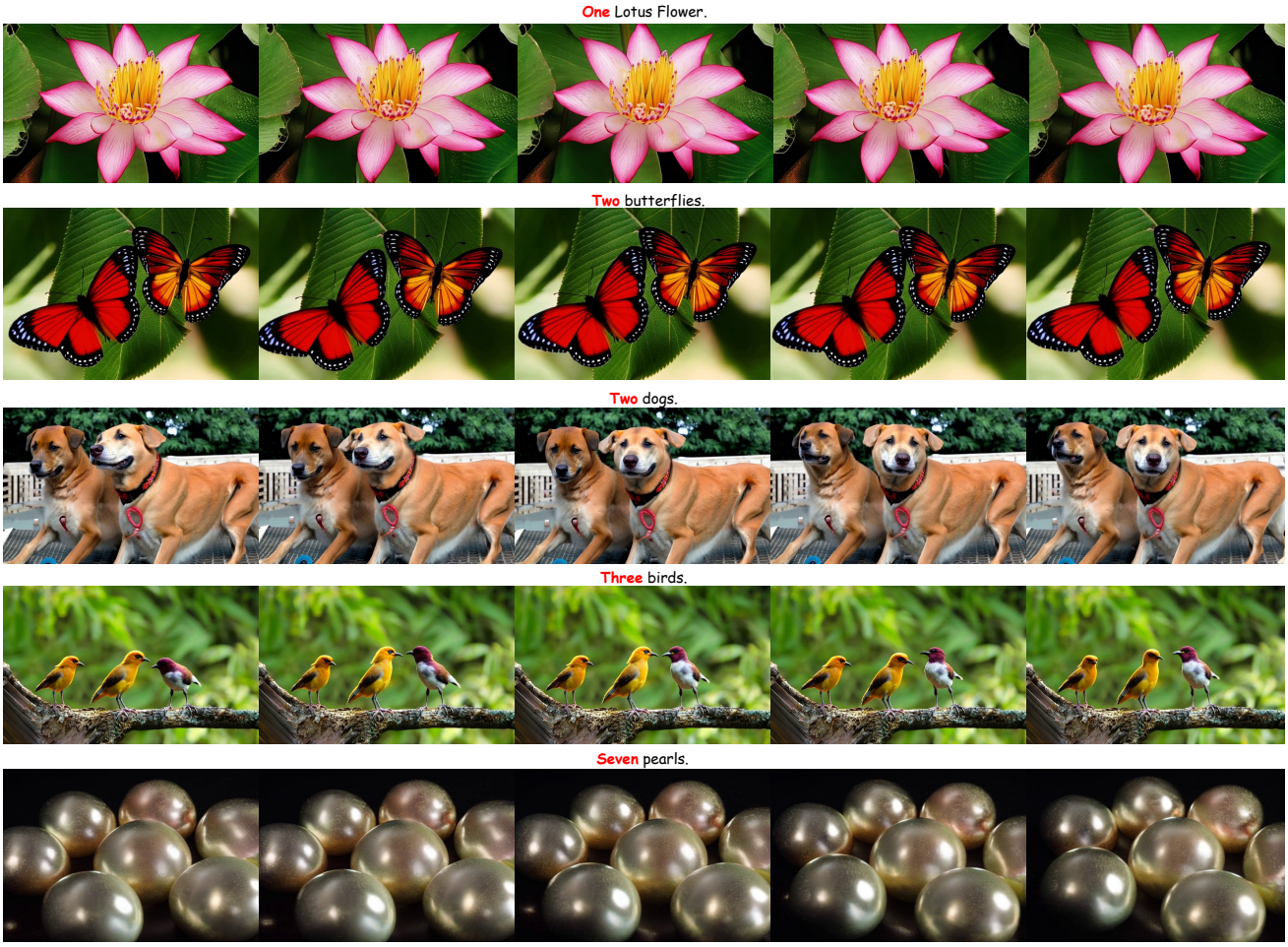


Figure 5. More examples in terms of counting.

