

# Frame-Level Captions for Long Video Generation with Complex Multi Scenes

Guangcong Zheng, Jianlong Yuan<sup>†</sup>, Bo Wang, Haoyang Huang, Guoqing Ma, Nan Duan

StepFun, China

<sup>†</sup>Project Leader

## Abstract

Generating long videos that can show complex stories, like movie scenes from scripts, has great promise and offers much more than short clips. However, current methods that use autoregression with diffusion models often struggle because their step-by-step process naturally leads to a serious error accumulation (drift). Also, many existing ways to make long videos focus on single, continuous scenes, making them less useful for stories with many events and changes. This paper introduces a new approach to solve these problems. First, we propose a novel way to annotate datasets at the **frame-level**, providing detailed text guidance needed for making complex, multi-scene long videos. This detailed guidance works with a Frame-Level Attention Mechanism to make sure text and video match precisely. A key feature is that each part (frame) within these windows can be guided by its own distinct text prompt. Our training uses **Diffusion Forcing** to provide the model with the ability to handle time flexibly. We tested our approach on difficult VBench 2.0 benchmarks ("Complex Plots" and "Complex Landscapes") based on the WanX2.1-T2V-1.3B model. The results show our method is better at following instructions in complex, changing scenes and creates high-quality long videos. We plan to share our dataset annotation methods and trained models with the research community. **Project page:** [Frame-Level Captions]

## 1 Introduction

The ability to create long video sequences from text instructions opens exciting doors for rich, evolving stories, such as turning scripts into videos, producing short films, or showing complex processes. Unlike short clips, long videos provide the needed duration for multiple connected scenes, detailed character interactions, and consistent plotlines that follow complex user requests [17, 42, 32, 23, 15, 20, 22]. However, creating high-quality, consistent, and accurate long videos from text is still a major challenge for current generative models.

A primary difficulty lies in the common autoregressive (step-by-step) methods used with diffusion models to make longer videos. Their sequential way of working naturally leads to errors accumulation over time. This shows up as lower visual quality, the video drifting away from the original text's meaning, and a loss of consistency as the video gets longer, seriously weakening the quality of extended generations [14, 18, 37, 36]. Furthermore, much current research on long videos deals with single, continuous scenes or slowly changing environments. This limited focus reduces the usefulness of long video generation for dynamic stories with many events, which is a key goal for creative applications. Standard methods using single, general (global) video descriptions struggle with small, quick changes, leading to timing issues [26, 18, 19]. While some approaches try to make multi-shot videos by first detecting scene cuts with tools like PySceneDetect and then processing these shots, as in Presto [30] and Long Context Tuning (LCT) [12], these methods can be complicated,

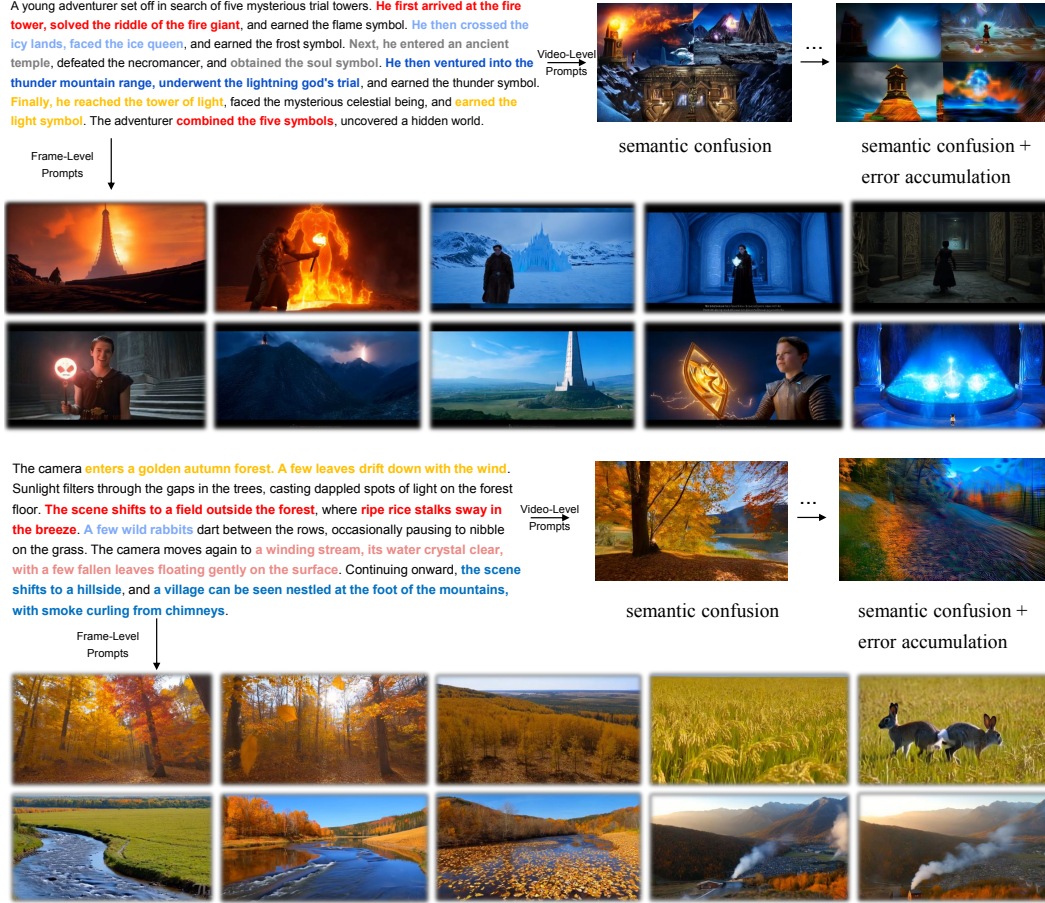


Figure 1: **Illustration of Semantic Confusion and Error Accumulation in Video Generation.** This figure contrasts 30s videos generated by conventional video-level prompt (right) and our proposed frame-level prompt (ten video key frames below), demonstrating improved generation quality of our method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

risk losing information, and depend heavily on good shot detection and captioning. They often still describe video at a "shot-text" level, which doesn't fully capture smooth, continuous changes. Error accumulation or drifting is recognized as 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 solve these basic problems, our work offers a new way of thinking, focused on highly detailed, frame-level text guidance and a novel non-sequential method for creating the video. Our first main contribution is an innovative **frame-level dataset annotation methodology**. We move beyond general or shot-level captions to provide very detailed text descriptions for each conceptual part (or latent segment) of a video. This rich information about meaning is essential for guiding models to understand and create the complex, changing details needed for stories with many scenes and detailed 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**. This directly addresses the limits of less detailed global or shot-level captions. This detailed annotation is designed to work closely with our **Frame-Level Attention Mechanism**, which clearly links each video segment’s visual features to its specific text description, improving content accuracy and consistency over time (Section 4.2).

