A. Implementation Details

Table 6. The detailed model configurations of Video Tokenizer.

Module	Configuration	Value	#Parameters
Encoder	Transformer Layer	12	85M
	Hidden Size	768	
	Attention Heads	12	
	IFrame Query Tokens	330	
	PFrame Query Tokens	74	
Decoder	Transformer Layer	12	85M
	Hidden Size	768	
	Attention Heads	12	
	Mask Tokens	1	
Quantizer	Codebook size	2048	25.4K
	Codebook Dimension	16	
	Similarity Metric	Cosine	
Total Parameters			170M

Table 7. The detailed model configurations of LanDiff.

Module	Configuration	Value	#Parameters
LLM	Transformer Layer	24	2B
	Hidden Size	2048	
	Attention Heads	16	
	MLP Dimension	11008	
	Activation	GELU	
	RoPE θ	10000	
	Text Drop Rate	0.1	
	Micro Conditioner Hidden Size	512	
Diffusion Backbone Module (Frozen)	Transformer Layer	30	2B
	Attention Heads	8	
	Hidden Size	1920	
	Time Embedding Size	256	
Diffusion Control Module	Transformer Layer	15	1B
	Attention Heads	8	
	Hidden Size	1920	
	Time Embedding Size	256	
Trainable Parameters			3B
Total Parameters			5B

A.1. Details of Video Tokenizer

The video tokenizer model follows a similar structure to TiTok [63]. However, to flexibly encode videos with different numbers of frames, we replace absolute position encoding with 3D RoPE position encoding [50]. We use interpolation of positional encoding to enable the encoder of the Theia [46] model to handle 480x720 resolution videos. On average, for a 480x720 resolution video of one second, our tokenizer generates about 200 tokens. In contrast, common tokenizers such as MagViT2 [62] generate about 10,000 tokens per second for videos of this resolution, and our sequence length is about 1/50 of MagViT2. The model configuration of the tokenizer and the parameters of each part are shown in Table 6. The model is a Transformer structure, with 12 layers in both the encoder and decoder, a hidden layer size of 768, and 12 heads. To improve

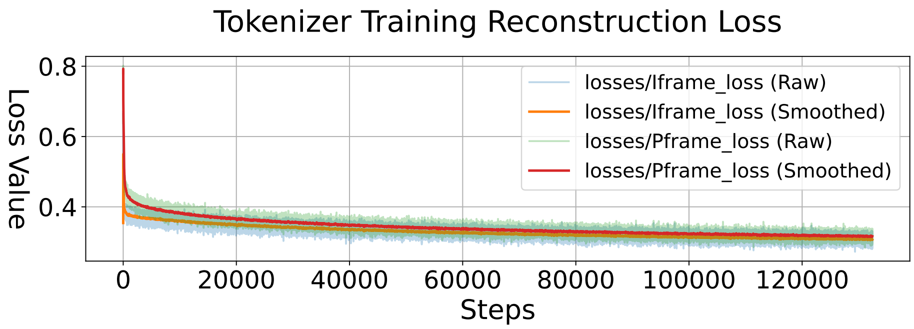

Figure 7. Training loss comparison of the video tokenizer. The plot illustrates the reconstruction loss trajectories for IFrame and PFrame components over training iterations. Despite the different token allocation strategies (330 tokens for IFrame vs. 74 tokens for PFrame), both frame types achieve comparable reconstruction quality.

the computational efficiency of the attention mechanism in the tokenizer, we employ flex attention[9]. In addition, inspired by EnCodec [8], to avoid discontinuities when encoding videos, we set a 20% overlap between groups. During training, the batch size is 96, we use the AdamW [36] optimizer, the learning rate is constant at $1e-4$, and the learning rate decay factor is 0. During training, we use Model Exponential Moving Average (EMA) to smooth the model parameters, and the decay rate of EMA is 0.8. The weights of the reconstruction loss and the commitment loss are both 1. Figure 7 shows the reconstruction loss trajectories for IFrame and PFrame components during training. Despite allocating significantly different token quantities (330 vs. 74), both frame types converge to comparable reconstruction quality. This validates our video frame grouping strategy that prioritizes key frames while minimizing tokens for intermediate frames. The results confirm our approach successfully balances reconstruction fidelity with computational efficiency.

