In the supplementary material, we provide more implementation details (Appendix A), including the hyperparameters used in training and inference. Then, we showcase additional comparisons with existing methods and more qualitative results (Appendix B). Furthermore, we discuss the social impacts and limitations (Appendix C).

A. Implementation Details

A.1. Hyperparameters

In Tab. 3, we provide an overview of the hyperparameter settings and conduct training based on the foundational text-to-video generation models of LTX-Video [22] and Wan-T2V [64]. The former allows for quick inference with limited resources; in an A100 single-card environment, without a dedicated acceleration strategy, it takes about 24 seconds to sample 40 steps for a video of approximately 5 seconds in duration. This meets the needs of general users for video processing. In contrast, Wan-T2V is a comprehensive performance video generation model that requires relatively more resources for training and inference, but it is capable of producing high-quality visuals and maintaining smooth temporal consistency.

B. Additional Results

B.1. More Visualization

In Fig. 6 and Fig. 7, we present more qualitative results based on Wan-T2V, which include tasks such as outpainting, inpainting, extension, grayscale, depth, scribble, pose, layout, face reference, and object reference.

B.2. Visualization Comparison

In Fig. 8, we present a visualization of the comparison of the VACE based on LTX-Video-2B [22] with others, including the extension task compared with I2VGenXL [77], CogVideoX [73], and LTX-Video-I2V [22]; the unconditional inpainting task compared with ProPainter [82]; the outpainting task with Follow-Your-Canvas [8] and M3DDM [17]; depth-controlled generation with Control-A-Video [10], VideoComposer [68], and ControlVideo [79]; pose-controlled generation with Text2Video-Zero [31], ControlVideo [79] and Follow-Your-Pose [40]; optical flow-controlled generation with FLATTEN [14]; and the reference task compared with commercially closed-source models Kling 1.6 [1], Pika 2.2 [49], and Vidu 2.0 [66].

C. Discussion

C.1. Limitations

First, the quality of generated content and the overall style are often influenced by the foundation model. This paper verifies this across different model scales: smaller models are advantageous for rapid video generation, but the quality and coherence of the videos are inevitably challenged; larger parameter models significantly improve the success rate of creative output, but the inference speed slows down, and resource consumption increases. Finding a relative balance between the two is also a key focus of our future work.

Secondly, compared to the foundational models for text-to-video generation, the current unified models have not been trained on large-scale data and computational power. This results in issues such as the inability to fully maintain identity during reference generation and a lack of complete control over inputs when performing compositional tasks. As discussed in the paper regarding full fine-tuning and additional parameter fine-tuning, when unified tasks begin to apply scaling laws, the results are promising.

In addition, the operational methods for the unified models, compared to image models, present certain challenges due to the inclusion of temporal information and various modalities in their inputs. This aspect creates a threshold for practical usage. Therefore, it is worth exploring how to effectively leverage the capabilities of existing language models or agent models to guide video generation and editing, thereby enhancing productivity.