To properly use such detailed and dynamic text prompts, models need to be trained to handle time flexibly. We achieve this using **Diffusion Forcing** (Section 4.3), a training strategy that shows the model video segments being denoised at different rates. This prepares it to manage varied timing patterns and allows for strong, adaptable inference. Building on these training improvements, we introduce our second major innovation: Parallel Multi-Window Denoising (PMWD), a new inference method designed to create very long videos (Section 4.4). PMWD divides the target long video into multiple overlapping sections (windows), usually matching the model’s training window size. Importantly, unlike step-by-step methods, all these windows are processed *at the same time (in parallel)* during each step of the diffusion denoising process. The data in the overlapping areas between windows is then averaged. This averaging not only ensures smooth connections but also allows information to flow in both directions, meaning later parts of the video can help refine earlier ones. A special feature of PMWD is that each conceptual frame, even within these simultaneously processed windows, can be guided by its own distinct, frame-level prompt. We test our approach thoroughly using highly challenging benchmarks, specifically the "Complex Plots" and "Complex Landscapes" prompt categories from VBench 2.0 [45]. We use the state-of-the-art open-source WanX2.1-T2V-1.3B model [32] as our base. Our experiments show that our combined frame-level approach is much better at following prompts when creating very long videos with multiple scenes, different characters, and complex actions.

In summary, our main contributions are:

- **Scalable Frame-Level Dataset Methodology:** We introduce an efficient and scalable approach for constructing datasets with dense, frame-by-frame textual annotations. This enables highly granular video-text alignment crucial for generating complex, multi-scene narratives without relying on traditional shot detection.
- **Frame-Level Attention for Precise Guidance:** We propose a novel attention mechanism that directly couples each video segment’s visual features with its unique frame-level prompt. This significantly enhances semantic fidelity, content accuracy, and temporal consistency in generated videos.
- **Parallel Multi-Window Denoising for Coherent Long Video Generation:** We develop PMWD, a inference strategy that processes a long video as multiple overlapping windows, denoised simultaneously in parallel. Guided by distinct frame-level prompts and leveraging overlap averaging for bidirectional context, PMWD effectively avoids the error accumulation common in sequential methods. This is enabled by training strategies like Diffusion Forcing that provide temporal flexibility.
- **State-of-the-Art Performance on Complex Videos:** Through comprehensive evaluations on challenging VBench 2.0 benchmarks ("Complex Plots" and "Complex Landscapes") using the WanX2.1-T2V-1.3B model, we demonstrate our integrated approach’s superior ability to follow intricate prompts in multi-element long videos, achieving high-fidelity results with minimal error accumulation.

## 2 Related Work

**Video Generation Dataset.** Large-scale video datasets have driven advancements in video generation, but many existing datasets like YouCook2 [47], VATEX [38], and ActivityNet [5] were not designed for this purpose and lack fine-grained annotations. Similarly, large-scale datasets like YTTemporal-180M [43] and HD-VILA-100M [40] suffer from low-quality captions generated through speech recognition, limiting their utility for high-quality video generation. Datasets like Panda-70M [8] offer extensive data but rely on simplistic global descriptions, which hinder the model’s ability to capture fine temporal details. Newer datasets, such as Koala-36M [35] and LongTake-HD [41], provide more detailed annotations but still rely on segment-level or shot-based annotations, limiting long-duration video generation. In contrast, our method introduces a frame-level captioning approach, where each frame is independently annotated with a description that maintains contextual relevance

to the preceding and succeeding frames. This ensures better alignment between visual content and text while preserving the temporal continuity and motion dynamics, ultimately improving the overall quality of long-form video generation.

**Long Video Generation.** Video generation has evolved from simple single-shot models to more complex long-form and multi-scene models. Early methods relied on GANs [9, 29, 31, 39], constrained by single-domain datasets. Diffusion models [3, 11, 2, 1, 10] introduced temporal layers, enabling motion modeling. DiT-based architectures [4, 25, 28, 17, 42, 32, 23, 15, 20] have achieved tremendous success in scaling diffusion transformers, significantly enhancing video quality. However, these models were limited to generating short clips. FreeNoise [27] and StreamingT2V [13] extended video sequences using auto-regressive methods and temporal attention mechanisms. Gen-L-Video [34] processes videos as sequences of overlapping short clips and employs a temporal co-denoising technique, wherein multiple predictions for each individual frame are averaged. Despite these advancements, challenges in content diversity and temporal consistency persisted. The Diffusion Forcing [6] paradigm addressed these issues by combining diffusion’s high-quality generation with auto-regressive models for sequence extension.

In multi-scene video generation, models like Mask2DiT [26], LCT [12], VideoStudio [21], SKYREELS-V2 [7], MovieDreamer [44], StoryAnchors [33], and VGoT [46] focused on scene-level consistency but struggled with temporal coherence across scenes. Recent methods, including StoryDiffusion [48] and MEVG [24], employed attention mechanisms to enhance visual and dynamic consistency. Our approach uses frame-level attention for dynamic scene extension without fixed scene durations, improving flexibility and coherence in long-form videos. Combined with Diffusion Forcing [6], our method ensures smooth scene transitions, extended video lengths, and maintains both visual richness and temporal consistency.