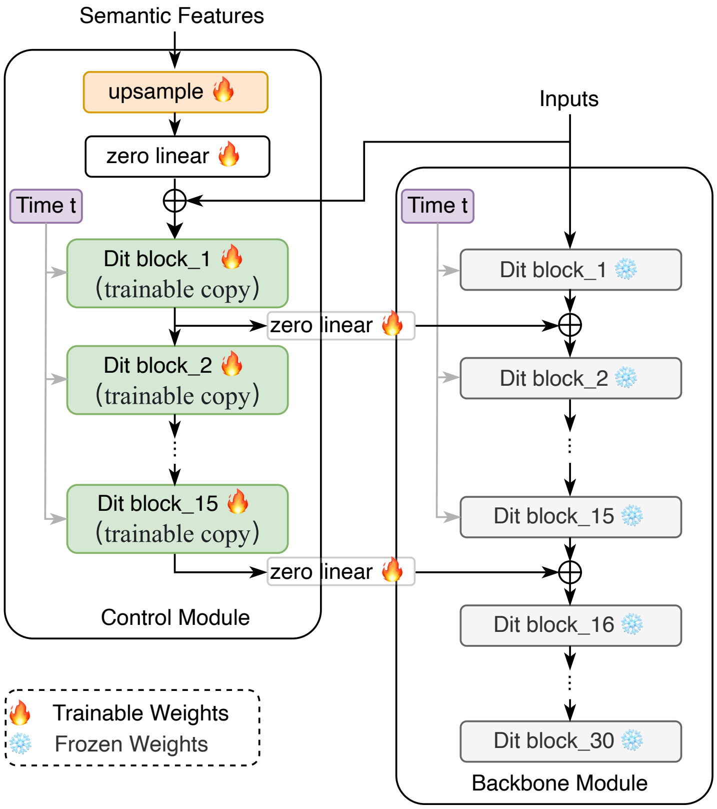

Figure 8. Proposed diffusion model structure. We use a ControlNet-style control module to guide the model to generate perceptual feature based on semantic features.

A.2. Details of LLM

We use a model structure similar to LLaMA [53] as the LLM. We set the text, motion score, and frames conditions in subsection 3.2 to null with probabilities of 10%, 50%, and 50%, respectively. The model has 24 layers, a hidden layer size of 2048, 16 heads, and an MLP hidden layer size of 11008. The batch size for model training is 4096, we use the AdamW optimizer, the learning rate is $1e-3$, the learning rate decay factor is 0.1, and we use a warm-up strategy for the first 1000 steps of training. We use a cosine learning rate decay strategy. We apply classifier-free guidance [19] for better generation quality, and the guidance scale is set to 6.5. We do not use top-k and top-p sampling.

A.3. Details of Diffusion Model

We copy the first 15 layers of the base model as the proposed trainable control module in subsection 3.3. We use a structure similar to the VQ-GAN [11] decoder as the upsampling module and change the upsampling method to pixelshuffle [47]. As shown in Table 7, the total number of parameters of the video detokenizer is 3B, and the number of parameters of the trainable control module is 1B. During inference, we follow the same sampling strategy as Yang et al. [60]. The batch size for training is 128, we use the AdamW optimizer, the learning rate is $1e-4$, and the learning rate decay factor is $1e-4$. To speed up training, we first train directly on the original features extracted from Theia. Then we use the quantized reconstructed features for training.

Table 8. **Performance comparison of Text-to-video (T2V) generation between our LanDiff and other state-of-the-art models on VBench benchmark.** The remaining 8 evaluation dimensions of VBench that are not shown in the main text. The best and second-best scores are highlighted in **bold** and underline, respectively. † indicates the scores we reproduced, while ‡ indicates the scores from the original papers, and other scores are from the VBench benchmark.

