

## 2 Supplementary Material: Frame-Level Captions for Long Video 3 Generation with Complex Multi Scenes

### 4 A Issues of Semantic Confusion and Error Accumulation

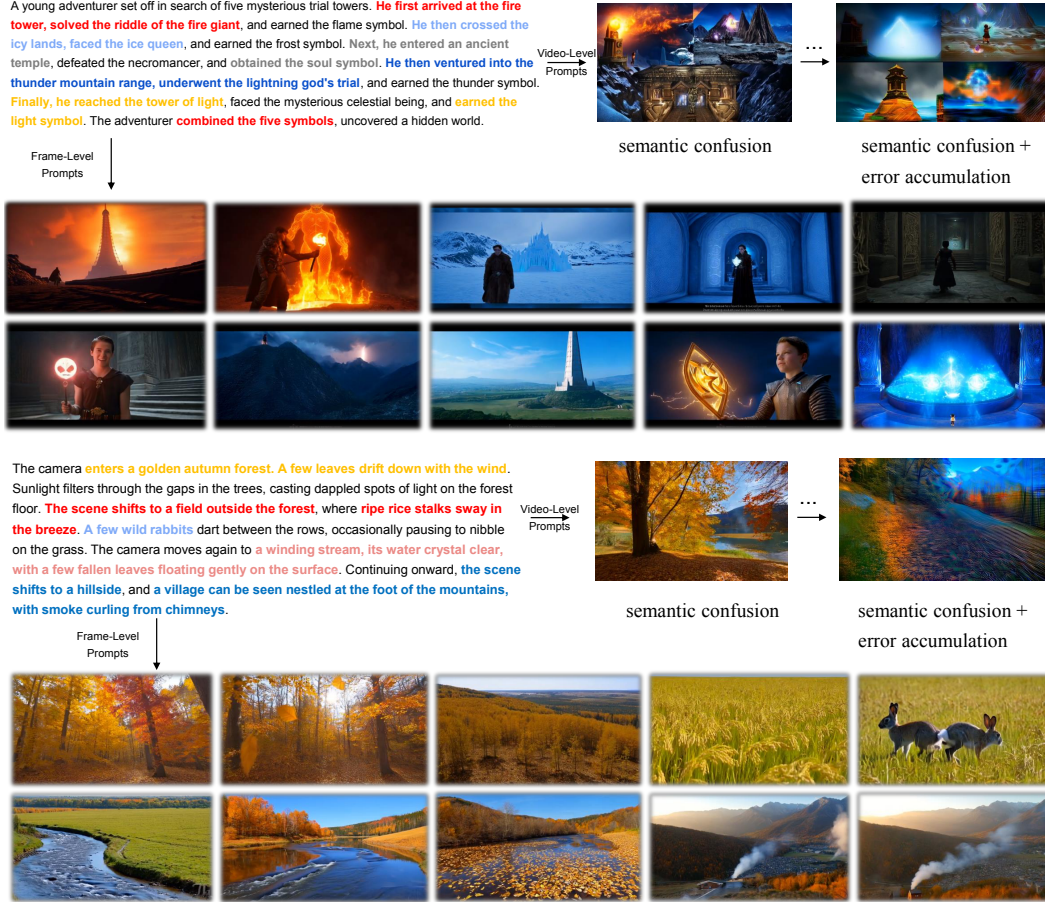


Figure 1: **Illustration of Semantic Confusion and Error Accumulation in Video Generation.** This figure contrasts 30s videos generated by conventional video-level prompt (right) and our proposed frame-level prompt (ten video key frames below), demonstrating improved generation quality of our method in aspects of semantic confusion and error accumulation.

Error *accumulation or drifting* is recognized a key issue in ultra-long video generation. This paper, however, focuses on *semantic confusion*, a less explored concept. This is because earlier methods for ultra-long video generation often addressed simpler scenarios, such as a single character in a consistent scene. Such scenarios inherently have lower semantic information density in their prompts, typically lacking complex interactions, multiple scene transitions, or varied camera perspectives. However, for complex scenes, such as those in narratives, films, or vlogs with multiple segments, using a single video-level prompt often leads to confusion, mainly due to the model’s low ability in focusing attention on diverse semantic elements when processing a single, overarching prompt. As illustrated in Fig. 1, this can cause significant semantic confusion and error accumulation, preventing the accurate generation of multiple distinct scenes. To address these challenges, we propose **dynamic prompts**, where each distinct event or segment in the video receives an individual prompt. Furthermore, our work utilizes *frame-level prompts*. Unlike *shot-level* approaches that describe entire shots, frame-level prompts detail the content on a per-frame basis. This method eliminates the need for explicit shot segmentation, significantly **reducing the difficulty and cost of data collection and annotation**. This, in turn, **facilitates the creation of larger datasets**.

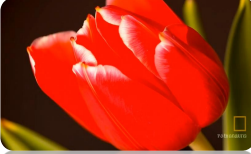
However, it is noted that frame-level prompting requires high-quality data. Consequently, the current performance of our method is **somewhat limited by the capabilities of existing Vision-Language**

22 **Models (VLMs).** Nevertheless, we anticipate this approach will show improved performance as  
 23 VLM technology advances and addresses current limitations.

## 24 **B More Application By enabling Frame-level Prompts**

### 25 **B.1 (Short Video) Key Frame / High Dynamics Generation Ability**

1. A highly detailed, cinematic time-lapse: a single, vibrant **red tulip** bud slowly unfurls its petals over several hours, revealing its delicate inner structure and rich color under soft studio lighting. Focus on the graceful, organic movement and intricate textures of the petals. Beautiful, mesmerizing, botanical art.



3. A young woman with **auburn hair**, wearing a chic trench coat, sits at an outdoor table of a bustling Parisian cafe on a bright autumn day. She **sips a coffee**, observing the lively street scene with a smile – people walking by, classic Parisian architecture, falling yellow leaves.



2. Overhead close-up shot in a professional kitchen: **A chef's hands meticulously plate a gourmet dessert.** They artfully arrange components like a chocolate sphere, berry coulis, edible flowers, and a sprinkle of powdered sugar on a pristine white plate. Focus on the precision, colors, and textures of the culinary art. Sophisticated, delicious visuals.

4. A spectacular nighttime drone light show: hundreds of coordinated, brightly colored LED drones form intricate, evolving 3D patterns – **perhaps a giant flying dragon or abstract geometric shapes.** A crowd below watches in awe. Stunning aerial choreography, vibrant colors against the night sky, festive.



1. A professional tennis player, in a vibrant sports outfit, executes a powerful forehand stroke during a championship match on a sunlit blue hard court. Beads of sweat are visible; intense focus on their face. The tennis ball blurs slightly from the speed. Focus on the athletic prowess, power, and precision. Bright stadium lighting.