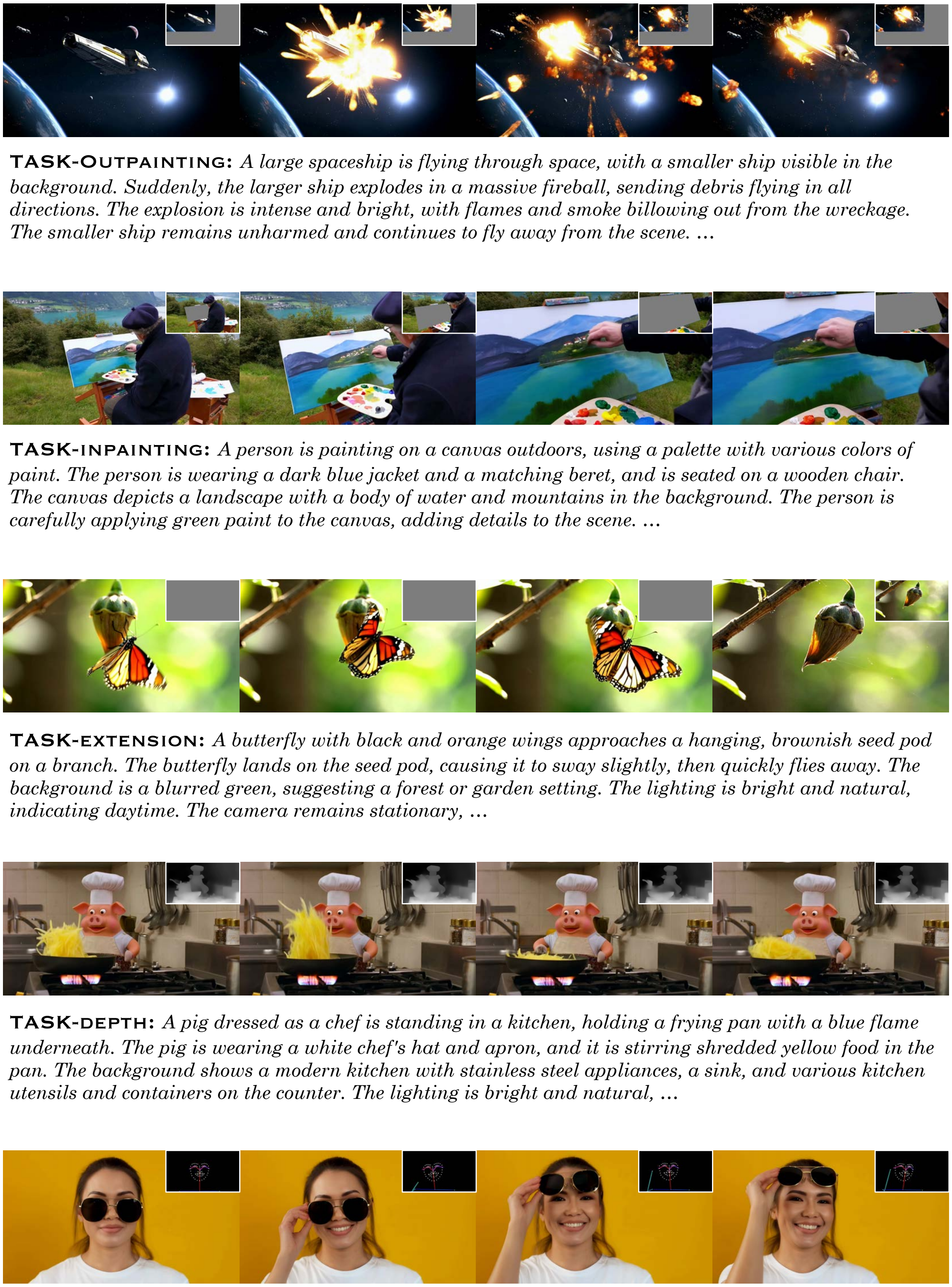
C.2. Societal impacts

From a positive perspective, intelligent video generation and editing can provide creators with a range of innovative tools, helping them to spark new ideas and enhance the artistic and innovative quality of video content. These technologies are gradually being applied across various industries; for example, in the business sector, video generation technology is transforming marketing and advertising strategies. Companies can quickly produce high-quality promotional videos, effectively communicating brand messages and attracting consumers. This ability to increase efficiency not only saves labor costs but also enables businesses to implement more creative marketing strategies, thus enhancing their market competitiveness.

However, with the proliferation of these technologies, certain social challenges have emerged. The convenience of video generation and editing may lead to the spread of misinformation and false content, undermining the public's trust in information. Additionally, when generating content, the technology may inadvertently reinforce existing biases and stereotypes, negatively impacting societal cultural perceptions. These issues prompt reflections on ethics and responsibility, calling for policymakers, technology developers, and various sectors of society to work together to establish appropriate regulations to ensure the healthy development of these technologies. We must also examine their potential impacts with a cautious attitude, actively exploring ways to balance innovation with social responsibility, so that they can deliver greater benefits to society.

Table 3. Hyperparameter selection for LTX-Video-based and Wan-T2V-based `VACE`.

Config	**#Model**	
	LTX-Video-based	Wan-T2V-based
Task	12 tasks + composition task	12 tasks + composition task
Batch Size / GPU	1	1/8
Accumulate Step	8	1
Optimizer	AdamW	AdamW
Weight Decay	0.1	0.1
Learning Rate	0.0001	0.00005
Learning Rate Schedule	Constant	Constant
Training Steps	200,000	200,000
Resolution	˜480p	˜720p
Shifting	Ture	True
Weighting Scheme	uniform	uniform
Sequence Length	4992	75600
Num Layers	28	40
Context Adapter	Res-Tuning	Res-Tuning
Context Layers	[0, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26]	[0, 5, 10, 15, 20, 25, 30, 35]
Concept Decouple	Ture	True
Pre-trained Model	LTX-Video-2b-v0.9	Wan2.1-T2V-14B
Sampler	Flow Euler	Flow Euler
Sample Steps	40	25
Guide Scale	3.0	4.0
Generation speed	˜24s	˜260s (8 gpus)
Device	A100×16	A100×128
Training Strategy	AMP / DDP / BFloat16	FSDP / Tensor Parallel / BFloat16