Model Name	Type	Model Size	Total Score	Quality Score	Semantic Score	aesthetic quality	appearance style	color	human action	imaging quality	overall consistency	temporal flickering	temporal style
Open Sourced Models													
InstructVideo	Diffusion	1.3B	76.61	81.56	56.81	52.55	20.16	77.14	85.20	68.01	19.91	98.19	21.26
Latte-1	Diffusion	0.7B	77.29	79.72	67.58	61.59	23.74	85.31	90.00	61.92	27.33	98.89	24.76
OpenSoraPlan V1.1	Diffusion	2.7B	78.00	80.91	66.38	56.85	22.90	89.19	86.80	62.28	26.52	99.03	23.87
Show-1	Diffusion	6.3B	78.93	80.42	72.98	57.35	23.06	86.35	95.60	58.66	27.46	99.12	25.28
OpenSora V1.2	Diffusion	1.1B	79.76	81.35	73.39	56.85	23.95	90.08	91.20	63.34	26.85	99.53	24.54
LTX-Video	Diffusion	1.9B	80.00	82.30	70.79	59.81	21.47	81.45	92.80	60.28	25.19	99.34	22.62
Mochi-1	Diffusion	10B	80.13	82.64	70.08	56.94	20.33	79.73	94.60	60.64	25.15	99.40	23.65
AnimateDiff-V2	Diffusion	1.3B	80.27	82.90	69.75	67.16	22.42	87.47	92.60	70.10	27.04	98.75	**26.03**
VideoCrafter-2.0	Diffusion	1.7B	80.44	82.20	73.42	63.13	25.13	**92.92**	95.00	67.22	**28.23**	98.41	25.84
CogVideoX-2B	Diffusion	2B	80.91	82.18	75.83	60.82	24.80	79.41	98.00	61.68	26.66	98.89	24.36
Emu3 ‡	LLM	8B	80.96	N/A	N/A	59.64	20.92	N/A	77.71	N/A	N/A	N/A	N/A
Vchitect-2.0-2B	Diffusion	2B	81.57	82.51	77.79	61.47	24.93	86.87	97.00	65.60	28.01	98.45	25.56
CogVideoX-5B	Diffusion	5B	81.61	82.75	77.04	61.98	24.91	82.81	**99.40**	62.90	27.59	98.66	25.38
DiT †	Diffusion	7B	81.85	82.70	78.42	60.00	24.95	78.62	98.20	63.80	27.85	99.13	26.10
RepVideo	Diffusion	2B	81.94	82.70	78.91	62.40	25.12	82.51	98.00	63.16	26.96	99.16	25.31
Vchitect-2.0[E]	Diffusion	2B	82.24	83.54	77.06	60.41	23.73	87.04	97.20	65.35	27.57	98.57	25.01
HunyuanVideo	Diffusion	13B	83.24	85.09	75.82	60.36	19.80	91.60	94.40	67.56	26.44	99.44	23.89
Close Sourced Models													
Pika-1.0	Diffusion	N/A	80.69	82.92	71.77	62.04	22.26	90.57	86.20	61.87	25.94	**99.74**	24.22
Kling	Diffusion	N/A	81.85	83.39	75.68	61.21	19.62	89.90	93.40	65.62	26.42	99.30	24.17
Jimeng	Diffusion	N/A	81.97	83.29	76.69	**68.80**	22.27	89.05	90.10	67.09	27.10	99.03	24.70
Gen-3	Diffusion	N/A	82.32	84.11	75.17	63.34	24.31	80.90	96.40	66.82	26.69	98.61	24.71
Hailuo	Diffusion	N/A	83.41	84.85	77.65	63.03	20.06	90.36	92.40	67.17	27.10	99.10	25.63
Sora	Diffusion	N/A	84.28	85.51	79.35	63.46	24.76	80.11	98.20	**68.28**	26.26	98.87	25.01
ARLON ‡	LLM+Diffusion	1.5B	N/A	N/A	N/A	61.01	N/A	N/A	N/A	60.98	27.27	99.37	25.33
ARLON †	LLM+Diffusion	5B	82.31	83.58	77.27	60.58	25.26	85.86	92.20	62.72	26.14	99.39	24.07
LanDiff	LLM+Diffusion	5B	**85.43**	**86.13**	**82.61**	64.78	**25.60**	91.09	97.20	65.69	27.43	99.43	25.26

A.4. More Analysis on VBench Benchmark

As shown in Table 8, we conduct a comprehensive comparison with state-of-the-art text-to-video generation models on VBench benchmark. The benchmark evaluates models across multiple dimensions including quality, semantics, aesthetics,