### 3 Frame-Level Dataset

Previous video-text datasets such as Panda70M [8] and Koala-36M [35] provide only global-level captions, resulting in coarse supervision that cannot reflect detailed visual changes within videos. LongTake-HD [41] offers shot-level sub-captions but still depends on explicit shot boundaries, making it difficult to model continuous motion and intra-shot dynamics. In contrast, our dataset uses frame-level uniform sampling and annotation, enabling dense and temporally continuous supervision. This design captures both subtle and significant changes without being limited by artificial segmentation, supporting more precise alignment between video frames and text descriptions. Overall, our dataset offers finer-grained, structurally consistent, and temporally faithful video-text supervision, facilitating improved learning of dynamic visual content.

**Large-scale frame-level video dataset construction.** We present a frame-level video dataset comprising 700,000 high-quality clips, designed to enhance fine-grained text-video alignment and provide dense semantic supervision for diffusion-based video generation models. The dataset systematically balances visual diversity, temporal continuity, and annotation precision, and can be further improved with larger scale in the future. We collect raw videos longer than 10 minutes from multiple platforms, remove near-duplicate content using perceptual hashing, and discard the first and last 10% of frames to ensure the sampled content is dynamic and semantically meaningful. Each processed video is evenly divided into four segments, from which an 8-second continuous clip is extracted. At a frame rate of 24 fps, one frame is sampled every 8 frames, resulting in approximately 24 frames per clip. This design balances temporal context and computational efficiency, and ensures compatibility with mainstream video VAE tokenization schemes, enabling precise one-to-one frame-token supervision.

**Adaptive frame-level annotation.** We utilize multimodal large language models to generate frame-level captions, automatically choosing between shared or independent descriptions based on the degree of visual change. Identical captions are assigned to frames with minimal differences, while significant changes trigger independent frame-level descriptions, achieving unified and adaptive semantic alignment. To enhance consistency and richness, we design structured annotation prompts that require each frame description to cover main subjects, actions, environment, shot size, and camera angle. Outputs strictly follow JSON format with no redundant commentary, ensuring precise semantic and structural alignment at the frame level. We provide the full frame-level annotation prompt, specifying formatting and content requirements to facilitate reproducibility and further



research; the template is too long to show here, see appendix. During inference, we use gemini pro 2.5 to convert a user input from short/detailed caption to a frame-level detailed caption. More details can be found in appendix.