TASK-OUTPAINTING: *A large spaceship is flying through space, with a smaller ship visible in the background. Suddenly, the larger ship explodes in a massive fireball, sending debris flying in all directions. The explosion is intense and bright, with flames and smoke billowing out from the wreckage. The smaller ship remains unharmed and continues to fly away from the scene. ...*

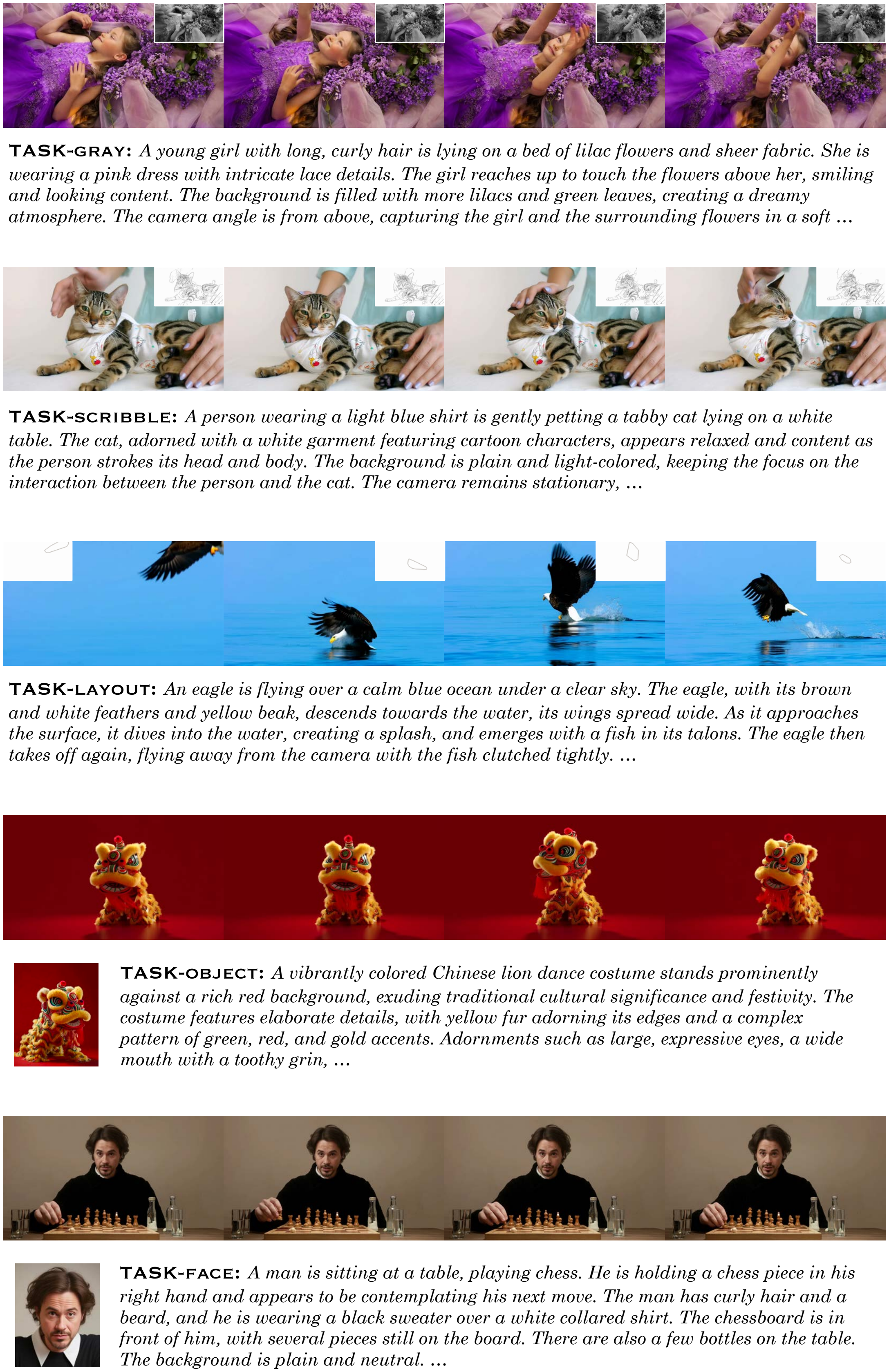
TASK-INPAINTING: *A person is painting on a canvas outdoors, using a palette with various colors of paint. The person is wearing a dark blue jacket and a matching beret, and is seated on a wooden chair. The canvas depicts a landscape with a body of water and mountains in the background. The person is carefully applying green paint to the canvas, adding details to the scene. ...*