2. A flock of thousands of starlings (a murmuration) performs an incredibly complex, fluid aerial ballet against a dramatic sunset sky. The birds move as one organism, creating mesmerizing, ever-changing dark shapes against the vibrant orange, pink, and purple clouds. Awe-inspiring natural phenomenon, beautiful synchronized movement.

3. Close-up, slow-motion detail of a honeybee diligently collecting nectar from a vibrant sunflower. Grains of yellow pollen dust its fuzzy body. The bee's delicate wings beat rapidly. Focus on the intricate details of the bee and flower, and the essential process of pollination. Macro beauty, natural world, vibrant colors.

4. A Formula 1 racing car, sleek and brightly colored, speeds down the main straight of a famous race track (e.g., Monaco or Silverstone) during a sunny race day, then brakes hard into a sharp corner, with sparks flying lightly from its undercarriage. Other cars are visible in the background. High speed, thrilling, focus on the car's engineering and the intensity of motorsport.

5. Time-lapse of a coastal scene showing the dramatic rise and fall of the tide over several hours. Water slowly engulfs rocks and then recedes, revealing tide pools and wet sand. The sun moves across the sky, changing the light and shadows on the landscape. Coastal beauty, powerful natural cycle, calming yet dynamic.

1. **New York's Grand Central Terminal** during morning rush hour, a captivating time-lapse. Streams of diverse commuters crisscross the sunlit main concourse, their movements blurred into dynamic light trails. Sunbeams dramatically pierce through the large arched windows, slowly shifting across the ornate architecture and marble floor. Cinematic, high contrast, incredible detail.



2. A vibrant, bustling night street food market in **Bangkok, Thailand.** Steam rises from numerous food stalls laden with colorful ingredients and sizzling woks. Cooks expertly prepare meals amidst flickering gas flames and hanging incandescent bulbs. A diverse crowd navigates the narrow, lively alleyway. Atmospheric, rich colors, shallow depth of field.



3. Dynamic low-angle shot from the **platform edge as a modern subway train** with sleek silver carriages and glowing interior lights rushes into a clean, well-lit station. Doors hiss open, a diverse group of passengers briskly exits and enters. The train's departure is equally swift. Smooth motion, reflections on the platform.



6. View from directly behind the drummer on a large concert stage: A massive, energetic crowd at a rock festival at night, thousands of hands in the air, faces illuminated by sweeping, multi-colored stage lights and occasional bursts of pyrotechnics. The band members are silhouetted in the foreground. Immersive, high energy, dynamic lighting, wide-angle shot.

Figure 2: Illustration of Key Frame Generation for Sequences with Multiple, Rapidly Changing Scenes, Conditioned on Semantically Distinct Text Prompts in 5s short video.

After fine-tuning with our proposed frame-level prompts, the model shows a much-improved ability to create complex videos. It can produce videos about **5 seconds** long (corresponding to 21 latents) that smoothly switch between as many as **6 different scenes**. This skill in handling **quick and varied scene changes**, as shown by the key frames presented in Fig. 2, is especially important. Although our training data, for similar 21-latent sections, usually presented only one or two scene changes—a relatively small number—the model’s strong performance when asked to create much more complex videos with multiple scenes shows that it has learned a general understanding of how to make logical shot and scene changes, rather than just copying the patterns common in the training data. In contrast, standard models often struggle when given prompts for such complex stories with multiple scenes, tending to create mixed-up or unclear videos. Our approach, however, shows a better ability to understand and accurately turn complex user requests into clear videos that feature multiple, well-defined scenes.

## B.2 (Short Video) Transition Generation (First-Last-Frame-to-Video) Ability

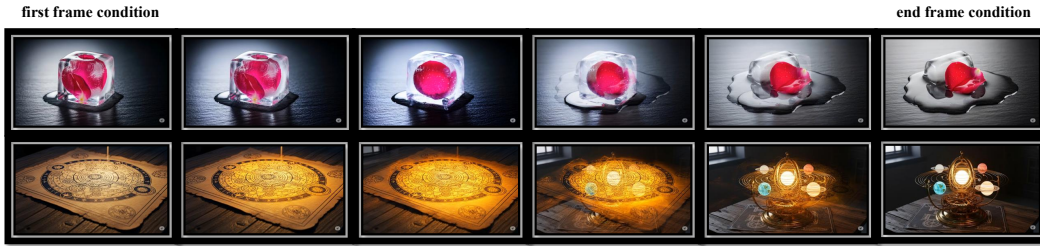


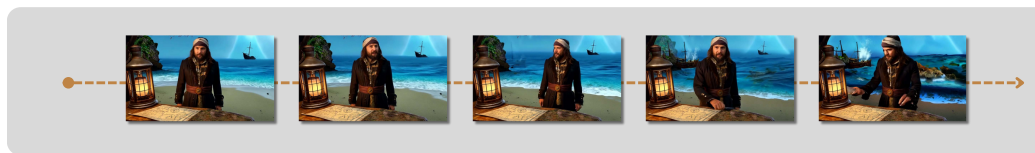
Figure 3: First-and-Last-Frame-to-Video Generation Enabled by Diffusion Forcing and Our Dynamic (Frame-Level) Prompts.

Diffusion forcing obviates the need for training an auxiliary encoder conditioning network or concatenating the first and last frame images within the model’s input. Instead, it only requires setting the timestep  $t = 0$  for these boundary frames, signifying them as clean or noise-free.

Furthermore, our proposed frame-level prompts significantly enhance the convenience and naturalness of first-and-last-frame video generation tasks. This approach is particularly advantageous as the initial and final frames often possess distinct semantic content, potentially originating from disparate sources such as individually selected web images. Frame-level prompts allow for independent textual descriptions for these two anchor frames. Subsequently, textual descriptions for the intermediate transitional frames can be generated or supplemented, a process that can be effectively assisted by Vision-Language Models (VLMs) or Large Language Models (LLMs), guided by high-level user intent.

