Mask^{2}DiT: Dual Mask-based Diffusion Transformer for Multi-Scene Long Video Generation

Supplementary Material

6. Supplementary

Auto-Regressive Scene Extension. Fig. 8 and Tab. 4 present the visualization and quantitative results of auto-regressive scene extension generated by Mask^{2}DiT, respectively, highlighting its effectiveness in extending videos based on a given set of visible scenes. In the provided examples, the rightmost three columns display frames generated by Mask^{2}DiT. These results highlight the model's ability to seamlessly extend scenes while maintaining temporal and semantic coherence. Notably, the transitions between scenes are smooth, and the generated frames exhibit consistent visual quality, effectively capturing the intended narratives within each segment. This showcases Mask^{2}DiT's capacity for generating longer, more complex videos with high fidelity. As for the implementation details, we fix the timesteps t corresponding to the visual token sequences of the conditional $n-1$ segments to 0, while the timesteps t for the n-th segment to be generated remain consistent with those during standard training.

Method	Visual Con. (%)	Semantic Con. (%)	Sequence Con. (%)
Ours	75.33	24.29	49.81

Table 4. Quantitative results of auto-regressive scene extension.

More quantitative results. To further validate the effectiveness of Mask^{2}DiT, we additionally report evaluation results for both fixed-scene and auto-regressive scene extension settings on EvalCrafter [23] and T2V-CompBench [34], as shown in Tab. 6. Overall, our method demonstrates superior consistency and achieves competitive performance across multiple metrics on standard single-scene video generation benchmarks. However, these benchmarks are not well-suited for our task.

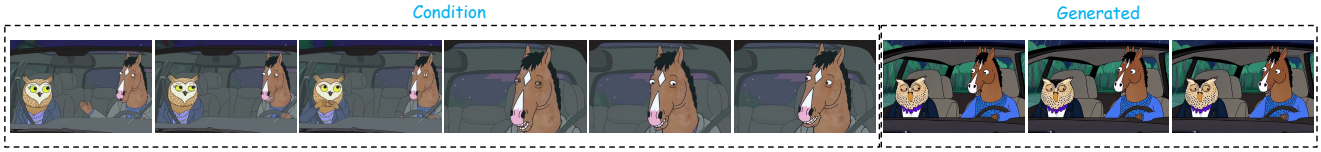
Method	Visual Con. (%)	Semantic Con. (%)	Sequence Con. (%)	Aesthetic Quality	Imaging Quality
Pre-training	49.04	**26.36**	37.70	-	-
SFT	73.22	23.54	48.38	57.44	72.37
Pre-training + SFT	**73.73**	23.73	**48.73**	**57.68**	**73.20**

Table 5. Effect of pre-training dataset.

Effect of Pre-training Dataset. To evaluate the impact of the pre-training dataset, we compare three configurations: using only the pre-training dataset, using only the SFT dataset, and a hybrid approach where the model is first pre-trained on the pre-training dataset and then fine-tuned with the SFT dataset, as shown in the Tab 5. Since CogVideoX is trained on single-scene data, its extension to multi-scene generation often results in failures. Pre-training allows the model to adapt to generating distinct scenes for different prompts. However, because the pre-training dataset consists of randomly concatenated single-scene videos without inherent logical relationships, it exhibits lower consistency compared to results achieved with the SFT dataset. Training exclusively on the SFT dataset improved Visual Consistency by 24.18% over the pre-training dataset, underscoring the critical role of multi-scene consistency in data for our approach. Finally, pre-training with the pre-training dataset followed by fine-tuning with the SFT dataset leads to a 0.35% improvement in Sequence Consistency, highlighting the complementary benefits of pre-training, particularly in scenarios where SFT data is limited. Furthormore, the pre-training phase contributes to enhanced visual quality in the generated videos, as evidenced by improvements in the aesthetic quality (**57.68** vs. 57.44) and imaging quality (**73.20** vs. 72.37) metrics reported by VBench [18]. These findings suggest that incorporating additional high-quality, semantically coherent multi-scene data could further enhance the performance of our model.

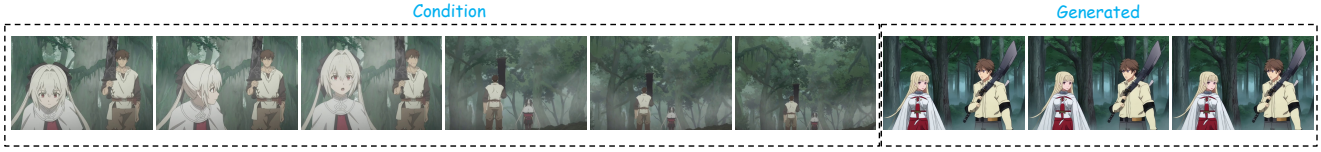
Performance Improvement with Larger Models. As shown in Tab 7, we compare the performance of a 5B model and observe that improvements in the foundation model lead to significant gains in multi-scene video generation. Specifically, the 5B model outperforms the 2B model with a 7.56% improvement in Sequence Consistency. To further illustrate the effectiveness of various foundation models, we present qualitative results in Fig. 10, 11, 12, 13, 14, 15, 16, 17, with the original videos included in the supplementary compressed package.

