Prompt-A-Video:
Prompt Your Video Diffusion Model via Preference-Aligned LLM

Supplementary Material

A. Limitations

The main limitation lies in the generalization of Prompt-A-Video, which requires new curated data for fine-tuning when adapting to new video generation models. We considered developing a universal prompt booster. However, we found that an LLM trained for one text-to-video model showed no effect when applied to another model, indicating generation models have different prompt domain preferences.

When applying a video prompt boost model (always generates long prompts) directly to text-to-image, it has text length truncation issue due to limited processing length of CLIP's text encoder, so the performance drops.

B. Reward-guided Prompt Evolution

Our reward-guided prompt evolution pipeline uses GPT-4o as evolutionary operator to automatically generate model-preferred prompts. As illustrated in Table 6, GPT-4o is guided to learn experience from historical refined prompts and generate new prompts of higher scores.

Table 5 shows an example of prompt refinement in the first iteration. Guided by system instructions and provided examples, GPT-4o generates three different enhanced descriptions, whose scores have improvements of different degrees compared to the original prompt.

Notably, our pipeline differs slightly from conventional evolutionary algorithms. Traditional evolutionary algorithms retain the top N results and sample K candidates from them for the next iteration based on their scores as probabilities, introducing randomness to avoid local optima. In our approach, since GPT's generation inherently possesses stochastic properties - producing varied outputs even with identical inputs - we directly advance the top N results to the next iteration.

Attractive blonde woman doing hand massage in a spa center	VQ: 2.47, TC: 2.66, DD: 2.84, TVA: 2.77, FC: 2.48, AES: 3.34, MPS: 2.7
A serene scene in a spa center where an attractive blonde woman is performing a hand massage. The woman has a focused expression and is working gently. The surroundings are tranquil, with soft lighting and calming decor. The environment suggests a soothing and relaxing experience. The video does not contain any text or drastic actions.	VQ: 2.63, TC: 2.73, DD: 2.92, TVA: 2.95, FC: 2.42, AES: 3.49, MPS: 3.67
A close-up of a blonde woman giving a hand massage in a quiet spa center. The woman's hands are applying gentle pressure, with a serene and focused expression on her face. The spa environment features soft lighting, adding to the calm and relaxing atmosphere. The scene is peaceful and intimate, designed to convey comfort and care. No text or significant movements are present in the video.	VQ: 2.58, TC: 2.77, DD: 2.88, TVA: 2.98, FC: 2.56, AES: 3.47, MPS: 3.04
An intimate view of a serene spa center with a blonde woman performing a hand massage. The woman appears focused and gentle, with the peaceful surroundings enhancing the calming effect. Soft, ambient lighting highlights the tranquility of the spa. The overall ambiance is relaxing, aiming to provide a sense of comfort and well-being. The video maintains a focus on the woman and her actions without any text	VQ: 2.58, TC: 2.69, DD: 2.77, TVA: 2.88, FC: 2.47, AES: 3.61, MPS: 3.43

Table 5. An example of the first iteration. The scores are normalized to 0 - 5.

C. Training Settings

For model training, we construct around 3,000 prompt pairs for each video generation model, and use them to fine-tune Llama3-Instruct in a chat format with LoRA. The instruction prompt for LLama3-Instruct is shown in Table 7. The supervised fine-tuning is deployed for 14 epochs, with a batch size of 16 and a learning rate of 1e-4. During the DPO phase, approximately 2,000 prompt triples are used for preference optimization. This stage is trained for 3 epochs with a batch size of 32 and a learning rate of 5e-5.

Given that VBench encompasses a large number of prompts, for ablation study, we meticulously curate a representative subset of 100 prompts in VBench. Prompts are selected for containing dynamic action from four dimensions of VBench: human action, overall consistency, appearance style, and subject consistency. Apart from five metrics from VideoScore, we incorporate two VBench metrics, aesthetic quality and imaging quality, due to their distinguishing values acorss methods.