## 50 C More Visualizations for Long Video with Complex Settings

### 1. Finetuned with Diffusion Forcing and Video-level Prompt, 5s video



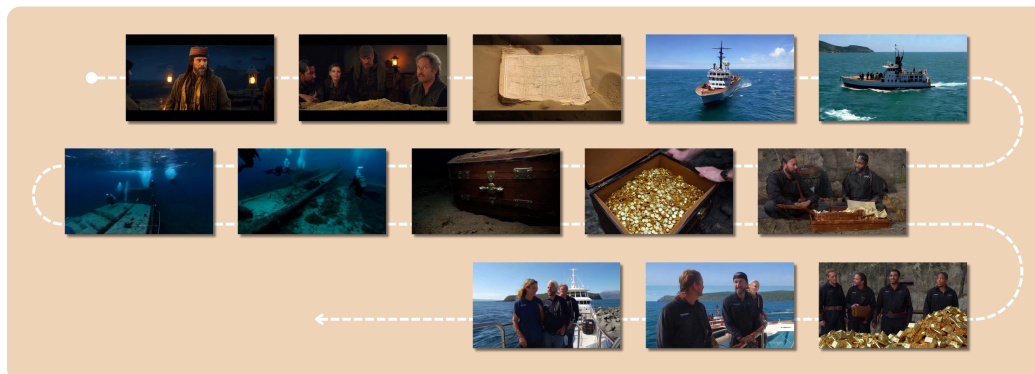
On a distant island, a **pirate captain** told his crew about a legendary treasure hidden in the wreck of a sunken **ship** deep in the sea. // On the first day, the crew found an **old map**. // On the second day, they sailed through a storm but discovered that the location on the map had become unclear. // On the third day, they dived into the sea and found the wreckage of the sunken ship. // On the fourth day, the crew found a locked chest inside the wreck. After much struggle, they finally opened the chest. // On the fifth day, the crew discovered that the treasure inside wasn't just gold and jewels; it also contained the pirate captain's final wish, leading them to decide to abandon the riches and search for their true life's purpose.

### 2. Finetuned with Diffusion Forcing and Video-level Prompt, 30s video



On a distant island, a **pirate captain** told **his crew** about a legendary treasure hidden in the wreck of a sunken ship deep in the sea. // On the first day, the crew found an **old map**. // On the second day, they sailed through a storm but discovered that the location on the map had become unclear. // On the third day, they dived into the sea and found the wreckage of the sunken ship. // On the fourth day, the crew found a **locked chest** inside the wreck. After much struggle, they finally opened the chest. // On the fifth day, the crew discovered that the treasure inside wasn't just gold and jewels; it also contained the pirate **captain's final wish**, leading them to decide to abandon the riches and search for their true life's purpose.

### 3. Finetuned with Diffusion Forcing and Frame-level Prompt, 30s video



On a distant island, a **pirate captain** told **his crew** about a legendary **treasure** hidden in the wreck of a sunken **ship** deep in the sea. // On the first day, the crew found an **old map**. // On the second day, they sailed through a storm but **discovered that the location** on the map had become unclear. // On the third day, they **dived into the sea** and **found the wreckage of the sunken ship**. // On the fourth day, the crew **found a locked chest inside the wreck**. After much struggle, they finally **opened the chest**. // On the fifth day, the crew discovered that the treasure inside wasn't just **gold and jewels**; it also contained the pirate **captain's final wish**, leading them to **decide to abandon the riches** and **search for their true life's purpose**.

**Figure 4: Comparative Video Generation Outputs.** This figure showcases videos generated by different methods: (a) a 5s video using a model finetuned with Diffusion Forcing and video-level prompt. (b) a 30s video using model finetuned with Diffusion Forcing (DF), FIFO, and a video-level prompt. (c) a 30s video using model finetuned with DF, PMWD (Parallel Multi-Window Denoising), and a dynamic prompts (frame-level). Note: Both 30s videos were generated with a fixed 21-latent window (consistent with training conditions). This limits the historical frame context for sliding window-based methods, potentially affecting ID consistency during multi-scene transitions or object occlusions. ID consistency could be enhanced using techniques like FramePack, incorporating additional compressed historical frames, or IP injection methods.

### 1. Finetuned with Diffusion Forcing and Video-level Prompt, 5s video



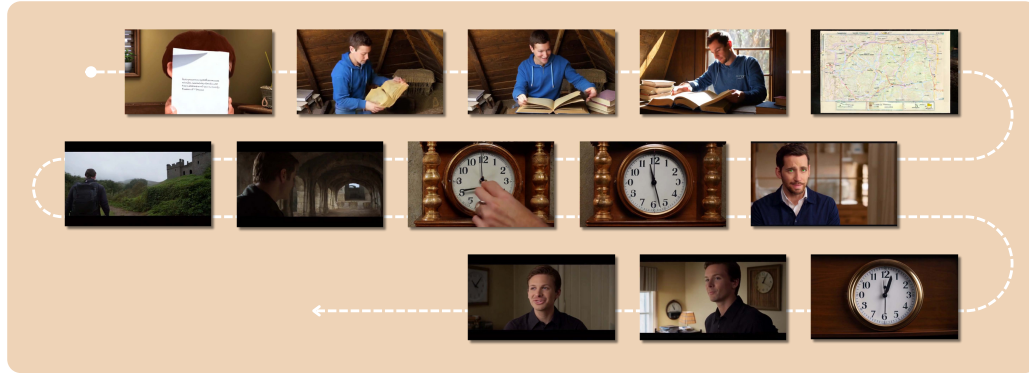
A young boy receives a letter from the future, saying that he will discover a secret that will change the world. // On the first day, he finds a dusty book filled with unknown symbols. // The next day, he finds a map in the book, guiding him to an abandoned castle. // On the third day, he enters the castle and finds a closed room containing an ancient clock. // On the fourth day, he follows the clock's instructions, activates a mechanism, and the second hand of the clock begins to turn backward. // On the fifth day, he realizes that the secret of the world does not lie in the outside world, but in how he perceives his past and future. This secret allows him to live in the present without being bound by time.

### 2. Finetuned with Diffusion Forcing and Video-level Prompt, 30s video



A young boy receives a letter from the future, saying that he will discover a secret that will change the world. // On the first day, he finds a dusty book filled with unknown symbols. // The next day, he finds a map in the book, guiding him to an abandoned castle. // On the third day, he enters the castle and finds a closed room containing an ancient clock. // On the fourth day, he follows the clock's instructions, activates a mechanism, and the second hand of the clock begins to turn backward. // On the fifth day, he realizes that the secret of the world does not lie in the outside world, but in how he perceives his past and future. This secret allows him to live in the present without being bound by time.