User study. In addition to objective evaluations, we have also designed a user study to subjectively assess the practical performance of various methods. Given 10 three-part prompts generated by ChatGPT, we adopt CogVideoX, StoryDiffusion equipped with CogVideoX-I2V, TALC, and Mask^{2}DiT to generate corresponding multi-scene videos separately. Each method is executed once with the same random seed, ensuring a fair comparison by eliminating randomness in generation. We ask 10 users from distinct backgrounds to evaluate the generated results across 4 dimensions based on the following question: "For a given three-scene text prompt, the following options present results from different multi-scene long video generation methods. Please evaluate the results in terms of text-video alignment (faithfulness to the given text prompt), video quality, and visual consistency across scenes. Based on these three criteria, please select the best result for each aspect and, finally, choose the overall best result considering all three aspects." Since abstention is allowed, we ultimately receive 266 valid votes. The final results are displayed in Tab. 9. Mask^{2}DiT



(a) "An owl and a horse are driving in a car at night. The owl gestures while talking to the horse, who is focused on driving.", "A cartoon horse in a gray sweater sits in a car at night, looking around as if reacting to something off-screen.", "An owl in a dark jacket and purple necklace sits anxiously in the passenger seat of a car driven by a horse in a blue sweater. They drive through a rainy, wooded area at night."

(b) "Two animated mice, Actor1 and Actor2, stand in a dimly lit space with a rough, concrete-like backdrop. Actor1 is completely white, while Actor2 has light grey shading. Actor1 points towards something off-screen to the left. Actor2 looks towards Actor1, its expression seemingly neutral.", "Two white rats face each other. The left rat blinks slowly, while the right one exhales visible puffs of air.", "Actor1 and Actor2, two anthropomorphic white rats, stand in a dimly lit area resembling a basement. They turn their heads towards each other."



(c) "A young woman with long, light hair and a white cloak stands in a misty forest with a young man carrying a large sword. She looks off to the side, seemingly curious or concerned.", "A man with brown hair, wearing a cream-colored, long-sleeved tunic and brown pants, and a pale-skinned woman with long white hair tied back with a dark ribbon, wearing a white and dark red outfit, stand at the edge of a misty forest with an uneven, foliage-littered floor.", "A woman with long, light hair, a white cape, and a red sash stands next to a man with brown hair and a cream shirt in a misty forest. He holds a large, dark sword over his shoulder. They both look to the left."

Figure 8. Visualization results of auto-regressive scene extension.

	Methods	VQA_A	VQA_T	flow score	Action binding	Consistent Attr Binding	Dynamic Attr Binding	Object Interactions
scene extension	Ours	58.70	57.27	0.64	1.76	2.93	0.0005	1.83
fixed scenes	CogVideoX	53.78	63.76	1.87	2.16	2.84	0.0076	2.01
	StoryDiffusion + CogVideoX-I2V	28.71	33.96	1.08	2.17	3.17	0.0075	**2.08**
	TALC	3.72	7.66	**14.99**	2.15	2.70	**0.0121**	2.01
	Ours	**75.48**	**64.61**	1.75	**2.30**	**3.28**	0.0065	2.02

Table 6. Evaluation for auto-regressive scene extension and video generation with a fixed number of scenes on two T2V evaluation benchmarks, *i.e.*, EvalCrafter [23] and T2V-CompBench [34].

Model	Visual Con. (%)	Semantic Con. (%)	Sequence Con. (%)	FVD (↓)
CogVideoX-2B	55.01	22.64	38.82	835.35
Ours	**70.95**	**23.94**	**47.45**	**720.01**
CogVideoX-5B	43.82	20.70	32.26	613.47
Ours	**89.21**	**20.81**	**55.01**	**607.64**

Table 7. Performance comparison with larger models. The 5B model improves multi-scene video generation by 7.56% over the 2B model, highlighting the impact of stronger foundation models.