### 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).
- 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": "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]",
  "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]"
}

```

#### 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.

### 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]",
  "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.

## 4 Method

Our work introduces a set of interconnected improvements for DiT-based video diffusion models, aimed at generating complex, long videos. We focus on enhancing how models understand text instructions for each video part, how they handle timing and changes, and how they create long, consistent videos during inference.

### 4.1 Overall Framework

Current Diffusion Transformer (DiT) based models are skilled at creating high-quality short videos. They typically use a VAE (Variational Autoencoder) to compress videos into compact latent data, and

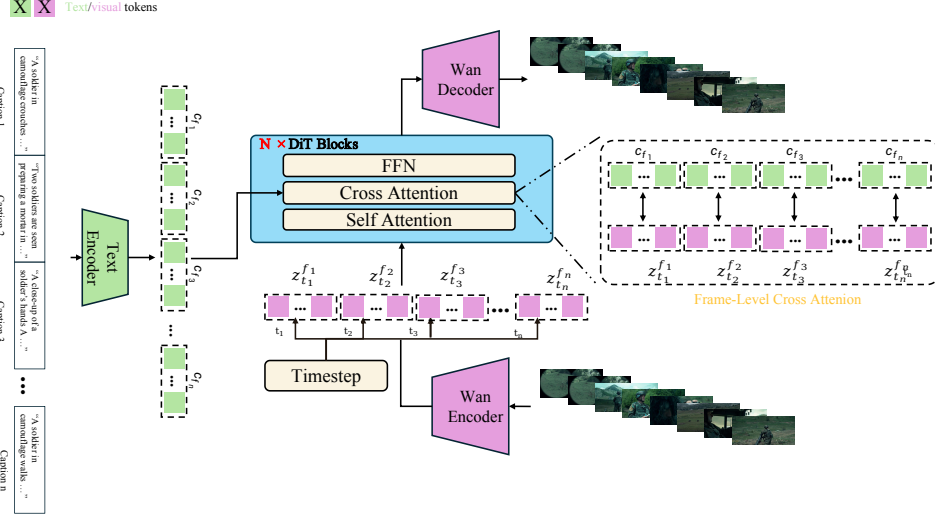


Figure 2: **Overview of the proposed frame-level training method.** Frame-Level Cross-Attention links the visual data of each video segment (latent token) directly to its own specific text description.

a DiT then generates video from these latents. However, when generating long videos with detailed stories and dynamic action, these models face several basic problems:

- **Imprecise Content Control:** Using a single text description (caption) for the entire video often leads to unclear or mixed-up details for different parts of the video, making it hard to accurately control specific events or elements across various scenes.
- **Limited Handling of Timing:** Standard models usually denoise all parts of the video at the same rate in each step. This restricts their ability to show varied motion speeds, different pacing in scenes, or sudden changes effectively.
- **Difficulty with Long Video Coherence:** When creating long videos by generating segments one after another from models trained on short clips, errors tend to build up. This can make the video lose consistency over time.

To address these key challenges, we propose three main contributions:

1. A **Frame-Level Cross-Attention** mechanism (Section 4.2) for precise, localized text-based control over the content of each video segment during training.
2. A **Diffusion Forcing** training strategy (Section 4.3) to teach models how to handle varied timing by exposing them to video segments denoised at different rates.
3. A **Parallel Multi-Window Denoising (PMWD)** inference method (Section 4.4), designed to generate long, coherent videos by significantly reducing the build-up of errors.

While these principles can be applied to many DiT-based video models, we demonstrate our methods by adapting and fine-tuning the WanX2.1-T2V-1.3B [32] framework, a well-known open-source model, as our base.

## 4.2 Frame-Level Cross Attention

To accurately control video content in line with detailed narratives, we introduce Frame-Level Cross-Attention. This method links the visual data of each video segment (latent token) directly to its own specific text description. Simultaneously, the DiT’s standard self-attention mechanism continues to capture overall temporal relationships, ensuring smooth motion. This approach provides both precise local content guidance and global video coherence.

Our process starts by assigning an independent text description to each conceptual "frame" (latent unit) of a video. When the original video is converted into latent data by the VAE, each resulting latent token  $z_f$  is directly paired with the embedding of its corresponding frame-level caption,  $c_f$ .

This creates a detailed, one-to-one mapping between text and video segments over time, offering exact guidance for generation. We modify the DiT’s cross-attention mechanism so that each latent token  $z_f$  attends exclusively to its paired caption embedding  $c_f$ , rather than to a single caption shared by the entire video. Formally, this is:

$$\text{CrossAttention}(q_f, c_f) = \text{Softmax} \left( \frac{q_f W_q (c_f W_k)^T}{\sqrt{d}} \right) (c_f W_v), \quad (1)$$

where  $q_f$  is the query projected from  $z_f$ , and  $W_q, W_k, W_v$  are learnable matrices. This targeted attention mechanism reduces the unclear meaning that can arise from global captions, greatly improving text-to-video alignment and allowing for precise control over dynamic content within each segment.

### 4.3 Diffusion Forcing for Temporal Flexibility

Creating long videos with dynamic action and varied pacing requires the model to handle time flexibly. Standard diffusion models are often limited in this area because they apply the same noise level to all video segments at each step of the denoising process, restricting their ability to generate diverse visual qualities, dynamic changes, or quick scene transitions. To give models this needed flexibility, we use a **Diffusion Forcing (DF)** strategy during training. This technique assigns an independent noise level to each video segment (latent token) in a training sequence. Specifically, we pick a reference segment, set its target noise removal stage (timestep), and then determine the noise stages for other segments in relation to it: preceding segments get "cleaner" (earlier) timesteps, and subsequent segments get "noisier" (later) timesteps. This approach maintains temporal smoothness while training the model to manage different denoising states simultaneously within one sequence.

This training approach makes the model highly adaptable at inference time. By adjusting a "step-size" parameter—which controls the allowed difference in noise schedules between adjacent segments—we can smoothly shift the generation style. We can opt for fully synchronized diffusion (small step-size, for high consistency) or for more dynamic, evolving outputs (large step-size, resembling autoregressive generation). This adaptability allows the model to produce either smooth, consistent videos or to progressively unfold complex scene transitions and actions as guided by the text prompts. Furthermore, already partially denoised historical segments can serve as stable conditions for generating later segments, aiding long-range consistency without forcing all segments to share the same noise level at the same time.

### 4.4 Flexible Inference Modes for Long Video Generation

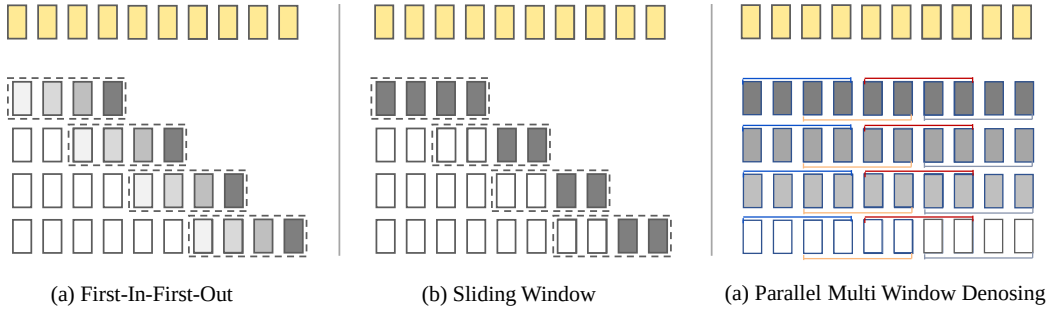


Figure 3: Three inference modes. Multiple yellow tokens represent different frame-level prompts. Less gray value represents lower noise.