TASK-EXTENSION: *A butterfly with black and orange wings approaches a hanging, brownish seed pod on a branch. The butterfly lands on the seed pod, causing it to sway slightly, then quickly flies away. The background is a blurred green, suggesting a forest or garden setting. The lighting is bright and natural, indicating daytime. The camera remains stationary, ...*

TASK-DEPTH: *A pig dressed as a chef is standing in a kitchen, holding a frying pan with a blue flame underneath. The pig is wearing a white chef's hat and apron, and it is stirring shredded yellow food in the pan. The background shows a modern kitchen with stainless steel appliances, a sink, and various kitchen utensils and containers on the counter. The lighting is bright and natural, ...*

TASK-POSE: *A young woman with curly hair and wearing a white shirt is standing against a yellow background. She is smiling and looking at the camera while holding her sunglasses up to her forehead with her right hand. The woman has dark skin, and her hair is styled in loose curls that fall around her shoulders. She is wearing large, round sunglasses with a gold frame and red lenses. ...*

Figure 6. **More visualization results** of Wan-T2V-based `VACE` framework.



TASK-GRAY: *A young girl with long, curly hair is lying on a bed of lilac flowers and sheer fabric. She is wearing a pink dress with intricate lace details. The girl reaches up to touch the flowers above her, smiling and looking content. The background is filled with more lilacs and green leaves, creating a dreamy atmosphere. The camera angle is from above, capturing the girl and the surrounding flowers in a soft …*

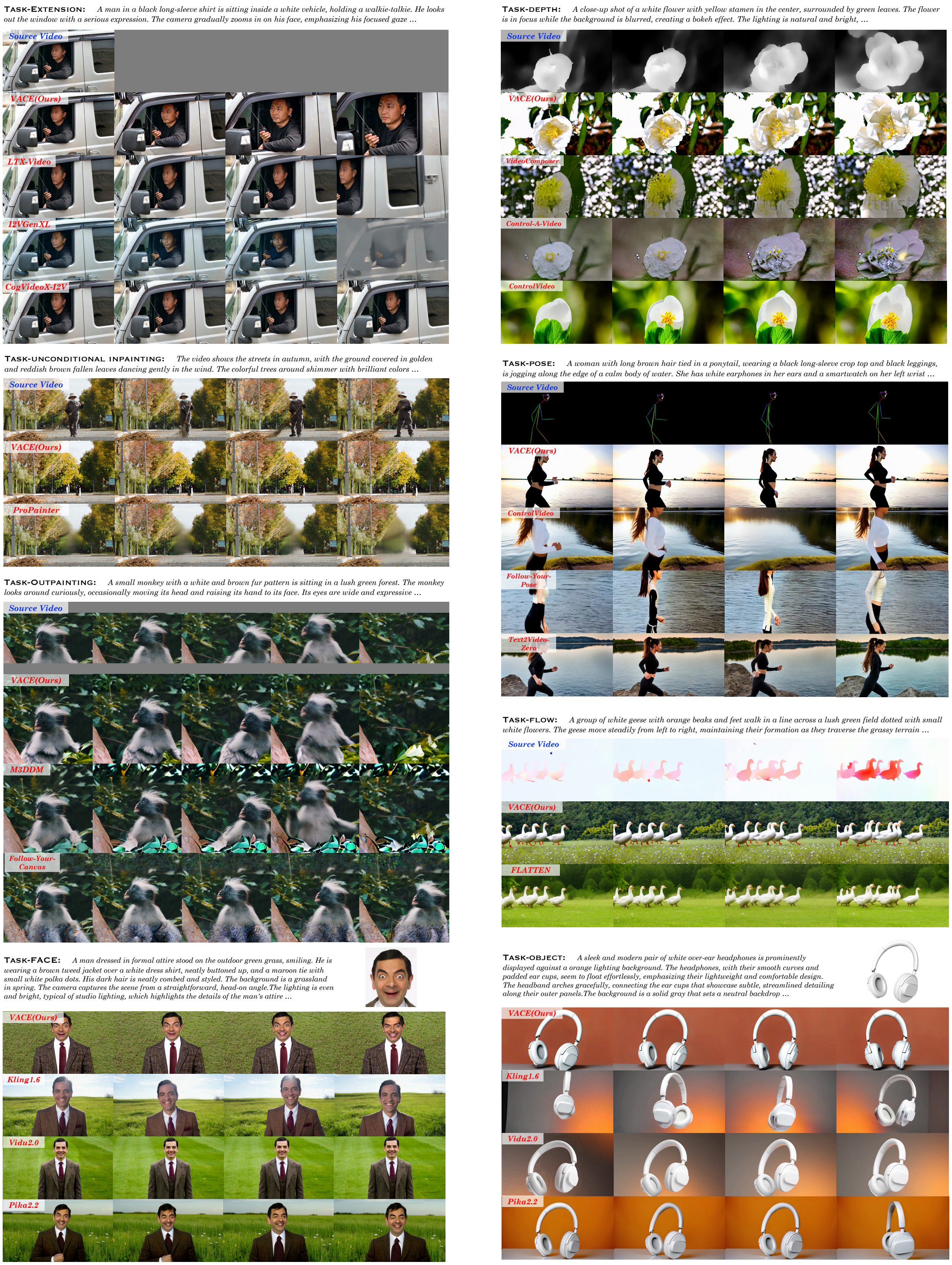
TASK-SCRIBBLE: *A person wearing a light blue shirt is gently petting a tabby cat lying on a white table. The cat, adorned with a white garment featuring cartoon characters, appears relaxed and content as the person strokes its head and body. The background is plain and light-colored, keeping the focus on the interaction between the person and the cat. The camera remains stationary, …*