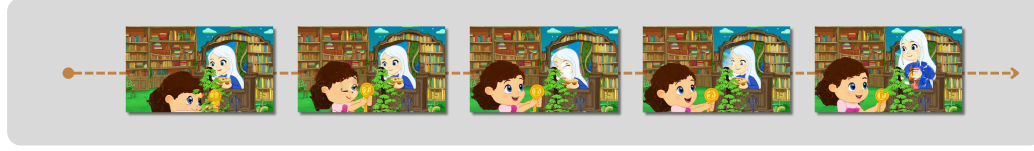
### 3. Finetuned with Diffusion Forcing and Frame-level Prompt, 30s video



A young boy receives a letter from the future, saying that he will discover a secret that will change the world. // On the first day, he finds a dusty book filled with unknown symbols. // The next day, he finds a map in the book, guiding him to an abandoned castle. // On the third day, he enters the castle and finds a closed room containing an ancient clock. // On the fourth day, he follows the clock's instructions, activates a mechanism, and the second hand of the clock begins to turn backward. // On the fifth day, he realizes that the secret of the world does not lie in the outside world, but in how he perceives his past and future. This secret allows him to live in the present without being bound by time.

Figure 5: Comparative Video Generation Outputs. This figure showcases videos generated by different methods: (a) a 5s video using a model finetuned with Diffusion Forcing and video-level prompt. (b) a 30s video using model finetuned with Diffusion Forcing (DF), FIFO, and a video-level prompt. (c) a 30s video using model finetuned with DF, PMWD (Parallel Multi-Window Denoising), and a dynamic prompts (frame-level). Note: Both 30s videos were generated with a fixed 21-latent window (consistent with training conditions). This limits the historical frame context for sliding window-based methods, potentially affecting ID consistency during multi-scene transitions or object occlusions. ID consistency could be enhanced using techniques like FramePack, incorporating additional compressed historical frames, or IP injection methods.

1. Finetuned with Diffusion Forcing and Video-level Prompt, 5s video



A little girl received a mysterious gift—the ancient time key. She used the key to open a sealed door and entered the time house. On the first day, she found a room filled with books, each with a different story. On the second day, she was taken to an ancient kitchen filled with strange smells, where a kind witch taught her how to make magic cookies that could change the past. On the third day, she found a mirror where her reflection looked completely different, even unrecognizable. On the fourth day, she planted a seed in the secret garden of the time house, and miraculously, it grew into a huge tree overnight. Finally, she discovered that she had always been living in a world unaffected by time, and the time house was her true home.

2. Finetuned with Diffusion Forcing and Video-level Prompt, 30s video



A little girl received a mysterious gift—the ancient time key. She used the key to open a sealed door and entered the time house. On the first day, she found a room filled with books, each with a different story. On the second day, she was taken to an ancient kitchen filled with strange smells, where a kind witch taught her how to make magic cookies that could change the past. On the third day, she found a mirror where her reflection looked completely different, even unrecognizable. On the fourth day, she planted a seed in the secret garden of the time house, and miraculously, it grew into a huge tree overnight. Finally, she discovered that she had always been living in a world unaffected by time, and the time house was her true home.

3. Finetuned with Diffusion Forcing and Frame-level Prompt, 30s video



A little girl received a mysterious gift—the ancient time key. She used the key to open a sealed door and entered the time house. // On the first day, she found a room filled with books, each with a different story. // On the second day, she was taken to an ancient kitchen filled with strange smells, where a kind witch taught her how to make magic cookies that could change the past. // On the third day, she found a mirror where her reflection looked completely different, even unrecognizable. // On the fourth day, she planted a seed in the secret garden of the time house, and miraculously, it grew into a huge tree overnight. Finally, she discovered that she had always been living in a world unaffected by time, and the time house was her true home.

Figure 6: Comparative Video Generation Outputs. This figure showcases videos generated by different methods: (a) a 5s video using a model finetuned with Diffusion Forcing and video-level prompt. (b) a 30s video using model finetuned with Diffusion Forcing (DF), FIFO, and a video-level prompt. (c) a 30s video using model finetuned with DF, PMWD (Parallel Multi-Window Denoising), and a dynamic prompts (frame-level). Note: Both 30s videos were generated with a fixed 21-latent window (consistent with training conditions). This limits the historical frame context for sliding window-based methods, potentially affecting ID consistency during multi-scene transitions or object occlusions. ID consistency could be enhanced using techniques like FramePack, incorporating additional compressed historical frames, or IP injection methods.

## 51 D Multiple Inference Mode Due to Precise Alignment and Flexibility of 52 Frame-Level Prompts

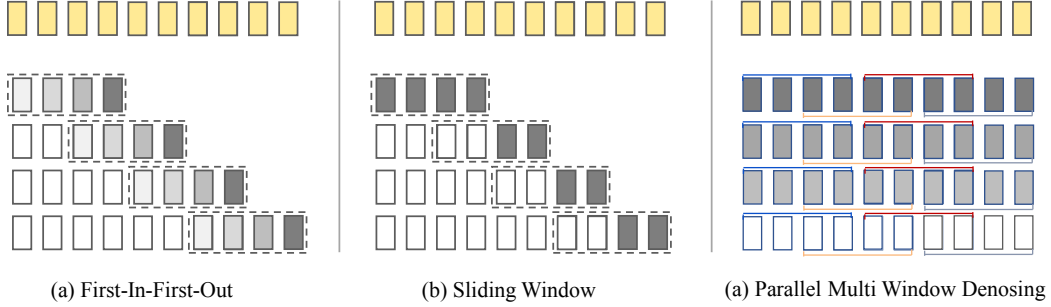


Figure 7: Three inference modes. Multiple yellow tokens represent different frame-level prompts. Less gray value represents lower noise.

53 Dynamic prompts, particularly frame-level prompts, offer significant **convenience** during inference  
54 because each latent unit directly maps to a corresponding prompt. This characteristic makes them  
55 highly suitable for methods such as FIFO, which involves single-latent window slides per denoising  
56 step, and for chunk-level auto-regressive approaches that support variable chunk sizes. Our proposed  
57 Parallel Multi-Window Denoising (PMWD) method also leverages this: for generating very long  
58 sequences, each latent within every parallel denoising window aligns easily with its specific prompt,  
59 facilitating effective information exchange in overlapping regions.

60 While these three approaches (FIFO, chunk-level auto-regression, and PMWD) are based on fixed-  
61 length sliding windows, this can present **limitations**. For instance, maintaining **ID consistency** can be  
62 challenging when the historical frames within the window are insufficient, or when objects disappear  
63 or are occluded in complex scenes. However, *the primary focus of this paper is the flexibility and*  
64 *potential offered by frame-level prompts across the entire pipeline—encompassing dataset collection*  
65 *and construction, through to training and inference stages.* Should enhanced ID preservation be a  
66 priority, our frame-level prompt system can be readily augmented with techniques that expand the  
67 historical context, such as those employed by FramePack or KV caching methods.

---

**Algorithm 1** Sliding Window Inference Mode for Long Video Generation for Model Trained With Diffusion Forcing and Frame-level Prompts.