The temporal flexibility gained from Diffusion Forcing during training allows for various inference methods to generate videos much longer than the training segments (the "*train short, test long*" approach). **Sequential Sliding Window Approaches.** Common methods for long video generation use a sequential sliding window. These include simple autoregressive techniques, where a new segment of  $M$  latents is generated based on  $N - M$  previous latents from an  $N$ -latent window (often re-noising the context), and more advanced methods like FIFO-Diffusion [16], which uses a queue with diagonally progressing noise levels for better temporal consistency. However, a core problem with all such step-by-step sequential methods is the unavoidable build-up of errors, which reduces quality and long-range consistency in very long videos.

**Parallel Multi-Window Denoising (PMWD).** To effectively overcome this error accumulation problem, we introduce PMWD. This novel inference strategy takes full advantage of our frame-level prompt system to generate long videos more as a complete whole, rather than piece by piece. For a target long video of  $L$  latents, we view it as  $K$  overlapping windows, each the length of a training segment. Crucially, all  $K$  windows are processed *in parallel* (at the same time) during each step of the diffusion denoising process. Every latent, whether new or historical context, is guided by its own dedicated frame-level prompt. This parallel, parallel method for the entire sequence inherently avoids the cascading error build-up seen in typical autoregressive techniques. Latents located in the overlapping regions between adjacent windows are averaged after each denoising step. This averaging, along with the parallel processing, allows information to flow in both directions (bidirectionally) between an earlier and a later window. Unlike methods where only the past influences the future, PMWD allows upcoming video segments to help refine earlier ones. This is especially useful for creating natural-looking scene changes and maintaining consistency in stories with multiple scenes.

---

**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** 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**  $\triangleright K$  window can denoise in parallel
- 9:      $s \leftarrow 0 + F_{\text{slide}} * k$   $\triangleright$  start position
- 10:     $e \leftarrow s + F_{\text{window}}$   $\triangleright$  end position
- 11:     $z_{t-1}[s : e] \leftarrow z_{t-1}[s : e] + D_\theta(z_t[s : e], t, c_{\text{text}}[s : e], c_{\text{others}})$   $\triangleright c_{\text{others}}$  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**  $t = 0 \triangleright$  all frames in the entire long video are denoised to be clean simultaneously

**return**  $z_t$

---

The  $c_{\text{others}}$  parameter in Algorithm 2 represents optional auxiliary conditioning information. This can include frames beyond the current processing window (e.g., distant historical frames) or injected data specifically designed to maintain identity (e.g., ID embeddings). The primary purpose of  $c_{\text{others}}$  is to potentially enhance ID preservation and temporal consistency in long video generation. In this paper, unless otherwise specified,  $c_{\text{others}}$  is not utilized, and the model’s receptive field during inference is confined to the window size employed during its training. Dynamic prompts, particularly frame-



---

**Algorithm 3** 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**  $z_{t_{\tau_0}}$   $\triangleright$  output only one clean frame in each slide of FIFO

---

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

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

## 5 Experimental Results

### 5.1 Experimental Setup

We fully fine-tune the open-source WanX-2.1-T2V-1.3B model with Diffusion Forcing technique on resolution 81x480x832 for 100,000 iterations using our internal dataset (detailed in Section 3) of dense frame-level annotations. Training occurred on H-series GPUs with a global batch size of 64.

### 5.2 Evaluation Dataset

We evaluate the model’s capability to generate complex videos by utilizing prompts from the VBench 2.0 benchmark, specifically focusing on the **Complex Plots** and **Complex Landscapes**.

**Complex Plots** assess the model’s ability to construct coherent and consistent multi-scene narratives based on prompts describing multi-stage events. These prompts often involve extended descriptions (150+ words) outlining a sequence of actions or a story with multiple acts, challenging the model to maintain plot consistency and logical flow throughout the generated video.

**Complex Landscapes** evaluate whether the model can faithfully translate long-form landscape descriptions (150+ words) into video, including multiple scene transitions dictated by camera movements. These prompts test the model’s understanding of spatial relationships and its ability to render dynamic changes in the environment as described in the text.

### 5.3 Evaluation Metrics

We evaluate video quality using metrics for overall video-text alignment and also propose a new metric for the issue of semantic confusion in multi scenes generation. Let  $P_g$  be the global prompt,  $V$  the generated video,  $\{P_1, \dots, P_F\}$  the sequence of  $F$  frame-level prompts, and  $\{V_1, \dots, V_F\}$  the corresponding sequence of generated frames.

**Standard VBench Evaluation.** To provide a comprehensive assessment of fundamental video quality aspects, particularly for the complex scenarios presented by our chosen VBench 2.0 prompt

categories (Complex Plots and Complex Landscapes), we incorporate a curated subset of established metrics from the VBench benchmark. This evaluation focuses on key indicators such as: aesthetic quality, image quality, and motion smoothness. These selected metrics offer standardized measures of the perceptual quality and spatio-temporal coherence of the generated videos.

**Video-Level Video-Text Similarity.** This standard metric evaluates overall coherence between  $P_g$  and  $V$ .  $\Phi_V(V)$  represents overall video features (uniformly sample 8 frames as input of ViClip).

$$\mathcal{S}_{\text{global}} = \text{Sim}(\Phi_T(P_g), \Phi_V(V)) \quad (2)$$

where we use a pre-trained vision-language model (e.g., ViCLIP) for text embeddings  $\Phi_T(\cdot)$  and video/latent 'frame' embeddings  $\Phi_V(\cdot)$ , with  $\text{Sim}(\cdot, \cdot)$  denoting cosine similarity.