TASK-LAYOUT: *An eagle is flying over a calm blue ocean under a clear sky. The eagle, with its brown and white feathers and yellow beak, descends towards the water, its wings spread wide. As it approaches the surface, it dives into the water, creating a splash, and emerges with a fish in its talons. The eagle then takes off again, flying away from the camera with the fish clutched tightly. …*

TASK-OBJECT: *A vibrantly colored Chinese lion dance costume stands prominently against a rich red background, exuding traditional cultural significance and festivity. The costume features elaborate details, with yellow fur adorning its edges and a complex pattern of green, red, and gold accents. Adornments such as large, expressive eyes, a wide mouth with a toothy grin, …*

TASK-FACE: *A man is sitting at a table, playing chess. He is holding a chess piece in his right hand and appears to be contemplating his next move. The man has curly hair and a beard, and he is wearing a black sweater over a white collared shirt. The chessboard is in front of him, with several pieces still on the board. There are also a few bottles on the table. The background is plain and neutral. …*

Figure 7. **More visualization results** of Wan-T2V-based `VACE` framework.



Task-Extension: *A man in a black long-sleeve shirt is sitting inside a white vehicle, holding a walkie-talkie. He looks out the window with a serious expression. The camera gradually zooms in on his face, emphasizing his focused gaze ...*

Task-unconditional inpainting: *The video shows the streets in autumn, with the ground covered in golden and reddish brown fallen leaves dancing gently in the wind. The colorful trees around shimmer with brilliant colors ...*

Task-Outpainting: *A small monkey with a white and brown fur pattern is sitting in a lush green forest. The monkey looks around curiously, occasionally moving its head and raising its hand to its face. Its eyes are wide and expressive ...*

Task-FACE: *A man dressed in formal attire stood on the outdoor green grass, smiling. He is wearing a brown tweed jacket over a white dress shirt, neatly buttoned up, and a maroon tie with small white polka dots. His dark hair is neatly combed and styled. The background is a grassland in spring. The camera captures the scene from a straightforward, head-on angle.The lighting is even and bright, typical of studio lighting, which highlights the details of the man's attire ...*

Task-depth: *A close-up shot of a white flower with yellow stamen in the center, surrounded by green leaves. The flower is in focus while the background is blurred, creating a bokeh effect. The lighting is natural and bright, ...*

Task-pose: *A woman with long brown hair tied in a ponytail, wearing a black long-sleeve crop top and black leggings, is jogging along the edge of a calm body of water. She has white earphones in her ears and a smartwatch on her left wrist ...*

Task-flow: *A group of white geese with orange beaks and feet walk in a line across a lush green field dotted with small white flowers. The geese move steadily from left to right, maintaining their formation as they traverse the grassy terrain ...*

Task-object: *A sleek and modern pair of white over-ear headphones is prominently displayed against a orange lighting background. The headphones, with their smooth curves and padded ear cups, seem to float effortlessly, emphasizing their lightweight and comfortable design. The headband arches gracefully, connecting the ear cups that showcase subtle, streamlined detailing along their outer panels.The background is a solid gray that sets a neutral backdrop ...*

Figure 8. **Qualitative comparisons** on various tasks based on LTX-Video-based `VACE` framework.