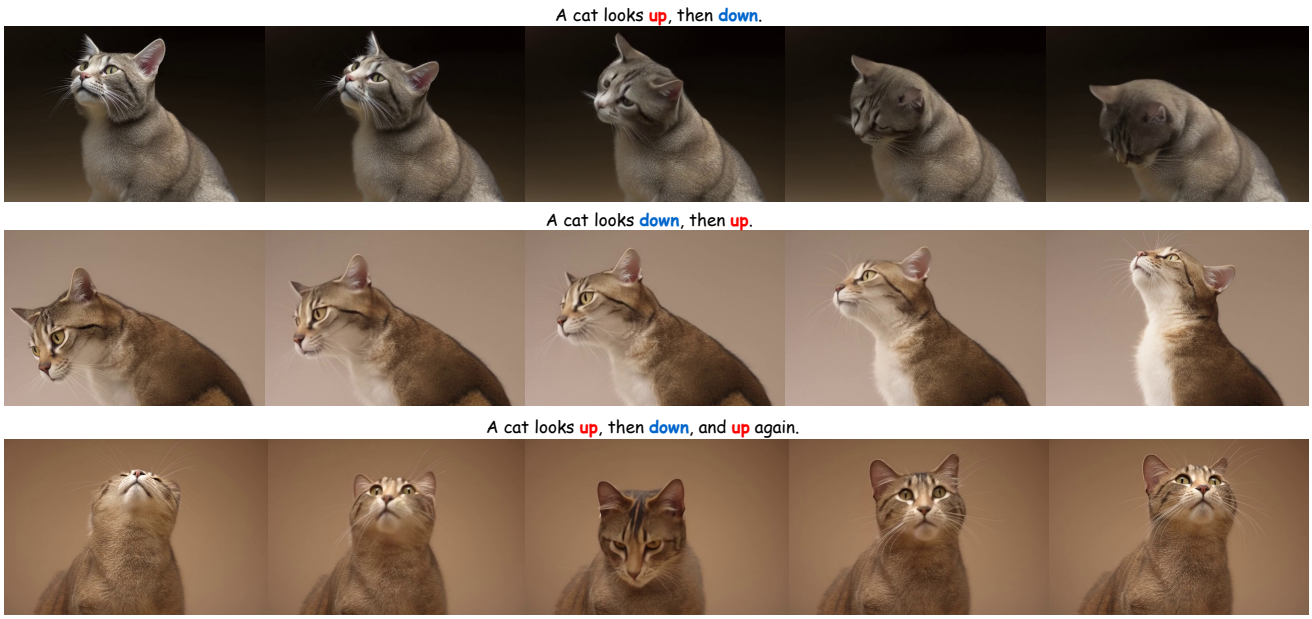
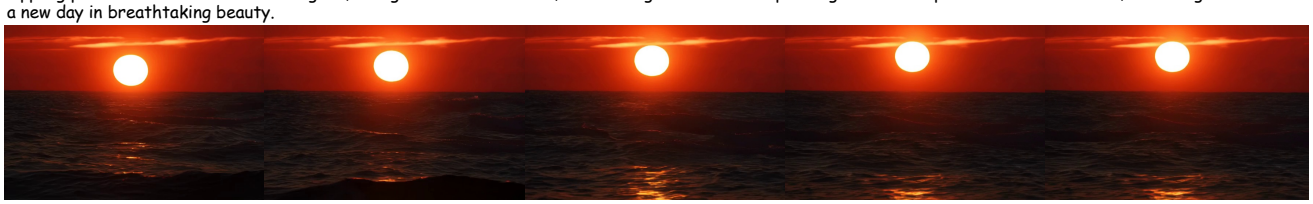


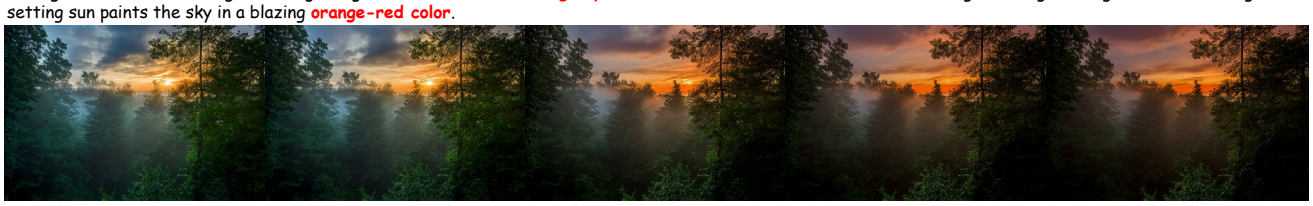
Figure 6. More examples in terms of action sequence over time.

As dawn breaks, the once-vivid stars begin to dim, their brilliance softening as the sky transitions from deep indigo to a pale, serene blue. One by one, the celestial lights vanish, retreating into the vast expanse above. The faint glow of the morning sun brushes the horizon, casting gentle hues of peach and gold. In their place, a tranquil light blue sky emerges, vast and endless, signaling the quiet start of a new day and leaving behind a faint memory of the night.