**Confusion Degree (CD).** Despite the widespread use of  $\mathcal{S}_{\text{global}}$ , this global metric may assign favorable scores even when content from different scenes are inappropriately combined. To pinpoint such temporal and semantic inaccuracies, we introduce the Confusion Degree (CD). A high CD score reveals a model's difficulty in maintaining a clear, sequential narrative, often resulting in a muddled or incoherent visual story. We first define two fundamental frame-level similarity metrics as follows:

$$\begin{aligned} S_{TT}(P_i, P_j) &= \text{Sim}(\Phi_T(P_i), \Phi_T(P_j)) \\ S_{TF}(P_i, V_j) &= \text{Sim}(\Phi_T(P_i), \Phi_V(V_j)) \end{aligned} \quad (3)$$

, where  $S_{TT}(P_i, P_j)$  represents *frame-level text-text similarity* and  $S_{TF}(P_i, V_j)$  represents *frame-level text-frame similarity*. Then  $\tilde{S}_{TT}(P_i, P_j) = S_{TT}(P_i, P_j)/S_{TT}(P_i, P_i)$  and  $\tilde{S}_{TF}(P_i, V_j) = S_{TF}(P_i, V_j)/S_{TF}(P_i, V_i)$  are applied as normalization function to ensure  $\tilde{S}_{TT}(P_i, P_i) = 1$  and  $\tilde{S}_{TF}(P_i, V_i) = 1$ .

The confusion degree of a text  $P_i$  in the generated video  $V$  is defined as:

$$\text{CD}(P_i) = \sum_{j \in \{1, \dots, F\}} \max(0, \tilde{S}_{TF}(P_i, V_j) - \tilde{S}_{TT}(P_i, P_j)) \quad (4)$$

where  $\tilde{S}_{TF}(P_i, V_j) - \tilde{S}_{TT}(P_i, P_j)$  indicates that the content of  $P_i$  is more aligned with frame  $V_j$  than its inherent semantic relationship with  $P_j$  would suggest, thereby signaling confusion. Then the confusion of a video  $V$  is defined as

$$\text{CD} = \frac{1}{F} \sum_{i=1}^F \text{CD}_i \quad (5)$$

, representing the average confusion degree across all frames. Lower CD values indicate superior narrative consistency and reduced semantic confusion throughout the video.

## 5.4 Comparison and Discussion

| Method                  | Video Length | Prompts Type      | Confusion Degree↓      | Video-level Text-Video Consistency↑ | Frame-level Text-Video Consistency↑ | Motion Smoothness↑ | Aesthetic Quality↑ | Image Quality↑ |
|-------------------------|--------------|-------------------|------------------------|-------------------------------------|-------------------------------------|--------------------|--------------------|----------------|
| DF + Video-level Prompt | 5s           | Complex Plot      | 0.2952 ± 0.0461        | 0.2100 ± 0.0410                     | 0.1635 ± 0.0282                     | 98.43              | <b>59.20</b>       | <b>67.92</b>   |
| DF + Video-level Prompt | 30s          | Complex Plot      | 0.2962 ± 0.0487        | 0.2053 ± 0.0368                     | 0.1518 ± 0.0258                     | <b>98.63</b>       | 52.02              | 58.07          |
| DF + Frame-level Prompt | 30s          | Complex Plot      | <b>0.1385 ± 0.0498</b> | <b>0.2196 ± 0.0309</b>              | <b>0.2054 ± 0.0231</b>              | 98.53              | 55.04              | 61.56          |
| DF + Video-level Prompt | 5s           | Complex Landscape | 0.2745 ± 0.0412        | 0.2101 ± 0.0341                     | 0.1831 ± 0.0227                     | 98.70              | <b>61.32</b>       | <b>59.61</b>   |
| DF + Video-level Prompt | 30s          | Complex Landscape | 0.2806 ± 0.0474        | 0.2066 ± 0.0351                     | 0.1723 ± 0.0230                     | 98.58              | 52.63              | 51.02          |
| DF + Frame-level Prompt | 30s          | Complex Landscape | <b>0.1528 ± 0.0479</b> | <b>0.2195 ± 0.0326</b>              | <b>0.2139 ± 0.0167</b>              | <b>98.99</b>       | 56.31              | 55.98          |

Table 1: Comparing video-level versus frame-level prompting for complex narrative videos. While global Video-Level Text-Video Consistency can yield misleadingly high scores despite internal scene blending or semantic confusion, metrics like Confusion Degree and frame-level consistency more effectively expose these flaws, highlighting the superior prompt adherence of frame-level strategies.

**Analysis of Video Generation under Complex Prompts.** Tab. 1 provides a comparative analysis of models trained and inferred using either global video-level or granular frame-level prompts. When generating short videos (e.g., 5 seconds) conditioned on a single **video-level prompt**, the model operates closer to an ideal scenario without temporal error accumulation. However, such prompts often lead to a high Confusion Degree (CD), as the model struggles to render extensive semantic information within a condensed timeframe, resulting in blended or muddled content.



Figure 4: Complex Plot Generation. This figure illustrates the impact of different prompting strategies on visual storytelling performance in a complex narrative task based on The Ugly Duckling. The first row shows results generated using diffusion forcing with a single global prompt (video-level), while the second row presents results from our proposed method that combines diffusion forcing with multiple tailored prompts (multi-prompting).

| Method                          | Prompt Level | Confusion Degree ↓ | Video-level ↑          | Frame-level ↑          | Motion ↑ | Aesthetic ↑ | Image ↑ |
|---------------------------------|--------------|--------------------|------------------------|------------------------|----------|-------------|---------|
|                                 |              |                    | Text-Video Consistency | Text-Video Consistency |          |             |         |
| First-In-First-Out (FIFO)       | Video        | 0.2962 ± 0.0487    | 0.2053 ± 0.0368        | 0.1518 ± 0.0258        | 98.63    | 52.02       | 58.07   |
| First-In-First-Out (FIFO)       | Frame        | 0.2416 ± 0.0514    | 0.2100 ± 0.0370        | 0.1660 ± 0.0266        | 98.81    | 51.32       | 59.25   |
| Sequential Sliding Window       | Frame        | 0.1773 ± 0.0550    | 0.2134 ± 0.0312        | 0.1842 ± 0.0227        | 98.80    | 52.04       | 60.11   |
| Parallel Multi-Window Denoising | Frame        | 0.1385 ± 0.0498    | 0.2196 ± 0.0309        | 0.2054 ± 0.0231        | 98.53    | 55.04       | 61.56   |

Table 2: Inference method comparison for 30s complex plot videos. Parallel Multi-Window Denoising (PMWD) achieves lower error accumulation (improved aesthetic/image quality) and better prompt adherence (reduced Confusion Degree, higher text-video consistency).

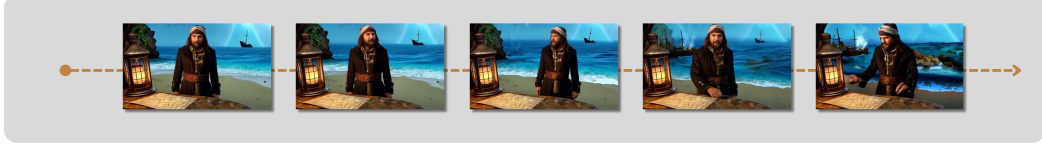