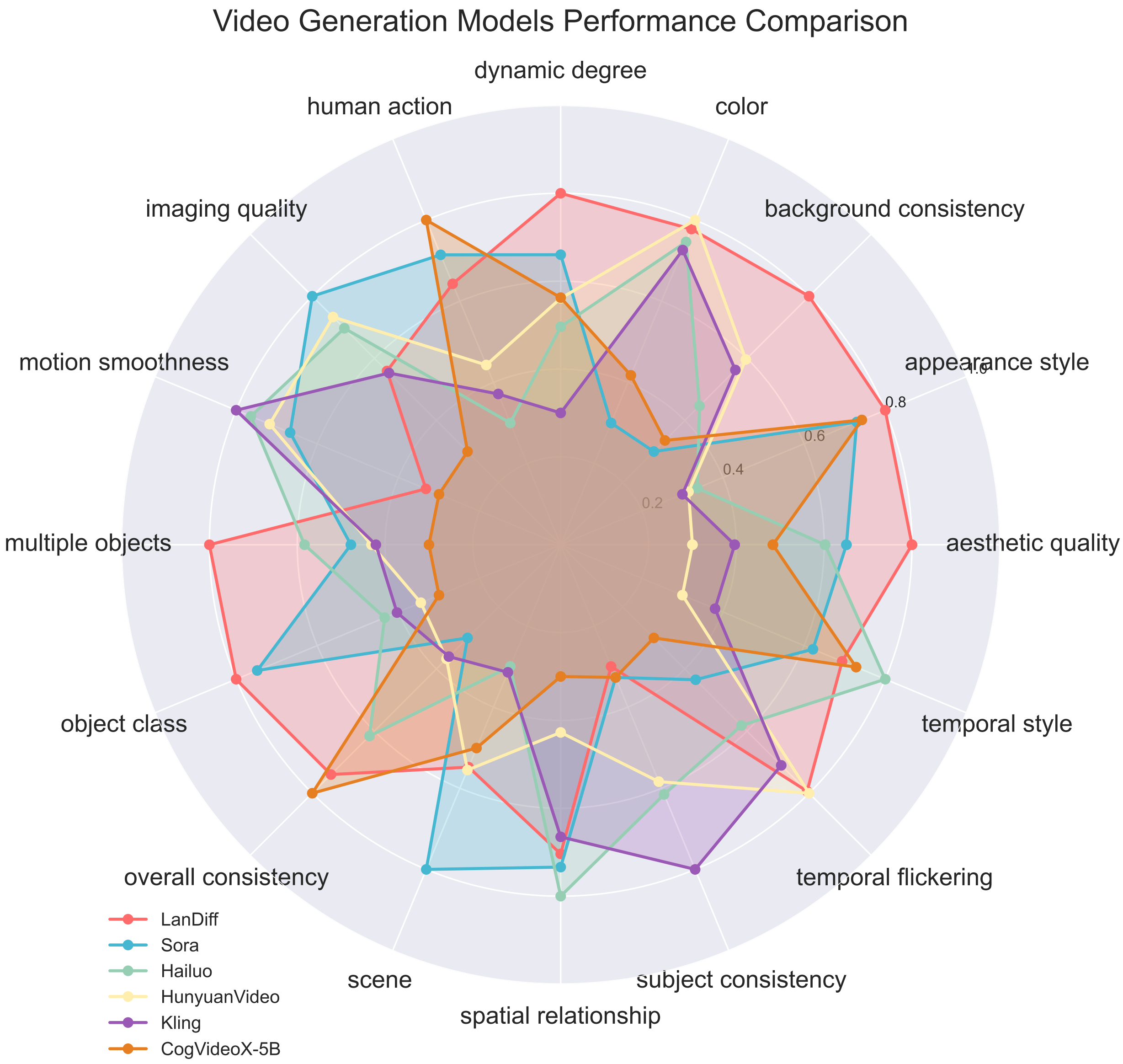

Figure 9. **Radar chart visualization of performance comparison across different dimensions on VBench.** The plot compares LanDiff against five competitive baselines: Sora, Hailuo, HunyuanVideo, Kling, and CogVideoX-5B. For better readability, the values in the radar chart have been normalized to a scale ranging from 0.3 to 0.8. The normalization was performed using the min-max scaling formula: $normalized = 0.3 + 0.5 \times \frac{value-min\_value}{max\_value-min\_value}$. The original raw performance data can be found in Table 3 and Table 8.

and temporal consistency. Our LanDiff achieves superior performance in most metrics, particularly excelling in overall quality (86.13) and semantic accuracy (82.61).

Among open-sourced models, there is a clear trend of performance improvement with model size, from Latte-1 (0.7B) to HunyuanVideo (13B). However, our hybrid LLM+Diffusion approach (5B) demonstrates that architectural innovation can be more impactful than simply scaling up model parameters. Notably, LanDiff outperforms much larger models like HunyuanVideo (13B) and Mochi-1 (10B) across most metrics.

As visualized in Figure 9, we compare LanDiff with five representative models across different evaluation dimensions. The radar chart reveals that LanDiff (shown in red) demonstrates well-balanced performance across all metrics, with notably strong results in quality score and semantic accuracy. While Sora (shown in blue) achieves competitive scores in imaging quality and scene, and HunyuanVideo excels in certain visual aspects, LanDiff maintains consistently superior performance across the entire spectrum of metrics. Notably, while there is typically a trade-off between dynamic degree and motion smoothness/subject consistency, LanDiff achieves a high level of dynamism while maintaining strong performance in stability metrics - with subject consistency and motion smoothness scores within 2.3% and 2.5% of the best-performing models respectively. Notably, while text-to-video models typically exhibit a trade-off between dynamic expressiveness and temporal stability, LanDiff successfully balances these competing objectives—achieving high dynamism while maintaining robust stability metrics, with subject consistency and motion smoothness scores deviating by only 2.3% and 2.5% respectively from the state-of-the-art in each category. The comprehensive comparison with these strong baselines, including both commercial (Sora, Hailuo, Kling) and open-source models (HunyuanVideo, CogVideoX-5B), further validates the effectiveness of our hybrid LLM+Diffusion approach.

B. Broader Impact

Text-to-video generation models such as LanDiff offer substantial potential for creative applications across entertainment, education, and content creation domains. Nevertheless, these technologies introduce ethical considerations and potential risks that warrant attention. The capability to generate photorealistic videos from textual descriptions could potentially be exploited to create misleading or deceptive content, including sophisticated deepfakes or videos that misrepresent individuals or events. Such misuse raises significant concerns regarding misinformation propagation, privacy violations, and potential harm to individuals or communities. To mitigate these risks, we recommend several safeguards: (1) incorporating visible watermarks or robust digital signatures in generated content to ensure transparency regarding its synthetic nature; (2) advancing and deploying sophisticated detection systems capable of identifying AI-generated content with high accuracy; (3) establishing comprehensive usage policies that explicitly prohibit harmful applications; and (4) implementing accessible reporting mechanisms for suspected misuse cases. Furthermore, we advocate for continued research into technical safeguards and responsible deployment frameworks specifically designed for generative video models. We emphasize that our research contribution aims to advance the field of multimodal generation for socially beneficial applications while acknowledging the necessity of addressing potential negative impacts through complementary technical innovations and policy measures.