---

**Require:** Denoiser  $D_\theta$ , Denoise Steps Per Slide  $T$ , Window Size  $F_{\text{window}}$ , Sliding Size  $F_{\text{slide}}$ , History Clean Latents  $z_0^{\text{history}}$ , History Condition Timestep  $T_{\text{history}}$ , frame-level prompts  $C_{\text{text}}$

```

1:  $F_{\text{history}} \leftarrow F_{\text{window}} - F_{\text{slide}}$   $\triangleright$  To maintain a fixed window size, we only see part of history frames
2:  $z_0^{\text{history}} \leftarrow z_0^{\text{history}}[-F_{\text{history}}:]$   $\triangleright$  see last  $F_{\text{history}}$  history latents only
3:  $c_{\text{text}} \leftarrow$  get last  $F_{\text{history}}$  prompts and  $F_{\text{slide}}$  new prompts from  $C_{\text{text}}$ .
4:  $z_T \sim \mathcal{N}(0, \mathbf{I}) \in \mathbb{R}^{F_{\text{slide}} \times H \times W \times D}$ 
5:  $z_t, t \leftarrow z_T, T$ 
6: repeat
7:    $z_t^{\text{history}} \leftarrow$  add noise( $z_t^{\text{history}}, T_{\text{history}}$ )
8:    $z_{t-1} \leftarrow D_\theta(z_t^{\text{history}}, z_t, t, c_{\text{text}})$ 
9:    $t \leftarrow t - 1$ 
10: until  $t = 0$   $\triangleright$  new frames are denoised to be clean
return  $z_0$ 

```

---

---

**Algorithm 2** First-In-First-Out Inference Mode for Long Video Generation for Model Trained With Diffusion Forcing and Frame-level Prompts.

---

**Require:** Denoiser  $D_\theta$ , Window Size  $F$ , frame-level prompts  $C_{\text{text}}$ , current latent  $z_{t_{\tau_1}, t_{\tau_2}, \dots, t_{\tau_F}}$

- 1:  $z_{t_{\tau_F}} \sim \mathcal{N}(0, \mathbf{I}) \in \mathbb{R}^{1 \times H \times W \times D}$
- 2:  $z_{t_{\tau_0}}, z_{t_{\tau_1}, t_{\tau_2}, \dots, t_{\tau_F}} \leftarrow \text{Slide one frame}(z_{t_{\tau_0}, t_{\tau_1}, \dots, t_{\tau_{F-1}}}, z_{t_{\tau_F}})$
- 3:  $c_{\text{text}} \leftarrow \text{outqueue the clean frame's prompt and inqueue a prompt for new noise frame from } C_{\text{text}}$
- 4:  $z_{t_{\tau_0}, t_{\tau_1}, \dots, t_{\tau_{F-1}}}, z_{t_{\tau_F}} \leftarrow D_\theta(z_{t_{\tau_1}, t_{\tau_2}, \dots, t_{\tau_F}}, t, c_{\text{text}})$

**return**  $z_{t_{\tau_0}}$  ▷ output only one clean frame in each slide of FIFO

---



---

**Algorithm 3** Parallel Multi Windows Denosing Inference Mode for Long Video Generation for Model Trained With Diffusion Forcing and Frame-level Prompts.

---

**Require:** Denoiser  $D_\theta$ , Denoise Steps, Window Size  $F_{\text{window}}$ , Long Video Length  $F$ , Parallel Numbers  $K$ , frame-level prompts  $C_{\text{text}}$

- 1:  $z_T \sim \mathcal{N}(0, \mathbf{I}) \in \mathbb{R}^{F \times H \times W \times D}$
- 2:  $z_t, t \leftarrow z_T, T$
- 3:  $F_{\text{overlapped}} \leftarrow (K \times F_{\text{window}} - F) / (K - 1)$
- 4:  $F_{\text{slide}} \leftarrow F - F_{\text{overlapped}}$
- 5: **repeat**
- 6:    $z_{t-1} \leftarrow 0 \in \mathbb{R}^{F \times H \times W \times D}$
- 7:    $\text{count} \leftarrow 0 \in \mathbb{R}^{F \times H \times W \times D}$
- 8:   **for all**  $k \in K$  **do** ▷  $K$  window can denoise in parallel
- 9:      $s \leftarrow 0 + F_{\text{slide}} * k$  ▷ start position
- 10:     $e \leftarrow s + F_{\text{window}}$  ▷ end position
- 11:     $z_{t-1}[s : e] \leftarrow z_{t-1}[s : e] + D_\theta(z_t[s : e], t, c_{\text{text}}[s : e], c_{\text{others}})$  ▷  $c_{\text{others}}$  is optional.
- 12:     $\text{count}[s : e] \leftarrow \text{count}[s : e] + 1$
- 13:   **end for**
- 14:    $z_{t-1} \leftarrow z_{t-1} / \text{count}$
- 15:    $t \leftarrow t - 1$
- 16: **until**  $t = 0$  ▷ all frames in the entire long video are denoised to be clean simultaneously

**return**  $z_t$

---

68 The  $c_{\text{others}}$  parameter in Algorithm 3 represents optional auxiliary conditioning information. This can  
69 include frames beyond the current processing window (e.g., distant historical frames) or injected data  
70 specifically designed to maintain identity (e.g., ID embeddings). The primary purpose of  $c_{\text{others}}$  is to  
71 potentially enhance ID preservation and temporal consistency in long video generation. In this paper,  
72 unless otherwise specified,  $c_{\text{others}}$  is not utilized, and the model's receptive field during inference is  
73 confined to the window size employed during its training.

## 74 E VLM and LLM for auto captioning and converting

### 75 E.1 Frame-level Captioner

76 We use Gemini pro 1.5 VLM for captioning.

#### System Prompt for Frame-Level Captioning

You are given a sequence of images representing a video. The video consists of exactly NUM\_FRAMES frames, and you must generate one description per frame. Your task is to produce a JSON object with exactly NUM\_FRAMES keys. Each key must be a sequential integer from 1 to NUM\_FRAMES and each key's value is a detailed, standalone description of that specific frame. Follow these guidelines:

##### 1. Output Format:

- The output must be a valid JSON object.
- The JSON object must have exactly NUM\_FRAMES keys, where each key is an integer representing the frame index (e.g., 1, 2, 3, ..., NUM\_FRAMES).