You need to refine user's input prompt. The user's input prompt is used for video generation task. You need to refine the user's prompt to make it more suitable for the task. Here are some examples of refined prompts:
a close-up shot of a woman standing in a room with a white wall and a plant on the left side. the woman has curly hair and is wearing a green tank top. she is looking to the side with a neutral expression on her face. the lighting in the room is soft and appears to be natural, coming from the left side of the frame. the focus is on the woman, with the background being out of focus. there are no texts or other objects in the video. the style of the video is a simple, candid portrait with a shallow depth of field.
a serene scene of a pond filled with water lilies. the water is a deep blue, providing a striking contrast to the pink and white flowers that float on its surface. the flowers, in full bloom, are the main focus of the video. they are scattered across the pond, with some closer to the camera and others further away, creating a sense of depth. the pond is surrounded by lush greenery, adding a touch of nature to the scene. the video is taken from a low angle, looking up at the flowers, which gives a unique perspective and emphasizes their beauty. the overall composition of the video suggests a peaceful and tranquil setting, likely a garden or a park.
a serene scene in a park. the sun is shining brightly, casting a warm glow on the lush green trees and the grassy field. the camera is positioned low, looking up at the towering trees, which are the main focus of the image. the trees are dense and full of leaves, creating a canopy of green that fills the frame. the sunlight filters through the leaves, creating a beautiful pattern of light and shadow on the ground. the overall atmosphere of the video is peaceful and tranquil, evoking a sense of calm and relaxation.
a scene where a person is examining a dog. the person is wearing a blue shirt with the word "volunteer" printed on it. the dog is lying on its side, and the person is using a stethoscope to listen to the dog's heartbeat. the dog appears to be a golden retriever and is looking directly at the camera. the background is blurred, but it seems to be an indoor setting with a white wall. the person's focus is on the dog, and they seem to be checking its health. the dog's expression is calm, and it seems to be comfortable with the person's touch. the overall atmosphere of the video is calm and professional.
...
The refined prompt should pay attention to all objects in the video. The description should be useful for AI to re-generate the video. The description should be no more than six sentences. The refined prompt should be in English.
User will provide an original prompt and your revised prompts, with their generated videos' scores (Visual Quality, Temporal Consistency, Dynamic Degree, Text Video Alignment, Factual Consistency, Aesthetic score, Image quality, 7 dimensions termed as VQ, TC, DD, TVA, FC, AES, MPS), and you need to give an improved prompt according to previous prompts and their scores on different dimensions.
Each prompt is tagged with an index, and the sentence labeled as 0 is the initial prompt. Each prompt is followed by (VQ, TC, DD, TVA, FC, AES, MPS) scores. You need build upon the most successful prompts and learn from the high-scoring prompts. You need to observe the scores of each prompt in different aspects, learn from the experiences of previous prompts, and combine their strengths to generate better prompts.
The new prompts should keep the same semantic meaning with original prompt, should not add extra scene changing or too many actions, which is hard for video generation.
Generate 3 paraphrases of the initial prompt which keep the semantic meaning and that have higher scores than all the prompts above. Respond with each new prompt in between <PROMPT>and </PROMPT>, e.g., <PROMPT>paraphrase 1</PROMPT>.

Table 6. The evolution instruction prompt for GPT-4o.

D. Detailed Results

Table 8 presents the comprehensive evaluation results across all VBench metrics. While our method achieves consistent improvements in Quality Score, the Semantic Score demonstrates minimal underperformance compared to LLM (GLM-4 or GPT-4o). This discrepancy primarily stems from two technical considerations: (1) During fine-tuning and DPO preference learning phases, the reward model predominantly prioritize visual fidelity metrics and aesthetic assessment criteria, resulting in suboptimal semantic consistency optimization; (2) The absence of style

”dialog”: [
{
”role”: ”user”,
”content”: ”You need to refine user’s input prompt. The user’s input prompt is used for video generation task. You need to refine the user’s prompt to make it more suitable for the task. 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. You will only ever output a single video description per user request. You should refactor the entire description to integrate the suggestions. Original prompt:\n” + **original prompt** + ”\n New prompt:\n”
},
{
”role”: ”assistant”,
”content”: **refined prompt**
}
]

Table 7. The instruction prompt for fine-tuning LLama3-Instruct.

descriptors in our training corpus imposes limited improvements on style consistency preservation.

E. Discussions

E.1. Model-agnostic prompt boost

We combine fine-tuning data corresponding to each generation model to train LLaMA together, aiming to capture universal patterns in video prompt enhancement, so that the prompt booster could generalize to a broader range of video generation models. However, the suboptimal performance indicates distinct prompt preferences across different models, highlighting the importance of our preference-aligned approach.

E.2. Negative prompts

Beyond prompt enhancement, we also explore approaches for negative prompt generation, which guides the model's focus towards preferred attributes and helps avoid unwanted aspects. While existing text-to-video models commonly employ fixed negative prompts, a question arises: Can input-specific negative prompts yield superior results by reducing particular undesired attributes associated with subjects or actions in each input? To investigate this, we implement two adaptive negative prompt generation methods.