Method	Visual Con. (%)	Semantic Con. (%)	Sequence Con. (%)	FVD (↓)
Ours(3s)	**87.62**	22.38	**55.00**	818.96
Ours(6s)	70.95	**23.94**	47.45	**720.01**

Table 8. Performance in shorter scene generation.

outperforms all state-of-the-art methods in all three evaluation aspects and overall preference by a big margin, demonstrating the broad application prospects of our method.

Ablation studies on probability p. We add ablation studies on probability p, as shown in Tab Tab. 10.

Ablation studies on generating shorter sequences. We directly apply our model trained on 6-second scenes to generate 3-second scenes, obtaining promising results, as shown in Tab. 8 and Fig. 9. The only limitation is a slight degradation in semantic consistency.

Limitations. The limitations are twofold. First, due to the constraints of the training data, our model is limited to generating animated videos. Second, the motion dynamics and durations for each scene require further investigation.

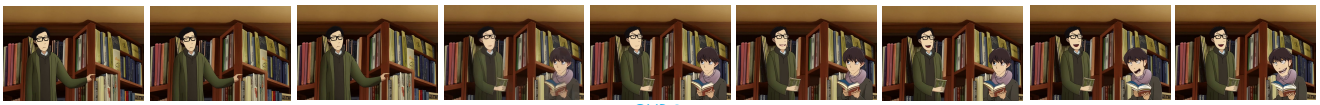


Figure 9. Visualization results of shorter scene generation.

Aspect	Visual Consistency↑	Semantic Consistency↑	Video Quality↑	Overall↑
CogVideoX	8.96	12.12	9.23	9.09
StoryDiffusion + CogVideoX-I2V	29.85	28.79	29.23	28.79
TALC	13.43	13.64	12.31	12.12
Ours	**47.76**	**45.45**	**49.23**	**50.00**

Table 9. Results for the user study in percentages.

	Video generation with a fixed number of scenes			Auto-regressive scene extension			Overall
p	Visual Con.	Semantic Con.	Sequence Con.	Visual Con.	Semantic Con.	Sequence Con.	Sequence Con.
0.1	72.37	23.36	47.86	72.63	23.38	48.01	47.94
0.3	**73.51**	23.61	**48.56**	73.89	23.87	48.88	48.72
0.5	73.21	23.82	48.52	75.33	**23.91**	49.62	**49.07**
0.7	71.63	**23.92**	47.78	**77.57**	22.81	**50.19**	48.98
0.9	67.96	23.54	45.75	75.73	23.50	49.61	47.68

Table 10. Ablation studies on the probability p.