77

- Do not include any extra keys or omit any frame indices.
2. Description Requirements:
    - Clearly describe the main characters, objects, and environmental features visible in the current frame.
    - Explain the primary actions and any visual changes involving characters or objects.
    - Provide additional details about the scene's environment (e.g., lighting, fog, tree shapes).
    - Emphasize any newly introduced visual elements or significant changes (e.g., scene transitions, lighting effects).
    - If there are no significant changes, still provide a complete, self-contained description of the scene without referencing previous frames.
    - Include shot size and camera angle details as follows:
      - Shot Size:
        - \* Long Shot: Wide view capturing the overall environment and spatial relationships.
        - \* Medium Shot: Focus on interactions (e.g., conversations, walking).
        - \* Close-Up: Emphasize details such as facial expressions, actions, or emotions.
        - \* Extreme Close-Up: Highlight minute details (e.g., eyes, hands).
      - Camera Angle:
        - \* Eye-Level: A neutral, natural perspective.
        - \* Low-Angle: Conveys a sense of power or dominance.
        - \* High-Angle: Suggests vulnerability or provides an overview.
        - \* Point-of-View: Represents the perspective of a specific character.
      - Special Angles (if applicable):
        - \* Bird's Eye View: An overhead perspective.
        - \* Handheld Shaking: Imparts a sense of realism.
        - \* Distant Tracking: Offers an observational feel.
  3. Output Example (Assuming there are 3 frames):
 

```
{
  "1": "An animated female character wearing a purple robe stands in the center of a forest, releasing bright purple energy from her palm that illuminates the surrounding trees. Dense fog softens the outlines of the trees. [Medium Shot, Eye-Level]",
  "2": "The intensity of the purple energy increases as the character's expression becomes more focused. Shadows shift among the trees, and a spider crawls into view. [Close-Up, Low-Angle]",
  "3": "The camera pulls back to reveal the entire forest bathed in purple light. Tall, eerie trees and faint silhouettes of giant spiders moving through the fog create an otherworldly scene. [Long Shot, Bird's Eye View]"
}
```
  4. Additional Strict Instructions:
    - Every image/frame must have one and only one corresponding description.
    - The JSON output must contain exactly NUM\_FRAMES keys, matching the number of video frames.
    - Do not combine multiple frame descriptions into one key or leave any frame without a description.
    - Do not add any extra text, explanations, or commentary outside of the JSON object.
  5. Important Notes:
    - Descriptions should be based solely on visual information; ignore any audio elements.
    - Each frame's description must be fully self-contained and not refer to previous frames.
    - Ensure your output adheres strictly to JSON format.

## 79 E.2 Frame-level Prompt Convertor

80 We use Gemini pro 2.5 for converting video-level prompt to frame-level prompt.

### System Prompt for Converting Video-Level Caption to Frame-Level Caption

As a story-telling prompt engineering specialist in temporal decomposition, transform the given global prompt into NUM\_PROMPTS frame-level prompts. You will be given the global prompt (English) of a 21s video. New generated frame-level prompts should provide clearer and explicit guidance for video frame generation by precisely decomposing temporal dynamics. Please output a JSON object with exactly NUM\_PROMPTS keys. Each key represents the temporal indices and must be a sequential integer from 1 to NUM\_PROMPTS. Each key's value is a detailed, standalone description. Follow these guidelines:

#### 1. Objective:

- Our primary goal is to enhance the clarity of video generation by providing more temporally precise guidance.
- The frame-level dynamic prompts should enable video generation models to better understand: Dynamic Spatial Relationship, Human Action, Motion Order Understanding, Dynamic Attribute, Complex Plot, Complex Scene

#### 2. Key Requirements:

- The original events described in the original prompt should be distributed as evenly as possible across the generated dynamic prompts.
- Decompose complex motions, actions, and spatial relationships into precise, self-contained, simple, frame-level descriptions.
- Include shot size and camera angle details as follows:
  - Shot Size:
    - \* Long Shot: Wide view capturing the overall environment and spatial relationships.
    - \* Medium Shot: Focus on interactions (e.g., conversations, walking).
    - \* Close-Up: Emphasize details such as facial expressions, actions, or emotions.
    - \* Extreme Close-Up: Highlight minute details (e.g., eyes, hands).
  - Camera Angle:
    - \* Eye-Level: A neutral, natural perspective.
    - \* Low-Angle: Conveys a sense of power or dominance.
    - \* High-Angle: Suggests vulnerability or provides an overview.
    - \* Point-of-View: Represents the perspective of a specific character.
  - Special Angles (if applicable):
    - \* Bird's Eye View: An overhead perspective.
    - \* Handheld Shaking: Imparts a sense of realism.
    - \* Distant Tracking: Offers an observational feel.
- Avoid unnecessary scene and camera changes.

#### 3. Input Format and Example:

"A dog is on the left of a table, then the dog runs to the front of the table."

#### 4. Assume NUM\_PROMPTS=7, then Output Frame-Level Dynamic Prompts Format and Examples (Key represents frame indices and Value represents frame-level detailed prompts):

```
{
  "1": "A dog is on the left of a table. [Long Shot, Eye-Level]",
  "2": "The same dog is on the left of the same table. [Long Shot, Eye-Level]",
  "3": "The same dog who is on the left of a table takes a small step forward the front of the same table. [Long Shot, Eye-Level]",
```

81

```

"4": "The same dog is now halfway between the left side and
the front of the same table. [Long Shot, Eye-Level]",
"5": "The same dog is now slightly in front of the same
table. [Long Shot, Eye-Level]",
"6": "The same dog is now slightly in front of the same
table. [Long Shot, Eye-Level]",
"7": "The same dog stands at the front of the same table.
[Long Shot, Eye-Level]"
}

```

##### 5. Strict Instructions:

- NUM\_PROMPTS=21
- Distribute the global prompt as evenly as possible across NUM\_PROMPTS, minimizing unnecessary camera angle and scene changes. Ensure that variations between adjacent prompts are as small as possible.
- The time difference between two adjacent prompts is only 1 second; therefore, ensure that dynamic descriptions depict slow changes that can realistically occur within 1 second.
- Maintain similar sentence structures as much as possible.
- For objects, people, characters, scenes, etc., ensure descriptions are detailed enough to uniquely identify them (ID). Especially when referring to the same object, person, character, or scene, explicitly state "the same" or clearly indicate which elements share the same ID as in previous frames.
- Do not add any extra text, explanations, or commentary outside of the JSON object.
- The output must be a valid JSON object with NUM\_PROMPTS keys.  
Please start with “{” and end with “}”. Only one json should be output.

82

## VBench 2.0 Evaluation Prompts Examples

### • Complex Plots, Example 16-th:

#### – video-level prompt:

"A little girl received a mysterious gift—the ancient time key. She used the key to open a sealed door and entered the time house. On the first day, she found a room filled with books, each with a different story. On the second day, she was taken to an ancient kitchen filled with strange smells, where a kind witch taught her how to make magic cookies that could change the past. On the third day, she found a mirror where her reflection looked completely different, even unrecognizable. On the fourth day, she planted a seed in the secret garden of the time house, and miraculously, it grew into a huge tree overnight. Finally, she discovered that she had always been living in a world unaffected by time, and the time house was her true home.