C. Limitation and Future Works

While LanDiff demonstrates significant advancements in text-to-video generation, several limitations remain to be addressed in future work. First, the scale of our language model (2B parameters) is substantially smaller than state-of-the-art text-only LLMs, potentially limiting the semantic understanding and generation capabilities. Future work will explore scaling our language models to larger parameter counts to enhance performance. Second, our choice of CogVideoX-2B as the underlying diffusion model establishes an upper bound on the quality of generated videos. We plan to investigate the development of more sophisticated diffusion backbones specifically optimized for video generation tasks to overcome this constraint. Third, we observe that LanDiff struggles with accurate text rendering within generated videos. This limitation likely stems from insufficient supervision of text features in the current semantic token representation. Future research will focus on developing more comprehensive video semantic tokens with enhanced text-specific supervision. Our current work primarily addresses text-to-video generation, but we envision extending LanDiff's capabilities to broader application scenarios. These include image-to-video generation, unified models for video understanding and generation, and interactive controllable video synthesis. These extensions would significantly expand the utility of semantic token-based language models in multimodal generation tasks and provide more flexible creative tools for users across various domains.

C.1. More Examples

A group of silver-colored fish with darker fins swim among green aquatic plants in an aquarium setting. The fish move gracefully through the water, navigating around the plants, which are of various sizes and shades of green. The aquarium environment is designed to mimic a natural habitat, with rocks and shadows in the background contributing to the underwater scene.

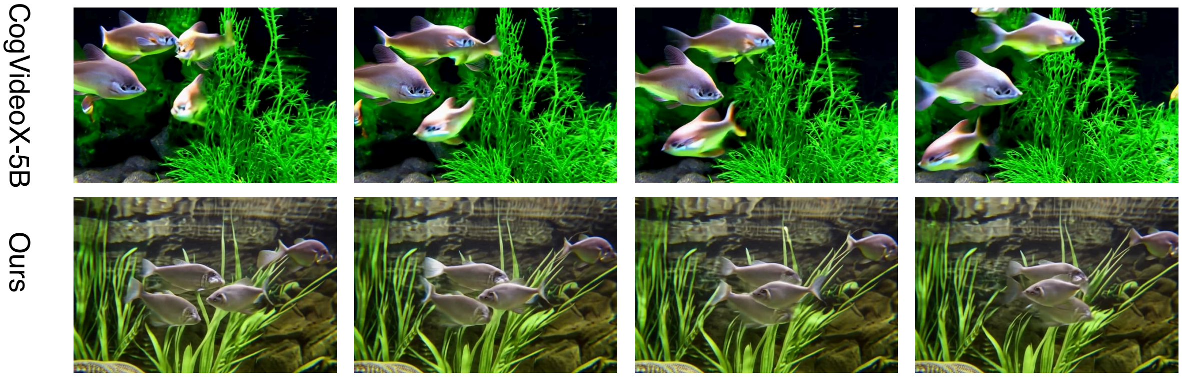

Figure 10. Examples of text to videos generation of LanDiff and CogVideoX-5B.

A life–sized ice sculpture of a playful dog, with intricate details and a joyful expression, stands in the middle of a sunlit, grassy field on a sweltering summer day. The ice dog, initially solid and vibrant, begins to melt under the relentless heat, with droplets of water forming on its surface. As the day progresses, the ice dog's form gradually diminishes, with its once sharp features becoming blurred and distorted. The melting process accelerates, and the ice dog's body starts to collapse, pooling into a puddle of water on the ground. By the end of the day, all that remains is a shallow puddle, reflecting the cloudless sky, with the memory of the once majestic ice dog now just a memory.

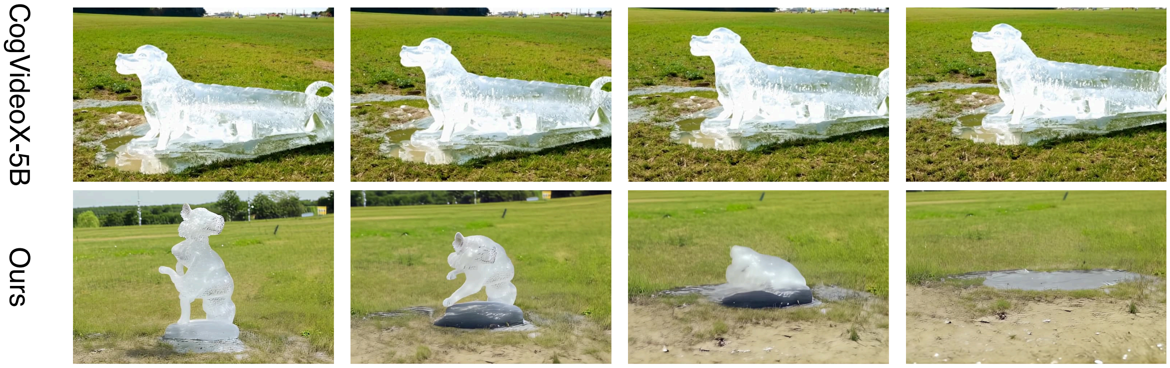

Figure 11. Examples of text to videos generation of LanDiff and CogVideoX-5B.