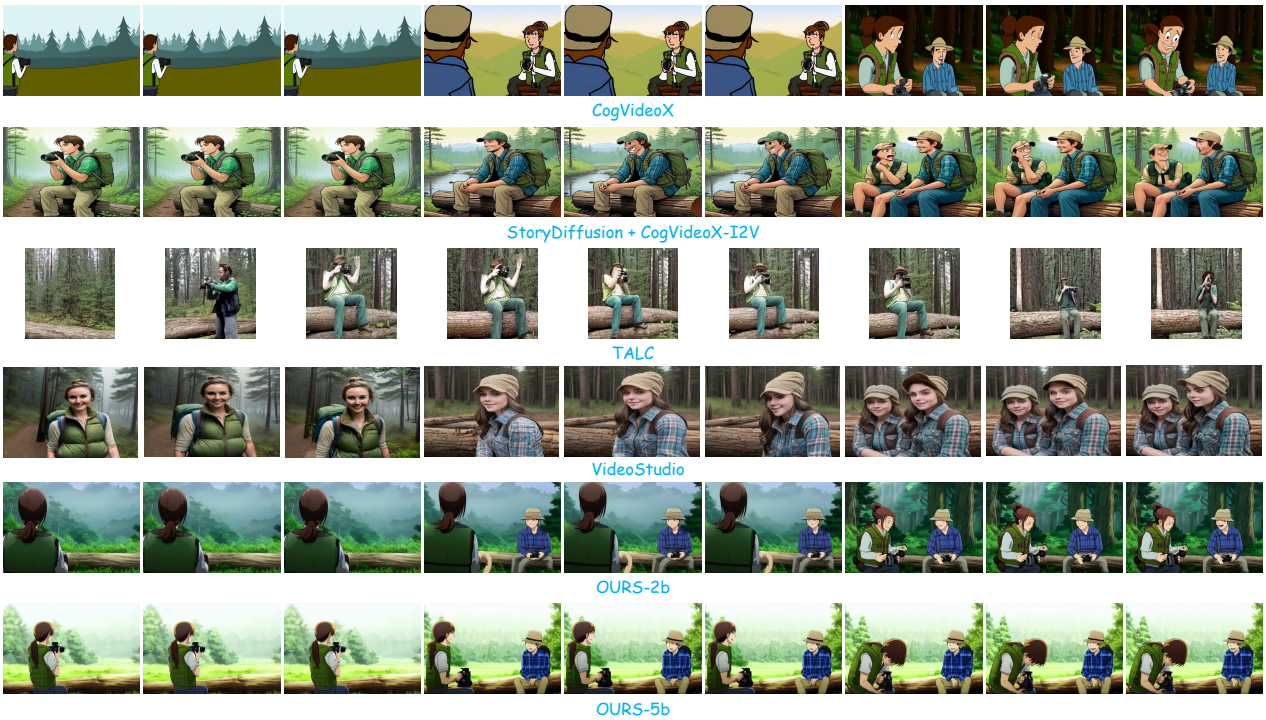


Figure 10. Visualization of Results from Different SOTA Methods. Video Captions: ”Actor1 (brown hair tied back, wearing a green hiking vest, holding a camera) stands at the edge of the clearing, framing a shot of the misty forest.”, ”Actor2 (wearing a beige hat, blue flannel shirt, sitting on a log) watches Actor1 (brown hair tied back, wearing a green hiking vest, holding a camera) with a smile, appreciating the tranquility of the early morning.”, ”Actor1 (brown hair tied back, wearing a green hiking vest, holding a camera) lowers the camera and shares a quiet laugh with Actor2 (wearing a beige hat, blue flannel shirt, sitting on a log) as they both take in the serene beauty of the forest.”

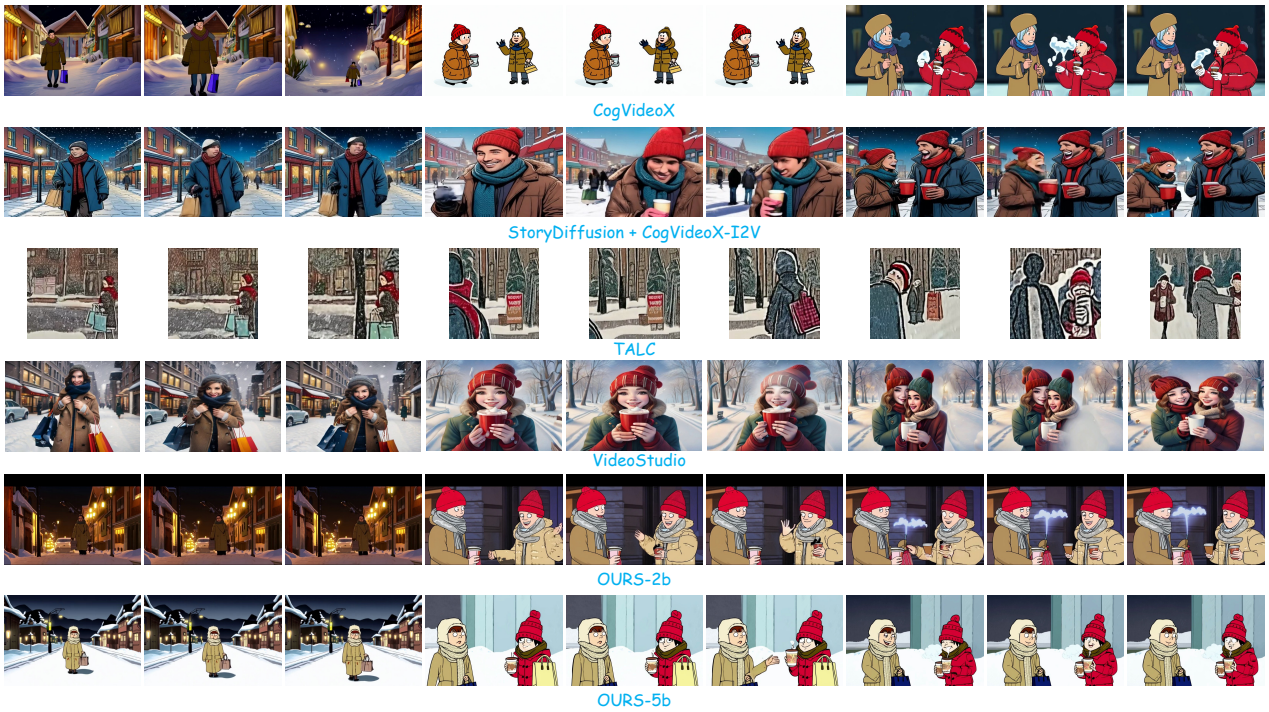


Figure 11. Visualization of Results from Different SOTA Methods. Video Captions: "Actor1 (gray backpack, green jacket, reading a map) stands at the platform, studying the station layout.", "Actor2 (blonde ponytail, red scarf, holding a suitcase) notices Actor1 (gray backpack, green jacket, reading a map) and approaches to offer directions.", "Actor1 (gray backpack, green jacket, reading a map) thanks Actor2 (blonde ponytail, red scarf, holding a suitcase), both exchanging a friendly smile in the busy station."