#### – converted frame-level prompt:

```

{
"1": "A little girl named Lily, with curly brown hair and
bright blue eyes, is shown in a warmly lit room. She
looks curious and observant. [Medium Shot, Eye-Level]",
"2": "The same Lily receives a mysterious gift, an ancient
golden time key, presented by an unseen hand. Her
bright blue eyes widen with intrigue as she takes the
key. [Close-Up on the golden key being passed to her,
then Medium Shot of Lily holding it, Eye-Level]",
"3": "Lily, now holding the same ancient golden time key
firmly, cautiously approaches a large, ornate, sealed
wooden door within the same house. [Medium Shot,
Eye-Level, tracking slightly with Lily as she walks]",
"4": "Lily carefully inserts the ancient golden time key
into the keyhole of the sealed wooden door. She turns

```

83

the key, and the door emits a soft click and begins to creak open, revealing a mysterious, inviting glow from within. [Close-Up on the key turning in the lock, then Medium Shot as the door opens, Eye-Level]",

"5": "Lily takes a tentative step through the now open doorway and enters a magical time house. Her face, illuminated by the glow, shows an expression of pure wonder and awe as she gazes at the mystical interior. [Point-of-View shot from Lily's perspective as she enters, then transitioning to a Medium Shot of Lily's amazed face, Eye-Level]",

"6": "On her first day in the magical time house, Lily finds herself standing at the entrance of a vast library. Bookshelves stretch up to an impossibly high, vaulted ceiling, packed with countless ancient-looking books. [Long Shot, High-Angle, emphasizing the immense scale of the library and Lily's smallness within it]",

"7": "The same Lily walks slowly and reverently between the towering bookshelves in the vast library, her head tilted upwards, admiring the sheer volume of books. The air is still and filled with the faint, sweet scent of old paper and magic. [Medium Shot, Eye-Level, tracking Lily smoothly as she explores deeper into the library]",

"8": "Lily pauses before a specific bookshelf and reaches out her hand, gently pulling a uniquely bound, ornate book from its place. As her fingers touch its cover, a faint, magical shimmer emanates from the book itself. [Close-Up on Lily's hand selecting the shimmering book, then Medium Shot of her holding it with both hands, Eye-Level]",

"9": "Lily sits on a comfortable, old-fashioned chair in the library and opens the shimmering book. As she carefully begins to flip through its delicate, ancient pages, soft, ethereal whispers seem to emanate from the book, and her curious expression deepens into fascination. [Close-Up focusing on Lily's expressive face and the open book in her lap, Eye-Level]",

"10": "On her second day, Lily is gently transported from the library to an ancient, warmly lit kitchen within the time house. The air here is filled with strange, enchanting scents of exotic spices and sweet baking. She looks around, her nose twitching with curiosity. [Medium Shot, Eye-Level, with a slow pan to show the charming details of the kitchen: copper pots, bubbling cauldrons]",

"11": "A kind-looking witch, with long, flowing silver hair, a warm, welcoming smile, and twinkling eyes, greets Lily in the ancient kitchen. The witch wears a velvet robe embroidered with celestial patterns. [Medium Shot of the witch gesturing a welcome, then a two-shot including Lily who looks up at her, Eye-Level]",

"12": "The same kind witch gestures towards a large wooden table laden with unusual and colorful ingredients: sparkling powders, glowing fruits, and oddly shaped herbs. She explains to Lily that they will make magic cookies that can change the past. Lily listens intently, captivated. [Close-Up on the magical ingredients shimmering on the table, then Medium Shot of the witch and Lily interacting at the table, Eye-Level]",

"13": "Guided by the witch's gentle instructions, Lily carefully begins to mix the ingredients for the magic

cookies in a large earthenware bowl. Her brow is furrowed in concentration as she stirs the mixture. [Close-Up on Lily's hands as she diligently mixes the dough, the ingredients swirling together, Eye-Level]",

"14": "Lily takes a small vial of shimmering stardust offered by the witch and carefully sprinkles it into the cookie dough. The dough instantly glows with a soft, golden light upon contact with the stardust. Lily gasps softly in delight. [Extreme Close-Up of the stardust falling like tiny stars into the bowl, then Close-Up on Lily's focused, joyful face and the glowing dough, Eye-Level]",

"15": "On her third day, while exploring another wing of the time house, Lily discovers a large, ornate magical mirror standing in a quiet, dimly lit, circular room. The mirror's frame is carved with intricate temporal symbols. [Medium Shot, Eye-Level, as Lily cautiously approaches the mysterious mirror, her reflection appearing normal initially]",

"16": "Lily stands directly before the magical mirror, looking intently at her own reflection. Slowly, her reflection begins to shimmer and distort, its features subtly changing, no longer perfectly matching her current appearance. Her eyes widen in surprise. [Close-Up on Lily's face looking into the mirror, her expression shifting from curiosity to astonishment, Eye-Level]",

"17": "The reflection in the magical mirror now fully transforms, showing a version of Lily from another time: this reflection has a mischievous grin, eyes that sparkle with a different light, and a surprising twirl of vibrant golden hair instead of her brown curls. It is almost unrecognizable. [Close-Up shot focused entirely on the mirror's surface, showcasing the details of the transformed, golden-haired reflection, Eye-Level]",

"18": "On her fourth day, Lily ventures into the enchanting secret garden of the time house. She holds a single, tiny, unassuming seed in the palm of her hand. The garden is lush, filled with fantastical, glowing plants and flowers of impossible colors. [Medium Shot, Eye-Level, following Lily as she steps into the vibrant, sun-dappled secret garden]",

"19": "Lily kneels down in a sunlit patch of the secret garden, amidst the magical flora. She uses her fingers to make a small hole in the rich, dark soil and carefully plants the tiny seed. [Close-Up on Lily's hands gently placing the seed into the earth and covering it, Low-Angle shot from ground level, emphasizing her connection to the earth]",

"20": "Lily stands back and watches in utter amazement as, in a rapid, magical sequence, the newly planted seed sprouts, a green shoot unfurls, and it grows astonishingly fast into a massive, majestic tree. Its leaves begin to shimmer with a soft, ethereal, multi-colored glow. [Medium Shot of Lily looking up in awe, transitioning to a visually stunning time-lapse effect of the tree's growth, concluding with a Low-Angle looking up at the fully grown, glowing tree against the sky]",