The first method involves leveraging the in-context learning capability of the LLM. Specifically, we manually create a small number of high-quality refined-negative prompt pairs. These pairs then serve as few-shot examples, guiding the LLM to generate adaptive negative prompts based on the input prompts. The other approach is to mimic the Prompt-A-Video posotive prompt generation process: first, fine-tune with curated prompt-negative prompt pairs, and then conduct preference alignment optimization.

The fixed negative prompt is shown in Table 9. As for adaptive negative prompts, considering our refined prompts from previous stages incorporate various positive modifiers, we can derive antonyms of these modifiers and negative descriptions of the subject or overall video characteristics to construct negative prompts. An exmaple is shown in Table 9.

Notably, we structure these negative prompts as comma-separated descriptors rather than complete sentences containing the original prompt's content, as our empirical analysis reveals that subject-containing negative prompts can interfere with accurate generation.

Our experimental results on the VBench for CogVideoX, as shown in Table 10, demonstrate that incorporating negative prompts can further enhance the performance of our videos. However, neither the in-context learning method nor the fine-tuning with preference alignment approach shows significant improvements compared to fixed negative prompts. It indicates that a comprehensive set of negative terms within fixed negative prompts is sufficient to achieve substantial performance improvements.

F. More Visualization

As shown in Figure 6, we qualitatively compare the videos generated with user prompts, GLM4-refined prompts and Prompt-A-Video. The results generated with Prompt-A-Video exhibit superior visual quality and enhanced motion magnitude.

Method	*Total Score*	*Quality Score*	*Semantic Score*	*subject consistency*	*background consistency*	*temporal flickering*	*motion smoothness*	*dynamic degree*	*aesthetic quality*	*imaging quality*	*object class*	*multiple objects*	*human action*	*color*	*spatial relationship*	*scene*	*appearance style*	*temporal style*	*overall consistency*
CogVideoX																			
Original prompts	78.74	80.98	69.75	95.41	96.25	96.47	97.33	63.89	59.07	63.68	80.38	54.5	85.00	87.46	51.02	43.02	23.40	23.31	25.90
Promptist	76.78	80.74	60.97	95.78	96.13	97.16	97.35	52.78	60.26	64.21	56.88	22.94	80.00	84.39	53.54	30.60	23.28	22.77	25.38
GLM-4	81.05	82.00	77.26	96.00	96.70	97.10	97.67	56.94	63.81	64.81	94.30	73.02	96.00	81.69	62.37	50.65	23.53	24.52	27.89
Prompt-A-Video	81.96	83.28	76.67	95.82	96.72	96.55	98.15	65.28	65.34	67.46	93.83	77.13	94.00	82.40	65.67	52.03	21.88	23.52	26.40
Open-Sora 1.2																			
Original prompts	76.56	78.26	69.76	84.95	93.92	97.22	96.13	83.33	54.74	58.06	82.04	45.50	87.00	86.64	58.05	36.26	24.08	24.91	26.13
Promptist	76.49	79.05	66.26	82.89	93.28	99.03	96.84	76.39	55.89	61.48	79.83	40.93	77.0	86.58	58.42	29.0	23.04	24.45	25.68
GPT-4o	78.70	79.23	76.58	89.91	96.03	98.37	96.94	62.50	56.26	58.72	90.82	57.83	94.0	92.36	65.14	55.96	22.80	24.42	26.98
Prompt-A-Video	78.46	79.78	73.20	88.16	95.83	97.87	97.06	72.22	57.18	59.8	92.41	54.58	92.00	92.11	61.15	39.94	22.94	24.57	25.52

Table 8. **Comparison with other prompt boost methods on VBench benchmark.**

Fixed negative prompt

The video is not of a high quality, it has a low resolution, and the audio quality is not clear. Strange motion trajectory, a poor composition and deformed video, low resolution, duplicate and ugly, strange body structure, long and strange neck, bad teeth, bad eyes, bad limbs, bad hands, rotating camera, blurry camera, shaking camera. Deformation, low-resolution, blurry, ugly, distortion.