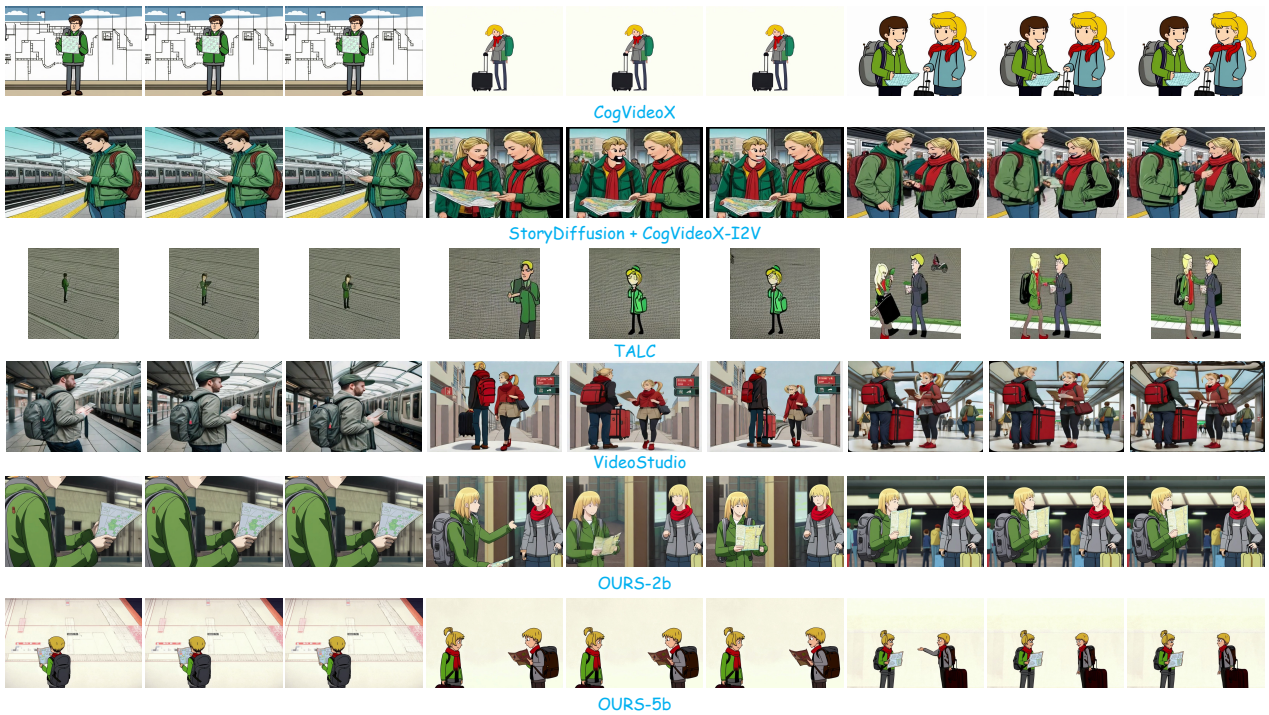


Figure 12. Visualization of Results from Different SOTA Methods. Video Captions: "Actor1 (wavy hair, plaid shirt, playing a ukulele) strums a soft melody by the bonfire.", "Actor2 (braided hair, cozy blanket, singing along) joins in, creating a serene harmony under the stars.", "Actor1 (wavy hair, plaid shirt, playing a ukulele) laughs softly as Actor2 (braided hair, cozy blanket, singing along) harmonizes, both enjoying the peaceful night."