"21": "Lily stands serenely before the magnificent, magically glowing tree in the heart of the secret garden. A gentle, knowing smile graces her face as she

```

        realizes that this wondrous time house, a place
        untouched by ordinary time, is her true, magical home.
        The scene is imbued with vibrant colors and profound
        mystical elements. [Medium Long Shot, Eye-Level,
        capturing Lily and the grandeur of the glowing tree,
        conveying a sense of peace, belonging, and enchantment]"
    }

```

- **Complex Landscape, Example 19-th:**

- **video-level prompt:**

"The camera begins with a vast lavender field stretching into the distance, the purple flowers swaying like waves in the breeze. The sun rises behind distant hills, casting a warm glow over the field, outlining each lavender bloom with golden edges. As the camera moves forward, a few bees can be seen buzzing through the flowers, collecting nectar, their hum echoing in the quiet morning air. The camera continues, capturing a few butterflies dancing gracefully through the air, their wings glistening in the sunlight. The camera shifts toward the distant hills, where the outlines of the mountains become clearer in the morning light, with a thin mist hovering over the peaks, giving the scene a dreamlike quality. Finally, the camera pulls back, showing the entire lavender field blending with the hills, where the purple flowers, golden sunlight, and distant mountains form a tranquil and beautiful landscape."

- **frame-level prompt:**

```

"1": "A vast lavender field stretches into the distance,
      where purple flowers sway like waves in the gentle
      breeze. [Long Shot, Eye-Level]",
"2": "The same vast lavender field continues to stretch
      into the distance, the same purple flowers swaying
      gently like waves in the same gentle breeze. [Long
      Shot, Eye-Level]",
"3": "The sun begins to rise behind distant hills, casting
      an initial warm, golden glow over the same vast
      lavender field. The same purple flowers are swaying.
      [Long Shot, Eye-Level]",
"4": "The sun rises further behind the same distant hills,
      intensifying the warm, golden glow over the same vast
      lavender field, outlining each lavender bloom with
      shimmering edges. The same purple flowers continue to
      sway. [Long Shot, Eye-Level]",
"5": "The camera begins to move forward slowly over the
      same vast lavender field, where the same purple flowers
      sway. The sun, now higher behind the same distant
      hills, casts a strong golden glow, outlining the
      blooms. [Long Shot, Eye-Level]",
"6": "The camera continues its slow forward movement over
      the same vast lavender field. The same purple flowers
      sway, and the golden glow from the sun behind the same
      distant hills strongly outlines the blooms. [Long Shot,
      Eye-Level]",
"7": "As the camera moves forward slowly over the same
      vast lavender field, a few bees become visible,
      beginning to buzz among the same swaying purple
      flowers. The golden sunlight from behind the same
      distant hills continues to illuminate the scene.
      [Medium Shot, Eye-Level]",
"8": "The camera, still moving forward slowly, focuses on
      the same few bees buzzing actively through the same
      swaying purple flowers, starting to collect nectar.
      Their hum begins to echo in the quiet morning air. The
      golden sunlight from the same distant hills is
      prominent. [Medium Shot, Eye-Level]",

```

"9": "While the camera glides forward over the same lavender field, a few butterflies begin to dance gracefully through the air among the same swaying flowers and the same buzzing bees. Their wings start to glisten in the same golden sunlight. [Medium Shot, Eye-Level]",

"10": "The camera continues to glide forward, capturing the same bees buzzing and collecting nectar from the same swaying flowers, and the same butterflies dancing gracefully, their wings glistening brightly in the same golden sunlight. [Medium Shot, Eye-Level]",

"11": "The camera glides smoothly forward, maintaining focus on the serene movement of the same bees as they collect nectar and the same butterflies as they dance among the same swaying lavender flowers, all under the golden sunlight. Their hum echoes. [Medium Shot, Eye-Level]",

"12": "The camera continues its gentle glide forward, showcasing the interaction of the same bees and the same butterflies with the same swaying lavender flowers. The golden sunlight illuminates their movements, and their hum echoes. [Medium Shot, Eye-Level]",

"13": "The camera begins to shift its direction, panning slowly from the flowers and insects towards the same distant hills. The golden sunlight still bathes the lavender field. The hum of bees is faintly audible. [Long Shot, Eye-Level]",

"14": "The camera continues to shift towards the same distant hills, where the outlines of the mountains start to become clearer in the morning light. A thin mist is visible, hovering over the peaks. The lavender field is in the foreground, with the golden sunlight. [Long Shot, Eye-Level]",

"15": "The camera's view is now focused on the same distant hills, the outlines of the mountains clearly defined in the morning light. The same thin mist hovers over the peaks, giving the scene a dreamlike quality. The lavender field stretches below, bathed in golden sunlight. [Long Shot, Eye-Level]",

"16": "The camera begins to pull back gradually from its focus on the distant hills and mist, starting to show more of the vast lavender field blending with the hills. The golden sunlight is still present across the scene. [Long Shot, High-Angle]",

"17": "The camera continues to pull back gradually, revealing more of the entire lavender field seamlessly blending with the same distant hills and mist-covered mountains. The golden sunlight illuminates the expansive scene. [Long Shot, High-Angle]",

"18": "The camera pulls back further, offering a wider view of the entire lavender field blending seamlessly with the same distant hills. The purple flowers, golden sunlight, and distant mountains are all visible in this wider composition. [Very Long Shot, High-Angle]",

"19": "The camera, having pulled back, now shows the entire vast lavender field blending seamlessly with the same distant hills. The same purple flowers, the same golden sunlight, and the same distant mountains form a tranquil and beautiful landscape. [Very Long Shot, High-Angle]",

"20": "The camera holds on the panoramic view of the entire lavender field blending with the distant hills.

The tranquil beauty of the purple flowers, golden sunlight, and mist-topped mountains evokes a sense of peace. [Very Long Shot, High-Angle]",  
"21": "The final shot is a serene, wide panorama of the vast lavender field, the golden sunlight, and the distant mountains, all blending into a tranquil and beautiful landscape, evoking a profound sense of peace and tranquility. [Very Long Shot, High-Angle]"

88

## 89 **F Broader Impacts**

90 Our work uses proprietary dataset. The dataset construction process explicitly includes deduplication,  
91 dynamic content filtering, and uniform sampling to ensure diversity and semantic richness. Annotation  
92 is fully performed using large language models (e.g., Gemini Pro 1.5), guided by structured prompts  
93 to ensure consistency and neutrality. To reduce potential misuse risks, the annotation process and  
94 model usage are designed with controlled access and reproducibility guidelines. While misuse is  
95 possible in principle, we emphasize that proper access control and responsible use are necessary to  
96 minimize unintended negative impacts.