The horizon glows with a fiery brilliance as a red sun begins its ascent above the calm sea. Its vibrant hue bathes the sky in shades of crimson and amber, casting a warm, ethereal light across the water. The sea, once cloaked in darkness, transforms into a shimmering expanse, reflecting the sun's fiery glow in rippling patterns. As the sun climbs higher, its light floods the world, illuminating the waves and painting the landscape with radiant warmth, heralding the arrival of a new day in breathtaking beauty.

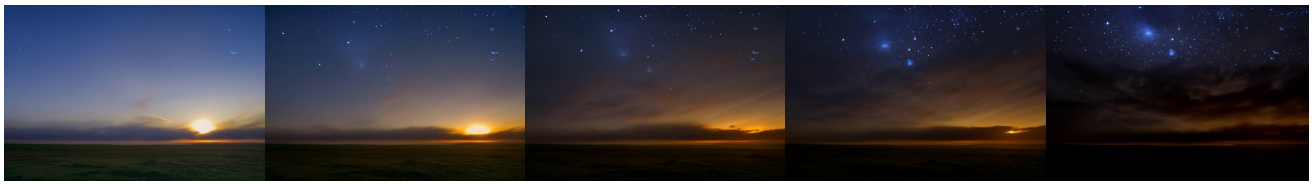


Among the forests, mist lingers among the green trees, and the sunlight penetrates the branches and leaves, shedding bits of golden light. In the evening, the setting sun paints the sky in a blazing orange-red color.



The fields are awakened by the golden sunlight, and a gentle breeze stirs up a green wave. In the evening, the setting sun puts on a coat of fiery red for the earth. At night, the fields are filled with starlight like water, and the Milky Way in the distance quietly guards this peaceful world.

On the vast plains, at dusk, the setting sun colors the clouds into a flaming golden red, and after nightfall, the deep starry sky shines like a jewel, making the whole world seem serene and mysterious.



With a ghostly blue glow, the whole world seems pure and mysterious. The setting sun dyed the sky red, and the ice reflected the warm orange light. At night, the stars are densely packed and the Milky Way crosses the dome of the sky, reflecting the coldness and quietness of the glacier.

Figure 7. More examples in terms of light changes, showcasing the illumination harmonization over time.

harmonization and the alignment with prompts.

Object Transformation. It means transforming an object into another, such as a car transforming into a superhero. This is a highly challenging task due to the complexity of capturing smooth transitions. As shown in Fig. 8, Mimir precisely understands the prompt and generates well.

E.3. More Comparison Results.

Unaligned Training Data. Since we use 500,000 clips that we collect ourselves during training, and in order to