Conversely, employing **frame-level prompts** demonstrates a marked improvement in prompt adherence, evidenced by lower CD scores alongside high frame-level consistency metrics. This enhanced ability to follow detailed, segmented instructions makes the frame-level prompting strategy more reliable and effective for generating coherent multi-scene long videos. Furthermore, metrics such as aesthetic and image quality serve as indirect indicators of error accumulation; significant degradation in these scores over time typically reflects compounding errors. This accumulation is an inherent consequence of the causal nature of sequential generation processes, a fundamental issue that even precise frame-level semantic guidance cannot resolve on its own when operating within such autoregressive frameworks. As shown in Fig. 4, our method demonstrates significantly improved coherence, reduced error accumulation, and less narrative confusion across the sequence. The images generated with multi-prompting maintain better stylistic and semantic consistency, showcasing its superiority over the global prompt approach in handling complex plot developments.

**Comparative Analysis of Long Video Inference Strategies.** In Tab. 2, we further analyze the efficacy of different inference strategies for long video generation, comparing our proposed Parallel Multi-Window Denoising (PMWD) with established sequential methods like FIFO-Diffusion and naive sliding windows. Sequential approaches, by their nature, tackle long video generation segment by segment. Naive sliding window techniques autoregressively generate a new chunk of latents conditioned on a limited history of prior latents (often re-noised to reduce error). FIFO-Diffusion [16] offers a more sophisticated sequential mechanism, processing a queue of latents with diagonally increasing noise levels to output one clean latent per step. While these methods incorporate mechanisms to manage error, they fundamentally struggle with the *inevitable accumulation of errors* over very long sequences. Our proposed Parallel Multi-Window Denoising (PMWD) is architecturally designed to overcome this critical limitation. Instead of sequential generation, PMWD processes the entire target long video (composed of multiple overlapping windows) *simultaneously* at each denoising step, with each latent guided by its specific frame-level prompt. The averaging of latents in overlapping regions is a key aspect of PMWD. This, combined with parallel processing, not only fuses information effectively but also transforms the strictly causal dependency of sequential models into a *bidirectional contextual influence*, where information from temporally subsequent windows can refine earlier ones. This capability is particularly advantageous for rendering naturalistic scene transitions and ensuring global narrative consistency.

## 6 Application By enabling Frame-level Prompts

### 6.1 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 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 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

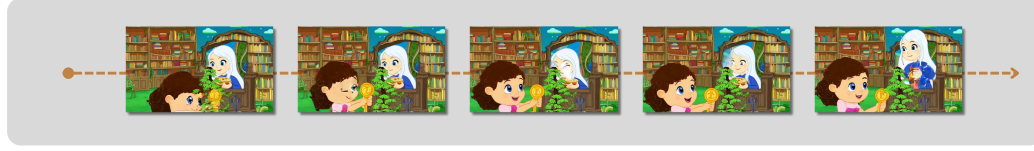


A young boy receives a letter from the future, saying that he will discover a secret that will change the world. // On the first day, he finds a dusty book filled with unknown symbols. // The next day, he finds a map in the book, guiding him to an abandoned castle. // On the third day, he enters the 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 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.



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

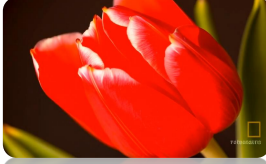


A little girl received a mysterious gift—the ancient time key. She used the key to open a sealed door and entered the time house. // On the first day, she found a room filled with books, each with a different story. // On the second day, she was taken to an ancient kitchen filled with strange smells, 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 7: 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.

## 6.2 (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.



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.



4. Extreme close-up, sharp focus: A skilled barista meticulously prepares a pour-over coffee. Hot water spirals from a gooseneck kettle onto freshly ground coffee in a ceramic dripper. Watch the coffee grounds bloom, rich aromas seemingly wafting, and a **steady stream of dark coffee** slowly dripping into a glass server below. Warm lighting, detailed textures of coffee and steam.

5. A **professional skateboarder** in a sun-drenched Venice Beach skatepark executes a complex kickflip over a set of stairs. Captured in dramatic slow-motion (e.g., 240fps playback), showcasing the board's rotation, the athlete's focused expression, and the subtle flex of the board upon landing. Bright California sunlight, dynamic action, sharp detail.

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 8: 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. 8, 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.

### 6.3 (Short Video) Transition Generation (First-Last-Frame-to-Video) Ability

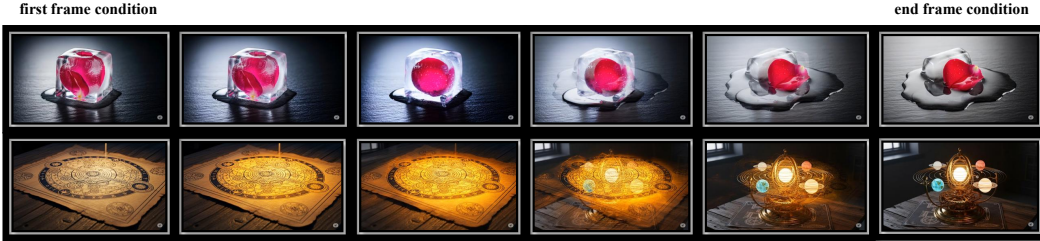


Figure 9: First-and-Last-Frame-to-Video Generation Enabled by Diffusion Forcing and Our Dynamic (Frame-Level) Prompts.

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.

## 7 Conclusion

Generating long, narratively complex videos with high fidelity remains challenging, primarily due to issues with coarse semantic guidance and the error accumulation inherent in common sequential generation techniques. We propose a comprehensive solution combining fine-grained frame-level annotations, novel training strategies, and a Parallel Multi-Window Denoising (PMWD) inference method. Our experiments on demanding VBench 2.0 benchmarks demonstrate that this integrated system significantly improves prompt adherence for complex, multi-scene narratives in ultra-long videos, achieving high-quality results with minimal error accumulation.



## References

- [1] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, et al. Lumiere: A space-time diffusion model for video generation. In *SIGGRAPH*, pages 1–11, 2024.
- [2] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.
- [3] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*, pages 22563–22575, 2023.
- [4] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1:8, 2024.
- [5] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, pages 961–970, 2015.
- [6] Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2024.
- [7] Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Juncheng Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengchen Ma, et al. Skyreels-v2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025.
- [8] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *CVPR*, pages 13320–13331, 2024.
- [9] Mengyu Chu, You Xie, Jonas Mayer, Laura Leal-Taixé, and Nils Thuerey. Learning temporal coherence via self-supervision for gan-based video generation. *TOG*, 39(4):75–1, 2020.