A close-up view of a Christmas tree reveals a variety of decorations including a purple ornament with a gold pattern, a gold textured ornament, a small white house-shaped ornament with red roof and gold details, and a brown pine cone. The tree branches are dense and green, providing a natural backdrop for the ornaments. The camera pans slightly across the scene, maintaining focus on the ornaments while subtly shifting the perspective.

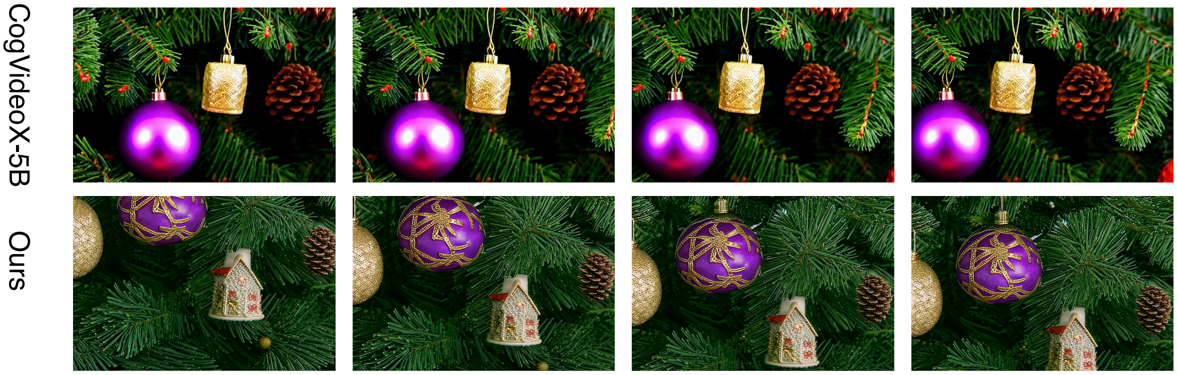

Figure 12. Examples of text to videos generation of LanDiff and CogVideoX-5B.

Two vibrant hot air balloons, one red and the other blue, are seen soaring through a clear blue sky, their baskets gently bumping against each other mid-air. The red balloon features intricate gold patterns on its surface, while the blue balloon boasts a white and silver design. As they collide, the passengers in the baskets, dressed in casual attire, react with surprise and excitement. The scene is set against a backdrop of a picturesque landscape, with lush green hills and a sparkling river below. The balloons' vibrant colors contrast beautifully with the azure sky, creating a visually stunning and dynamic scene.

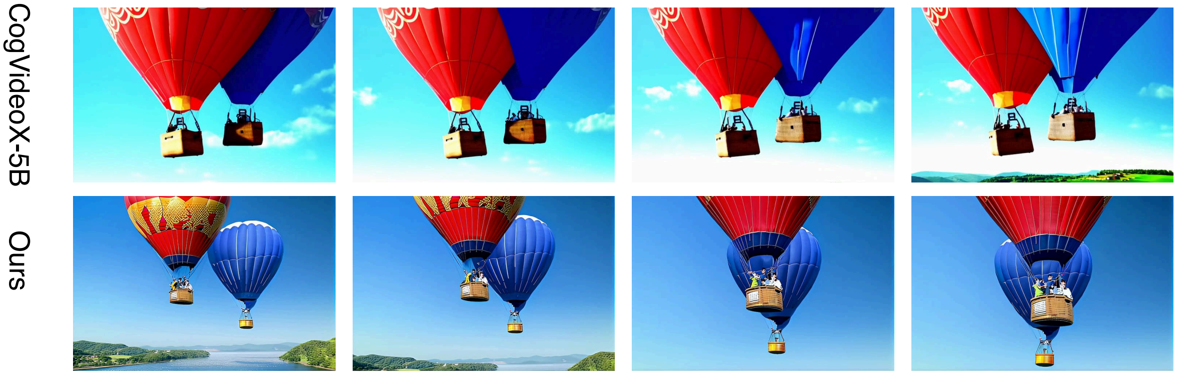

Figure 13. Examples of text to videos generation of LanDiff and CogVideoX-5B.

A colossal, human–shaped cloud towers over the earth, its massive form casting a shadow across the landscape. The cloud man's features are distinct, with a stern expression and outstretched arms. Suddenly, the cloud man releases a barrage of lightning bolts, illuminating the sky as they streak towards the earth. The scene is set against a backdrop of a stormy sky, with dark clouds and distant thunder adding to the dramatic atmosphere.