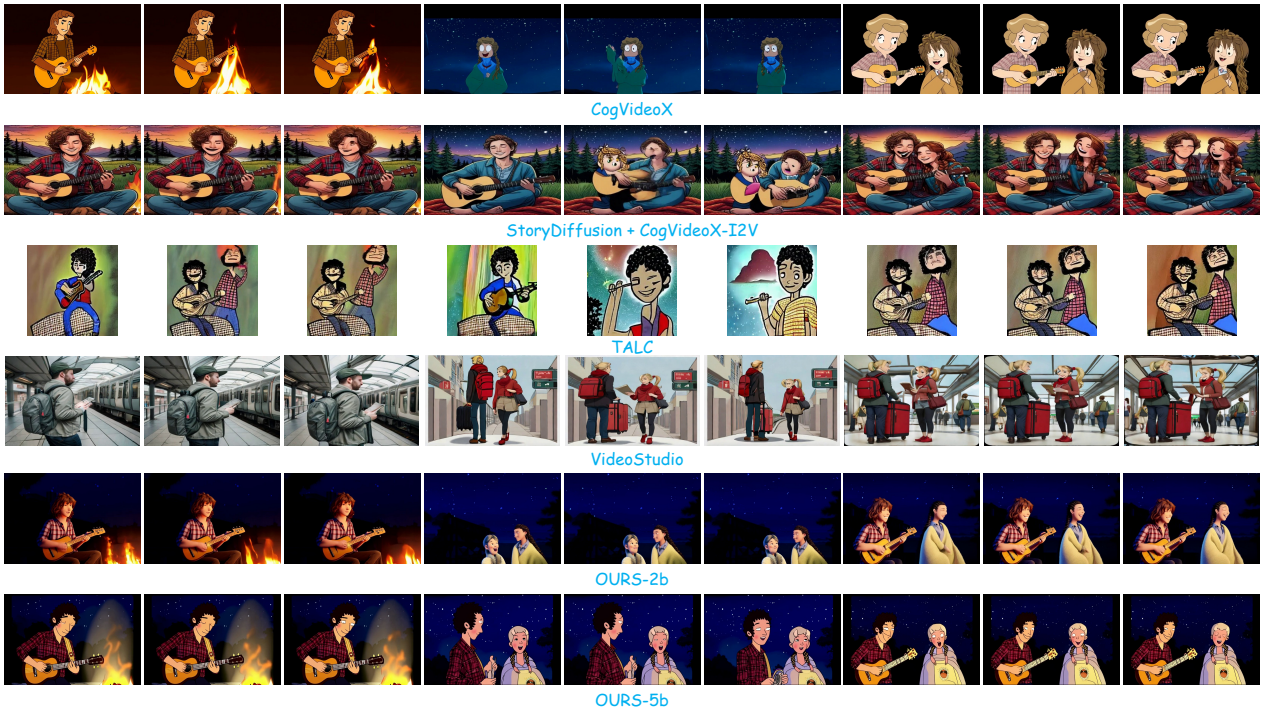


Figure 13. Visualization of Results from Different SOTA Methods. Video Captions: "Actor1 (wavy hair, plaid shirt, playing a ukulele) strums a soft melody by the bonfire.", "Actor2 (braided hair, cozy blanket, singing along) joins in, creating a serene harmony under the stars.", "Actor1 (wavy hair, plaid shirt, playing a ukulele) laughs softly as Actor2 (braided hair, cozy blanket, singing along) harmonizes, both enjoying the peaceful night."

Figure 14. Visualization of Results from Different SOTA Methods. Video Captions: "Actor1 (dark curly hair, navy sweater, reading a plaque) studies the information about an ancient artifact.", "Actor2 (wearing a scarf, glasses, admiring an exhibit) stands nearby, sharing an intrigued look with Actor1 (dark curly hair, navy sweater, reading a plaque).", "Actor1 (dark curly hair, navy sweater, reading a plaque) and Actor2 (wearing a scarf, glasses, admiring an exhibit) exchange thoughts about the exhibit, both captivated by the history around them."