Adaptive negative prompt example

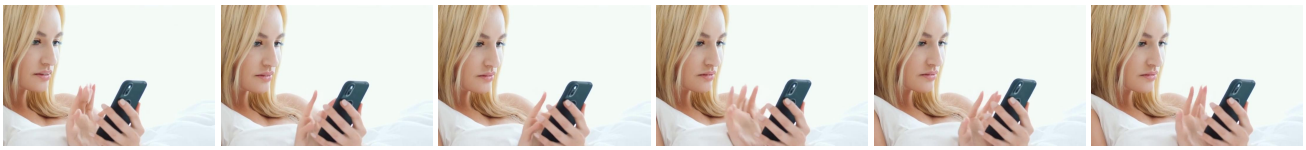
refined prompt: As the sun gently peeks through the vibrant window curtains, I sit comfortably in my plush, velvety chair, surrounded by an array of artfully arranged cosmetics. I begin by applying a lightweight, radiant foundation, seamlessly blending it into my skin with a fluffy brush. Next, I define my eyes with a rich, earthy eyeshadow, gradually building the palette with a pop of shimmering champagne on the lid. A swipe of deep, berry-stained lipstick completes my morning glow. I finish with a quick mist of hydrating toner and a light dusting of translucent powder, leaving my complexion fresh and flawless. The video captures my calm and focused routine, highlighting the beauty in the simplicity of a morning makeup ritual, as the natural light dancing through the room highlights the subtle, yet elegant enhancement of my daily beauty regimen.

negative prompt: The video is not of a high quality, it has a low resolution. Strange motion trajectory. bad hands, missing fingers. deformation, distortion. Dark and unclear, blur, ugly, watermark, static. The person has bad anatomy, bad eyes, bad teeth, long and strange neck, bad hands, text, error, ugly appearance, deformed body, poorly drawn face, long body. The makeup is poorly applied, with uneven lines and clashing colors. The lip color is tacky and over-saturated, and the toner makes the skin look dull and oily. The camera work is shaky and poorly framed, with harsh lighting that accentuates imperfections.

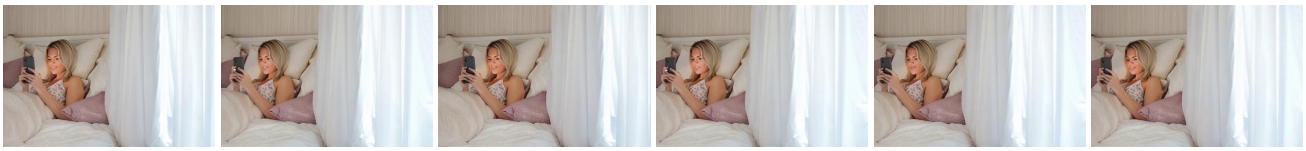
Table 9. Negative prompt strategies.

Prompts	SC	BC	AQ	IQ	MS	DD
Prompt-A-Video	0.948	0.960	0.627	0.662	<u>0.982</u>	**0.52**
+ NP (fixed)	0.955	0.964	**0.639**	<u>0.682</u>	**0.986**	0.46
+ NP (ICL)	<u>0.958</u>	<u>0.965</u>	<u>0.637</u>	**0.683**	**0.986**	<u>0.50</u>
+ NP (tuning)	**0.959**	**0.966**	0.634	<u>0.682</u>	**0.986**	0.48

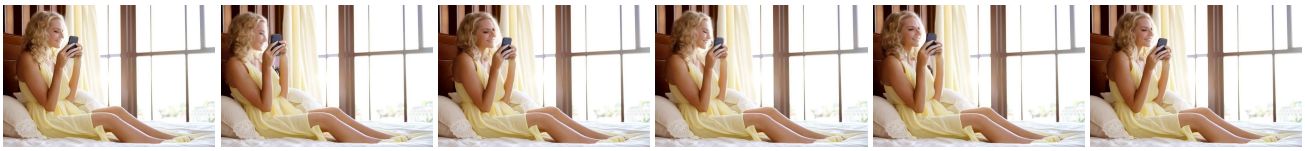
Table 10. Ablation studies about different negative prompts generation methods. NP indicates negative prompts.