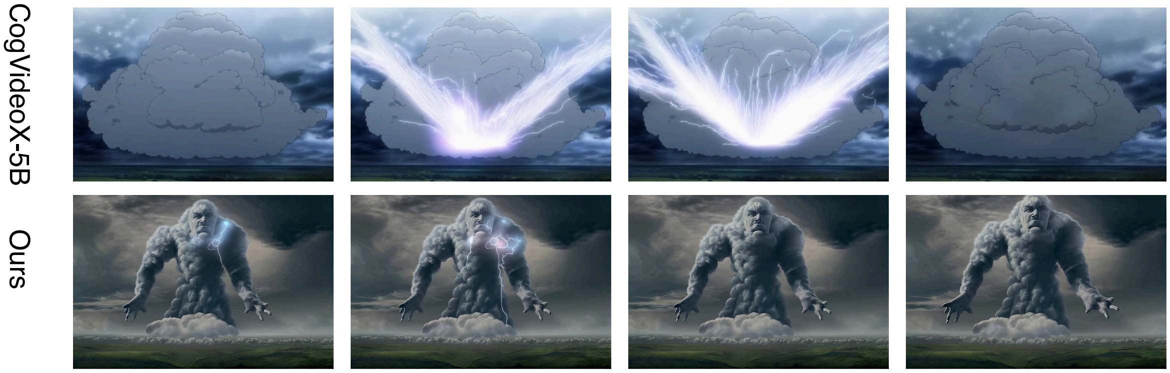

Figure 14. Examples of text to videos generation of LanDiff and CogVideoX-5B.

A sleek white sailboat glides gracefully across a calm, azure sea, its sails billowing gently in the breeze. Above, a silver airplane soars through a clear blue sky. The boat's hull reflects the sunlight, creating a shimmering effect on the water's surface. The airplane, seen in a high–altitude flyover, casts a shadow that momentarily aligns with the boat's path, creating a fleeting connection between sea and sky. The scene is captured in a wide shot, ensuring both the boat and airplane are prominently centered, emphasizing their contrasting yet harmonious presence in the vast expanse.

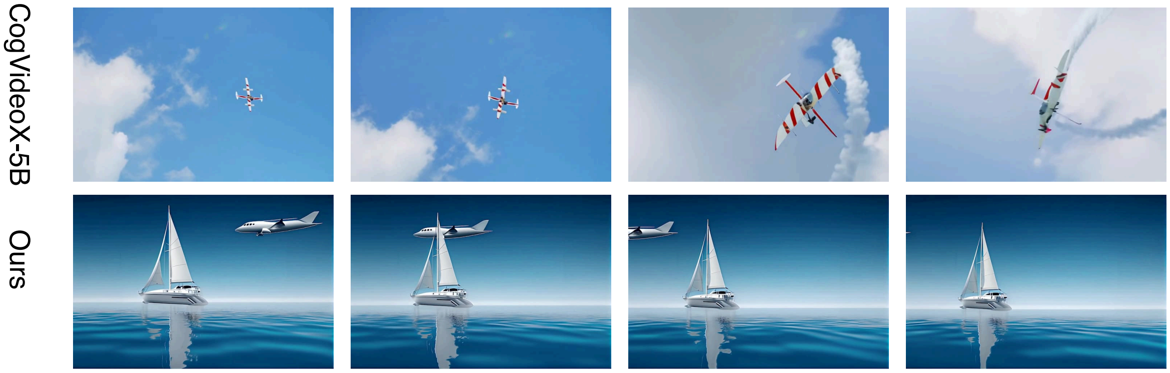

Figure 15. Examples of text to videos generation of LanDiff and CogVideoX-5B.