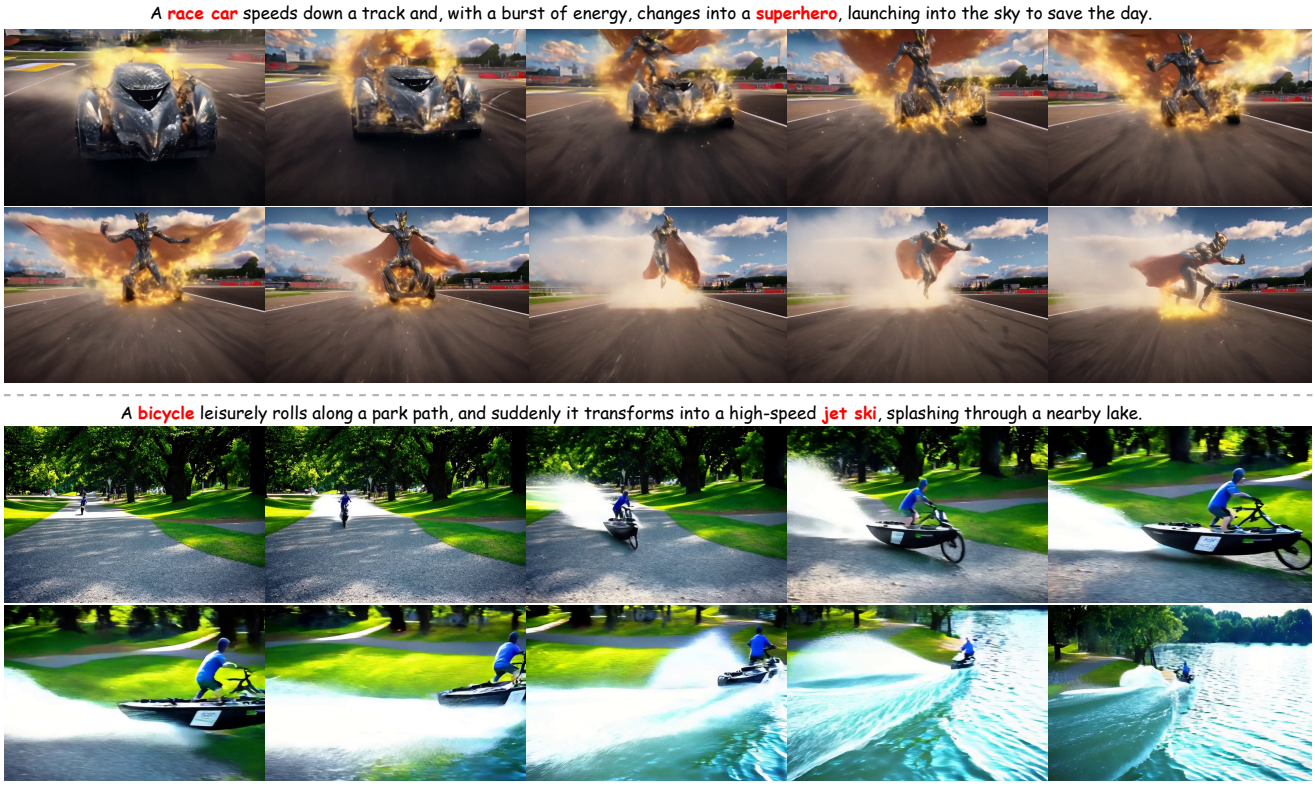


Figure 8. More examples in terms of object transformation over time.

ensure a fair comparison, we train CogVideoX-5B with the same 500,000 clips as our *Baseline*. As shown in the second row of Tab. 1, Mimir (the last row) outperforms the baseline across all metrics, which demonstrates that the improvements stem from the design of our model rather than the data.

Method	Object	Color	Spatial	Temporal
Baseline	90.82%	85.29%	68.25%	25.19%
Mimir	**92.87%**	**86.50%**	**78.67%**	**26.22%**

Table 1. Training with the same training data (500,000 clips).

Extra Information and Fairness. In Mimir, we introduce extra answer tokens and the four instructions to enhance our performance. The additional information is a key contribution of this paper, and our focus is on exploring how to effectively utilize it to enhance video generation. To verify the performance improvement brought by this additional information, we have fairly validated the effectiveness of each module and extra information in Tab. 3 of the main paper. An alternative approach is to directly expand the original prompt with detailed information prompted from the four instructions, and then use the expanded prompt to generate videos. Therefore, we use the same Phi-3.5 to recaption prompts based on the same instructions, and then input them into OpenSora and CogVideoX-5B for evaluation (marked with # in Tab. 2, where "Re." means re-training). As shown in Tab. 2, despite these models having extra information, their performance still lags behind Mimir.

Method	Object	Color	Spatial	Temporal
OpenSora #	90.83%	85.10%	77.89%	24.08%
CogvideoX #	89.75%	85.10%	66.87%	25.63%
Re. CogvideoX #	90.91%	85.67%	66.57%	25.56%
Mimir	**92.87%**	**86.50%**	**78.67%**	**26.22%**

Table 2. Compare using the expanded prompt.

Influence of Extra Params. Because we introduced a decoder-only LLM, which can be regarded as extra parameters, we directly integrated Phi-3.5 into the base model to validate the performance influence of these additional parameters, as shown in the second row of (the main paper's Tab. 3). This integration causes a performance decline because it disrupts the original pre-trained semantic space. Furthermore, we report the efficiency impact: the 480P+49 frames generation time increased from 208.6s to 211.9s (+3.3s), and the checkpoint size increased from 21.51 GB to 29.15 GB (+7.64 GB).

About SANA. SANA addresses the issue of excessively large values in LLM outputs through the normalization. Mimir, developed concurrently, focuses on integrating LLMs into the existing framework in a **non-destructive** manner, prioritizing the preservation of the original model's functionality. By leveraging SANA's normalization (with citation), our ablation studies have confirmed its effective-

Method	Object	Color	Spatial	Temporal
SANA*	2.64%	0%	1.20%	2.80%
Mimir	**92.87%**	**86.50%**	**78.67%**	**26.22%**

Table 3. Compared to SANA.

ness and show that Mimir further enhances performance. When SANA is directly applied to video generation training, the results, as shown in Fig. 9 and Tab. 3, reveal notable limitations.