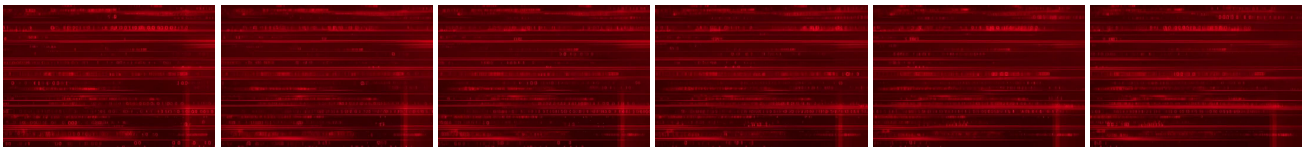
Original prompt: Beautiful blonde woman using smartphone in bed



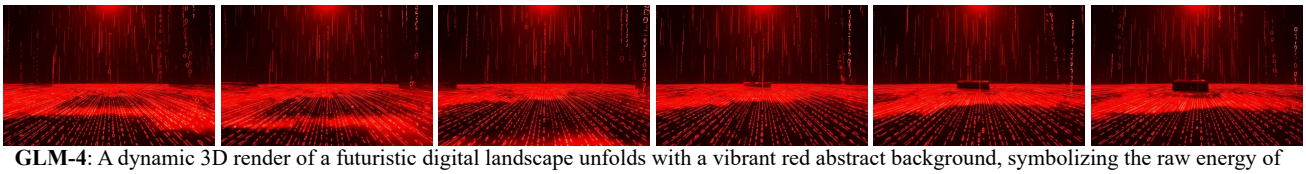
GLM-4: A stunning blonde woman, with her hair cascading over her shoulders, relaxes on a plush, white duvet. She's nestled comfortably among a collection of soft, pastel pillows. Clad in a delicate, floral nightgown, she's engrossed in the glow of her smartphone, the screen illuminating her features with a gentle light. The morning sun filters through the white, sheer curtains, casting a warm, ambient glow in the cozy bedroom setting.



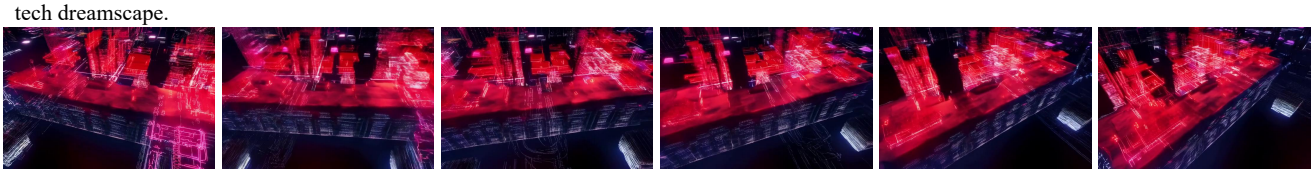
Prompt-A-Video: A stunning blonde woman, with sun-kissed tresses cascading down her frame, sits cozily in a plush, cream-colored bed. She's lounging against a rich, wooden headboard, where a pillow with intricate lace trim lies beneath her. She's dressed in a soft, pale yellow summer sundress that drapes elegantly around her, the hem fluttering slightly as she sways to the rhythm of the music playing on her smartphone. The sunlight streaming through the large, floor-to-ceiling windows highlights the gentle curls of her hair and the warm glow of the device in her hand, as she smiles subtly to herself, enjoying the peaceful morning moment, surrounded by the serene calm of her serene bedroom.



Original prompt: Digital server data. color red abstract background of virtual space with binary animation. matrix style futuristic 3d render



GLM-4: A dynamic 3D render of a futuristic digital landscape unfolds with a vibrant red abstract background, symbolizing the raw energy of virtual space. Swirling patterns of binary code cascade down the screen in a mesmerizing animation, reminiscent of the iconic matrix style. At the center, sleek servers hum silently, their surfaces glinting with holographic interfaces. The scene pulses with life, as if the very fabric of the digital realm is being woven before your eyes, with streams of ones and zeroes dancing in perfect harmony across the expanse of this high-tech dreamscape.



Prompt-A-Video: In a dystopian metropolis of glowing fibers, a majestic urban landscape unfolds. A colossal digital server sprawls like a concrete behemoth, its metal hide sheathed in a vibrant red hue that evokes both warnings and vitality. The abstract binary animations dancing upon its surface are a visual representation of the intricate code that governs this synthetic universe. The 3D render captures the Matrix-inspired cityscape at dusk, where neon lights reflect off the server's steel skin, casting a web of light upon the concrete foundations that anchor this futuristic architecture. The visuals conclude with a sweeping aerial view, highlighting the server as a sentinel of progress, standing tall amidst a metropolis that pulsates with the rhythm of innovation.

Figure 6. Videos generated using CogVideoX with user prompts, GLM4-refined prompts and Prompt-A-Video.