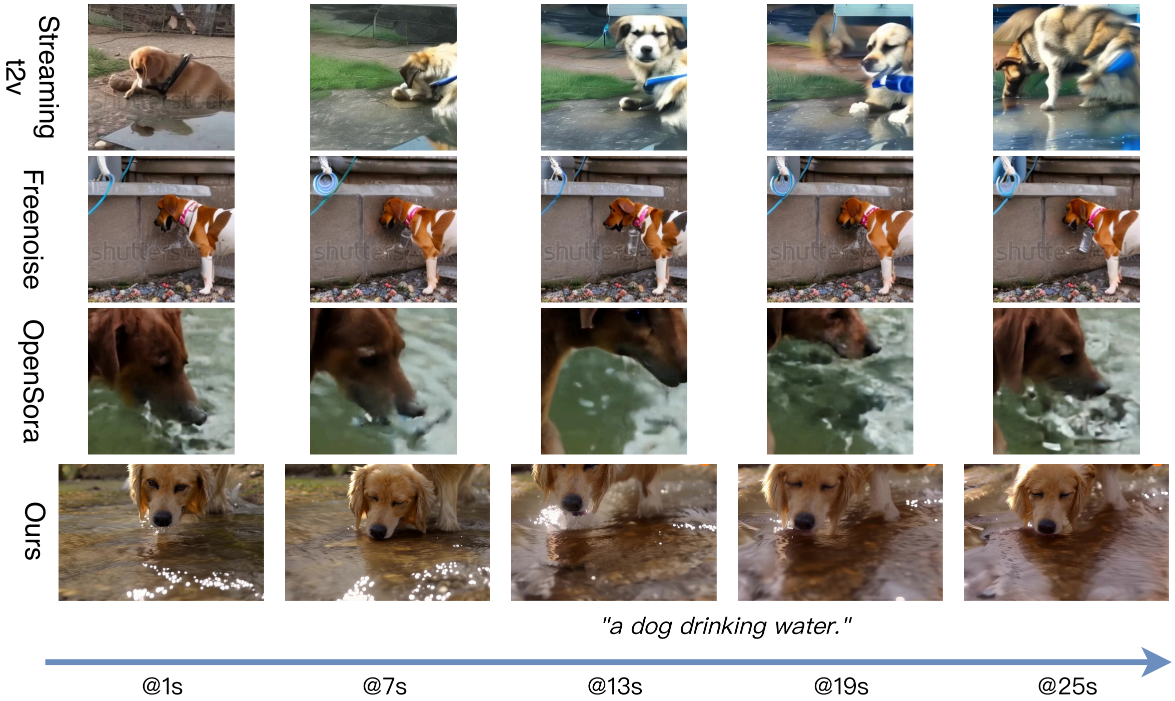

Figure 16. Examples of text to long video generation.

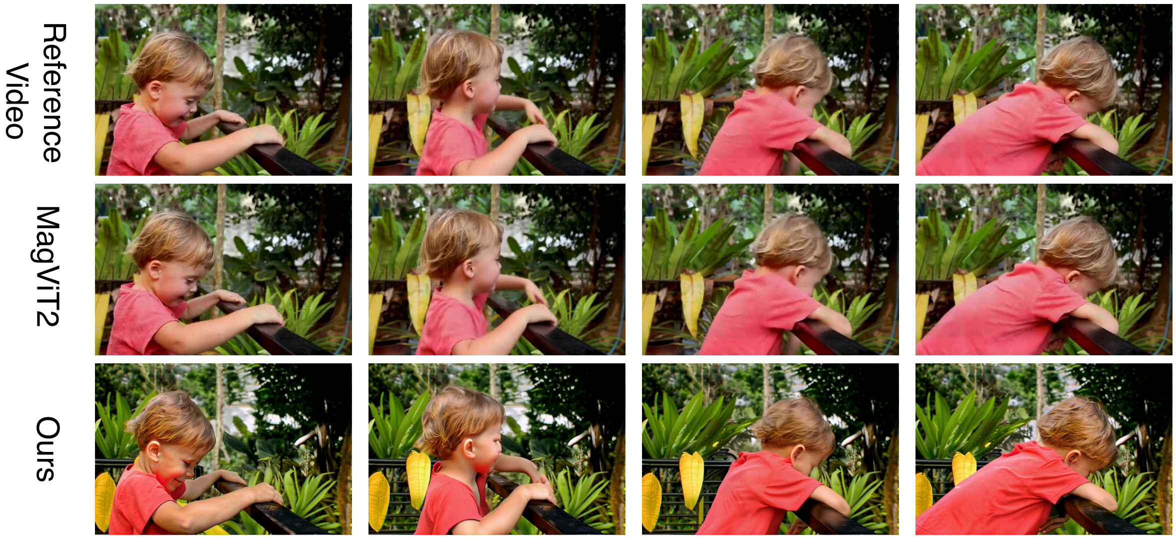

Figure 17. Visualization results of video reconstruction using video tokenizer.

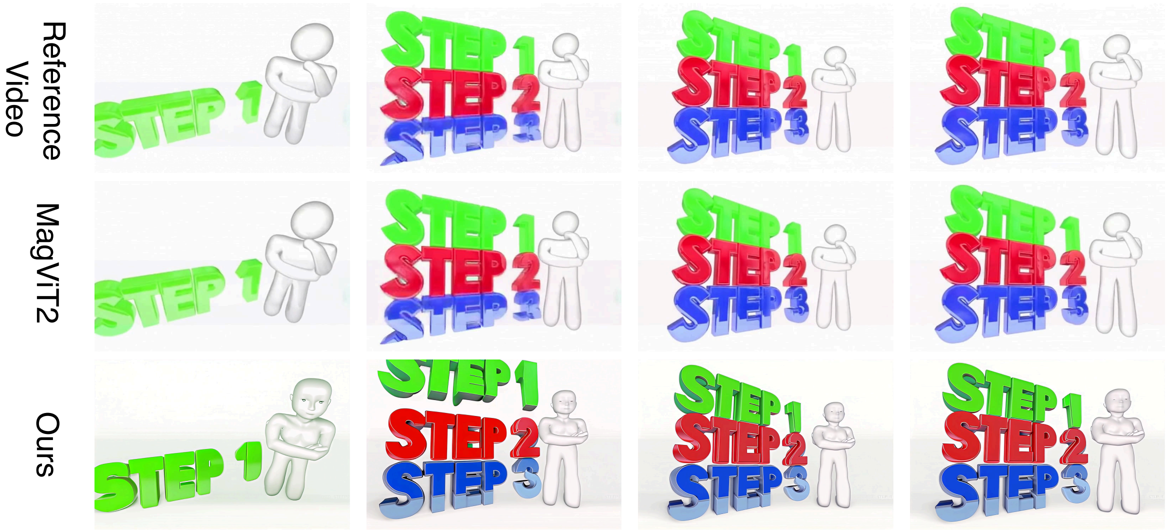

Figure 18. Visualization results of video reconstruction using video tokenizer.