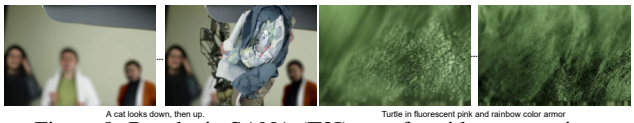


Figure 9. Results in SANA (T2I) way for video generation.

F. Discussion

F.1. Advantages from Decoder-Only Architecture.

Our intention in highlighting the advantages in our paper is to emphasize the strengths of LLMs. Given that decoder-only LLMs currently dominate the NLP field, we discussed them collectively. Regarding T5, as an earlier encoder-decoder LLM, its application in generative models typically uses only its encoder branch for extracting semantic features. It is a relatively simple approach which is incompatible with decoder structures. Our focus is to integrate powerful decoder-only LLMs into diffusion models while enabling collaboration with text encoder like T5 to achieve improved video generation performance.

F.2. LLM Advantages Contributes to T2V.

The LLM advantages (*e.g.,* the imagenation) mentioned in this paper specifically refer to the understanding of objects, colors, motion, spatial relationships, and how they represent in generated videos. We have provided qualitative proofs as follows: Fig. 5 of main paper/Fig. 5 demonstrate the imagine ability to objects, Fig.5 of main paper/Fig. 3 to colors, Fig.5 of main paper/Fig. 4 to spatial relationships, Fig.6 of main paper/Fig. 6 to motions. Besides, Tab. 1 of main paper demonstrates quantitative proofs across these four aspects.

F.3. Component's Effect

We list explanations of 3 core components bellow: *(1) LLM Integration:* By retaining both query and answer tokens, we preserve the imaginative potential of LLMs, enabling precise representation of elements like color and motion in videos. *(2) Non-Destructive Fusion:* The zero convolution prevents training from compromising video quality. *(3) Semantic Stabilizer:* Using fixed instructions to guide LLM responses ensures temporal stability in videos. Further, we have conducted ablation studies on all above components, and results in Tab. 3 of main paper demonstrate their respective contributions to improving generated videos.

F.4. Scalability & Computational.

Our non-destructive Token Fuser ensures Mimir, which is still a diffusion transformer, retains scalability in parameters and dataset size.

F.5. Limitations

While our current work has made significant strides, it also possesses certain limitations. Firstly, the generated videos are typically limited to short durations (a few seconds to tens of seconds). This is primarily due to the significant computational resources and storage requirements needed for generating longer videos. Additionally, extending the video length may exacerbate temporal inconsistencies, such as discontinuities in actions or backgrounds across frames, which can detract from the overall quality and realism. Secondly, the effectiveness of our T2V model is heavily dependent on the quality and diversity of the training data. In domains where the training dataset lacks coverage—such as specific professional scenarios—the model's performance can be suboptimal. This limitation highlights the importance of expanding and diversifying training datasets to improve the model's generalizability across a broader range of applications.

F.6. Social Impact

Our proposed T2V (Text-to-Video) model demonstrates strong potential for generating high-quality, contextually accurate video content directly from textual descriptions. This technology offers significant benefits across various domains, enabling more accessible, creative, and automated video generation workflows. However, like any generative technology, our T2V model also raises concerns about potential misuse. Malicious actors could exploit it to produce deceptive or harmful video content, such as fake news or misleading advertisements, amplifying the spread of misinformation on social media platforms. This misuse could lead to detrimental societal consequences, including the erosion of trust in digital media. Despite ongoing advancements in generative content detection technologies, challenges remain, especially in scenarios involving complex, high-quality synthetic videos. To address this, we are committed to promoting responsible use of T2V technology and actively contributing to the research community. We aim to share our generated results to support the development of more robust detection algorithms, fostering a safer digital environment capable of mitigating the risks associated with increasingly sophisticated generative models.

References

[1] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 2

[2] Yuming Fang, Hanwei Zhu, Yan Zeng, Kede Ma, and Zhou Wang. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3677–3686, 2020. 2

[3] Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. Cogvlm2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024. 2

[4] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023. 2

[5] 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

[6] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021. 2

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 2

[8] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models, 2023. 2

[9] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *arXiv preprint arXiv:2307.06942*, 2023. 2

[10] Jialian Wu, Jianfeng Wang, Zhengyuan Yang, Zhe Gan, Zicheng Liu, Junsong Yuan, and Lijuan Wang. Grit: A generative region-to-text transformer for object understanding. In *European Conference on Computer Vision*, pages 207–224. Springer, 2025. 2

[11] 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. 1