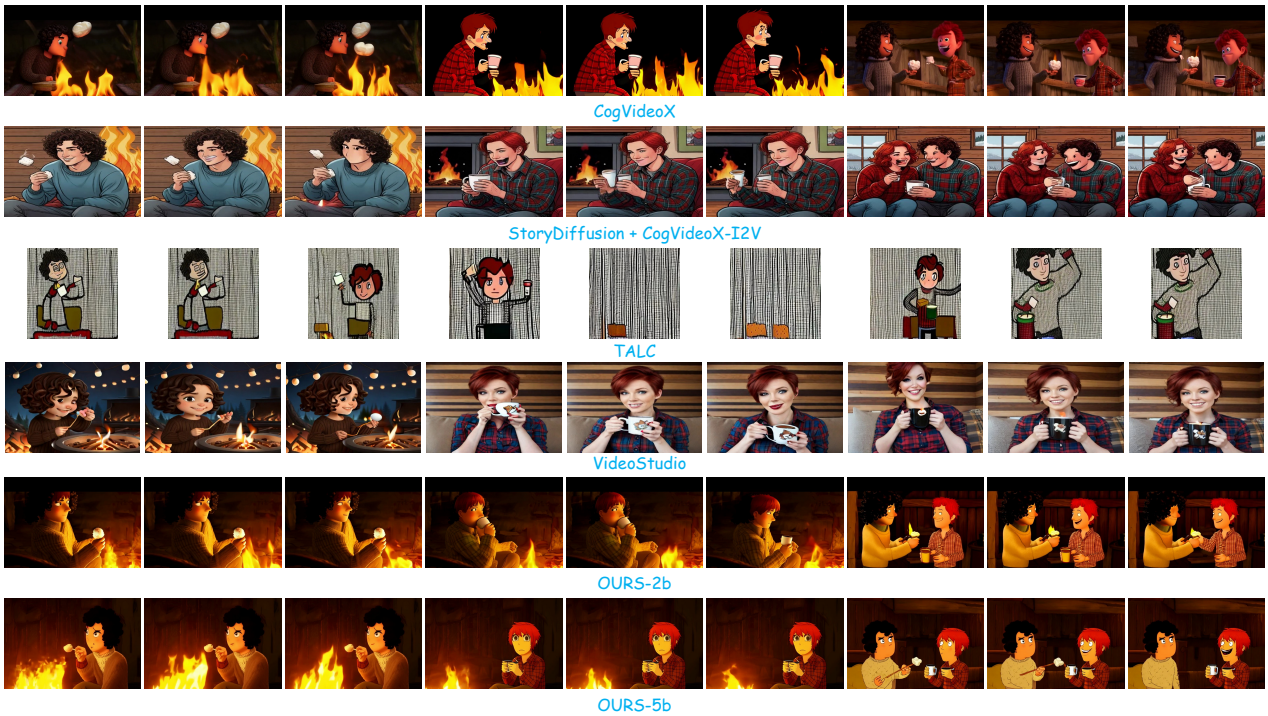


Figure 15. Visualization of Results from Different SOTA Methods. Video Captions: "Actor1 (dark curly hair, cozy sweater, roasting marshmallows) holds a marshmallow over the flames, watching it toast.", "Actor2 (short red hair, plaid shirt, holding a mug of hot cocoa) sits nearby, savoring the warmth of the fire.", "Actor1 (dark curly hair, cozy sweater, roasting marshmallows) offers a marshmallow to Actor2 (short red hair, plaid shirt, holding a mug of hot cocoa), both sharing laughter in the cozy cabin."