- [10] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Factorizing text-to-video generation by explicit image conditioning. In *ECCV*, pages 205–224. Springer, 2024.
- [11] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023.
- [12] Yuwei Guo, Ceyuan Yang, Ziyang Yang, Zhibei Ma, Zhijie Lin, Zhenheng Yang, Dahua Lin, and Lu Jiang. Long context tuning for video generation. *arXiv preprint arXiv:2503.10589*, 2025.
- [13] Roberto Henschel, Levon Khachatryan, Daniil Hayrapetyan, Hayk Poghosyan, Vahram Tadevosyan, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Streamingt2v: Consistent, dynamic, and extendable long video generation from text. *arXiv preprint arXiv:2403.14773*, 2024.
- [14] Emiel Hoogeboom, Alexey A Gritsenko, Jasmijn Bastings, Ben Poole, Rianne van den Berg, and Tim Salimans. Autoregressive diffusion models. In *International Conference on Learning Representations*, 2023.
- [15] Haoyang Huang, Guoqing Ma, Nan Duan, Xing Chen, Changyi Wan, Ranchen Ming, Tianyu Wang, Bo Wang, Zhiying Lu, Aojie Li, et al. Step-video-ti2v technical report: A state-of-the-art text-driven image-to-video generation model. *arXiv preprint arXiv:2503.11251*, 2025.
- [16] Jihwan Kim, Junoh Kang, Jinyoung Choi, and Bohyung Han. Fifo-diffusion: Generating infinite videos from text without training, 2024. URL <https://arxiv.org/abs/2405.11473>.

- [17] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.
- [18] Chengxuan Li, Di Huang, Zeyu Lu, Yang Xiao, Qingqi Pei, and Lei Bai. A survey on long video generation: Challenges, methods, and prospects, 2024. URL <https://arxiv.org/abs/2403.16407>.
- [19] Shuang Li, Yihuai Gao, Dorsa Sadigh, and Shuran Song. Unified video action model, 2025. URL <https://arxiv.org/abs/2503.00200>.
- [20] Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024.
- [21] Fuchen Long, Zhaofan Qiu, Ting Yao, and Tao Mei. Videostudio: Generating consistent-content and multi-scene videos. In *ECCV*, pages 468–485. Springer, 2024.
- [22] Guoqing Ma, Haoyang Huang, Kun Yan, Liangyu Chen, Nan Duan, Shengming Yin, Changyi Wan, Ranchen Ming, Xiaoni Song, Xing Chen, Yu Zhou, Deshan Sun, Deyu Zhou, Jian Zhou, et al. Step-video-t2v technical report: The practice, challenges, and future of video foundation model, 2025. URL <https://arxiv.org/abs/2502.10248>.
- [23] Guoqing Ma, Haoyang Huang, Kun Yan, Liangyu Chen, Nan Duan, Shengming Yin, Changyi Wan, Ranchen Ming, Xiaoni Song, Xing Chen, et al. Step-video-t2v technical report: The practice, challenges, and future of video foundation model. *arXiv preprint arXiv:2502.10248*, 2025.
- [24] Gyeongrok Oh, Jaehwan Jeong, Sieun Kim, Wonmin Byeon, Jinkyu Kim, Sungwoong Kim, and Sangpil Kim. Mevg: Multi-event video generation with text-to-video models. In *ECCV*, pages 401–418. Springer, 2024.
- [25] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023.
- [26] Tianhao Qi, Jianlong Yuan, Wanquan Feng, Shancheng Fang, Jiawei Liu, Siyu Zhou, Qian He, Hongtao Xie, and Yongdong Zhang. Mask2dit: Dual mask-based diffusion transformer for multi-scene long video generation. *arXiv preprint arXiv:2503.19881*, 2025.
- [27] Haonan Qiu, Menghan Xia, Yong Zhang, Yingqing He, Xintao Wang, Ying Shan, and Ziwei Liu. Freenoise: Tuning-free longer video diffusion via noise rescheduling. *ICLR*, 2023.
- [28] Team Seaweed, Ceyuan Yang, Zhijie Lin, Yang Zhao, Shanchuan Lin, Zhibei Ma, Haoyuan Guo, Hao Chen, Lu Qi, Sen Wang, et al. Seaweed-7b: Cost-effective training of video generation foundation model. *arXiv preprint arXiv:2504.08685*, 2025.
- [29] Ivan Skorokhodov, Sergey Tulyakov, and Mohamed Elhoseiny. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *CVPR*, pages 3626–3636, 2022.
- [30] Gabriel Tseng, Ruben Cartuyvels, Ivan Zvonkov, Mirali Purohit, David Rolnick, and Hannah Kerner. Lightweight, pre-trained transformers for remote sensing timeseries. *arXiv preprint arXiv:2304.14065*, 2023.
- [31] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. In *CVPR*, pages 1526–1535, 2018.
- [32] Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.
- [33] Bo Wang, Haoyang Huang, Zhiyin Lu, Fengyuan Liu, Guoqing Ma, Jianlong Yuan, Yuan Zhang, and Nan Duan. Storyanchors: Generating consistent multi-scene story frames for long-form narratives, 2025. URL <https://arxiv.org/abs/2505.08350>.

- [34] Fu-Yun Wang, Wenshuo Chen, Guanglu Song, Han-Jia Ye, Yu Liu, and Hongsheng Li. Gen-l-video: Multi-text to long video generation via temporal co-denoising. *arXiv preprint arXiv:2305.18264*, 2023.
- [35] Qiheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, et al. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. *arXiv preprint arXiv:2410.08260*, 2024.
- [36] Weihang Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. Lvbench: An extreme long video understanding benchmark, 2024.
- [37] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent, 2024. URL <https://arxiv.org/abs/2403.10517>.
- [38] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *ICCV*, pages 4581–4591, 2019.
- [39] Yaohui Wang, Piotr Bilinski, Francois Bremond, and Antitza Dantcheva. Imaginator: Conditional spatio-temporal gan for video generation. In *WACV*, pages 1160–1169, 2020.
- [40] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In *CVPR*, pages 5036–5045, 2022.
- [41] Xin Yan, Yuxuan Cai, Qiuyue Wang, Yuan Zhou, Wenhao Huang, and Huan Yang. Long video diffusion generation with segmented cross-attention and content-rich video data curation. *arXiv preprint arXiv:2412.01316*, 2024.
- [42] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.
- [43] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Neurips*, 34:23634–23651, 2021.
- [44] Canyu Zhao, Mingyu Liu, Wen Wang, Jianlong Yuan, Hao Chen, Bo Zhang, and Chunhua Shen. Moviedreamer: Hierarchical generation for coherent long visual sequence. *arXiv preprint arXiv:2407.16655*, 2024.
- [45] Dian Zheng, Ziqi Huang, Hongbo Liu, Kai Zou, Yinan He, Fan Zhang, Yuanhan Zhang, Jingwen He, Wei-Shi Zheng, Yu Qiao, and Ziwei Liu. Vbench-2.0: Advancing video generation benchmark suite for intrinsic faithfulness, 2025. URL <https://arxiv.org/abs/2503.21755>.
- [46] Mingzhe Zheng, Yongqi Xu, Haojian Huang, Xuran Ma, Yexin Liu, Wenjie Shu, Yatian Pang, Feilong Tang, Qifeng Chen, Harry Yang, et al. Videogen-of-thought: A collaborative framework for multi-shot video generation. *arXiv preprint arXiv:2412.02259*, 2024.
- [47] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, volume 32, 2018.
- [48] Yupeng Zhou, Daquan Zhou, Ming-Ming Cheng, Jiashi Feng, and Qibin Hou. Storydiffusion: Consistent self-attention for long-range image and video generation. *Neurips*, 37:110315–110340, 2024.