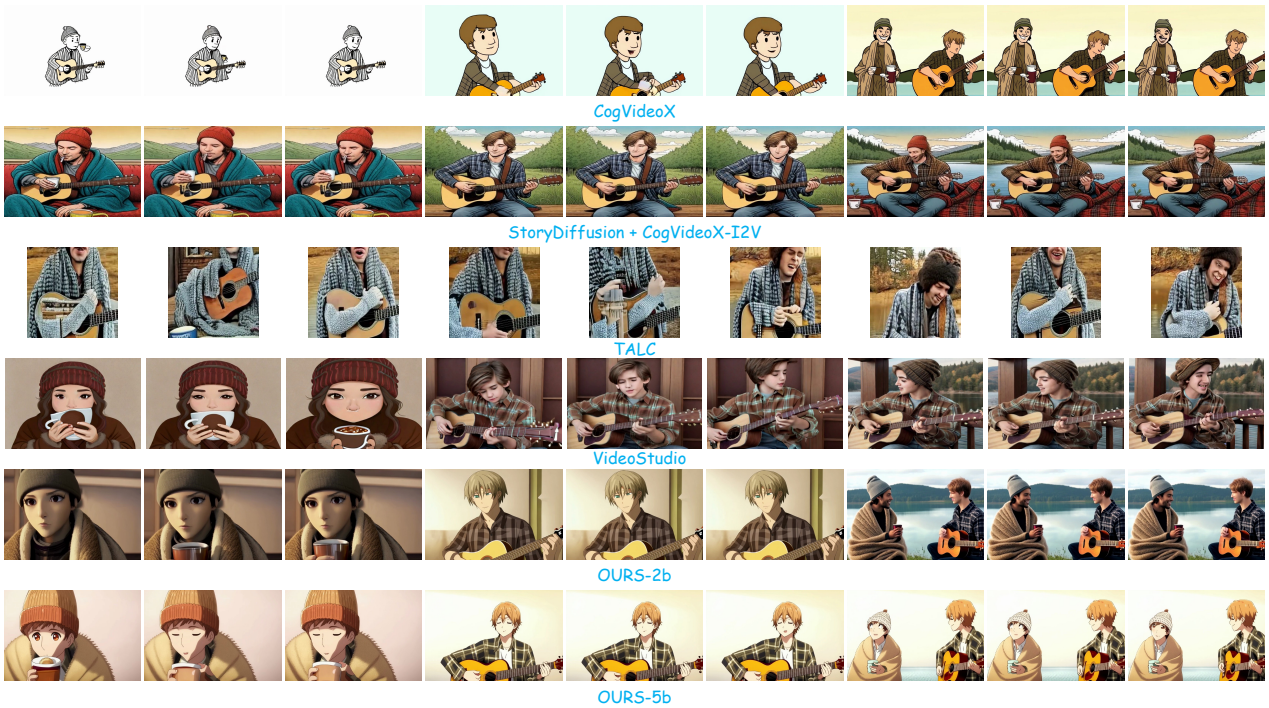


Figure 16. Visualization of Results from Different SOTA Methods. Video Captions: "Actor1 (wearing a beanie, wrapped in a wool blanket, holding a cup of tea) sips their drink, listening to the soft sounds of the guitar.", "Actor2 (light brown hair, in a flannel shirt, strumming a guitar) plays a gentle tune, filling the peaceful air.", "Actor1 (wearing a beanie, wrapped in a wool blanket, holding a cup of tea) smiles and relaxes, enjoying the moment with Actor2 (light brown hair, in a flannel shirt, strumming a guitar) by the lake."

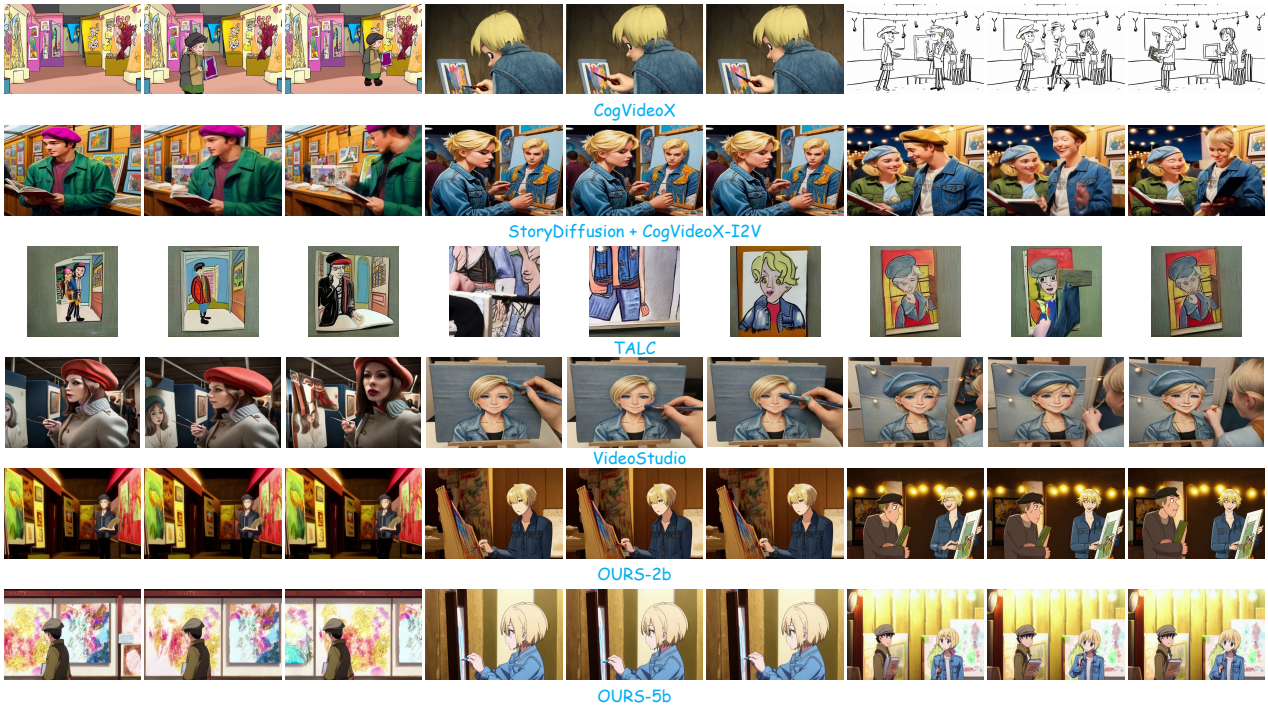


Figure 17. Visualization of Results from Different SOTA Methods. Video Captions: "Actor1 (wearing a beret, carrying a sketchbook, admiring artwork) walks through the booths, captivated by the vibrant colors.", "Actor2 (short blonde hair, denim jacket, painting on a small canvas) concentrates, adding finishing touches to their piece.", "Actor1 (wearing a beret, carrying a sketchbook, admiring artwork) stops to compliment Actor2 (short blonde hair, denim jacket, painting on a small canvas), both exchanging smiles under the string lights."