CI-VID: A Coherent Interleaved Text-Video Dataset

Yiming Ju[1*], **Jijin Hu**[2*], **Zhengxiong Luo**[1*], **Haoge Deng**[2*], **hanyu Zhao**[1], **Li Du**[1], **Wenbo Xiao**[3], **Chengwei Wu**[1], **Donglin Hao**[1], **Xinlong Wang**[1 †], **Tengfei Pan**[1 †]
[1] Beijing Academy of Artificial Intelligence
[2] Beijing University of Posts and Telecommunications
[3] University of New South Wales
{ymju, tfpan, wangxinlong}@baai.ac.cn

Abstract

Text-to-video (T2V) generation has recently attracted considerable attention, resulting in the development of numerous high-quality datasets that have propelled progress in this area. However, existing public datasets are primarily composed of isolated text–video (T–V) pairs and thus fail to model inter-clip relationships. To address this limitation, we introduce CI-VID, a dataset that moves beyond isolated T2V generation toward text-and-video-to-video (T&V2V) generation. CI-VID contains over 340,000 samples, each comprising a semantically coherent video sequence with interleaved text captions that capture both clip-level content and inter-clip relationships. To validate its effectiveness, we design a comprehensive, multi-dimensional benchmark incorporating human evaluation, VLM-based assessment, and similarity-based metrics. Experimental results demonstrate that models trained on CI-VID significantly improve both accuracy and content consistency in multi-clip video generation. This enables the creation of story-driven content with smooth transitions and strong semantic coherence. The dataset is available at https://github.com/ymju-BAAI/CI-VID.

1. Introduction

Recent advances in Artificial Intelligence Generated Content (AIGC) have been largely driven by growing data and compute [19]. In the field of computer vision, the success of recent text-to-video (T2V) models, such as Sora [4], VideoPoet [20], Emu3 [37], CogVideoX [41], and VideoTetris [34], has notably expanded the possibilities for visual content generation, enabling the automatic creation of hyper-realistic videos based on human instructions.

Researchers have contributed many high-quality video generation datasets to advance the field, including OPEN-VID [26], InternVid [38], ShareGPT4Video [7], Vript [40], Koala-36M [36], among others. Although these datasets provide high-quality video clips paired with text captions, most consist solely of isolated text–video (T–V) pairs in a one-to-one correspondence, without modeling inter-clip relationships. This one-to-one pairing paradigm presents two main limitations:

1. Models trained solely on isolated T-V pairs struggle to generate semantically coherent multi-clip videos. Existing datasets typically segment videos at shot boundaries and annotate each shot independently. However, real-world videos often consist of multiple semantically coherent shots that collectively depict a complete scene. For example, Figure 1 shows a tutorial on modifying a black hanger, where the procedure unfolds through a sequence of clips, each conveying a part of the overall process. Similarly, in storytelling, movies, and news broadcasts, scenes are rarely captured in a single continuous shot but rather constructed from a series of semantically linked clips. Owing to the constrained one-to-one correspondence in current datasets, models trained on them often fail to maintain coherent characters, visual styles, and scene transitions across multi-clip generations.

2. Independent T–V pairs do not support text-and-video-to-video (T&V2V) generation. In video extrapolation tasks, prior methods typically rely solely on preceding visual frames as input [35], often leading to repetitive outputs [13] and limited semantic control. To meaningfully guide extrapolated content, textual inputs are essential in addition to visual inputs as conditioning signals (T&V2V generation). However, as most existing datasets consist of isolated T–V pairs, they are inherently inadequate for learning generation conditioned jointly on visual and textual inputs, thus limiting their applicability in T&V2V generation tasks.

These limitations hinder the use of existing datasets in research on complex video generation tasks beyond unit-level T2V, such as story generation, video rewriting, and

[*]Equal Contribution. [†]Corresponding Author.

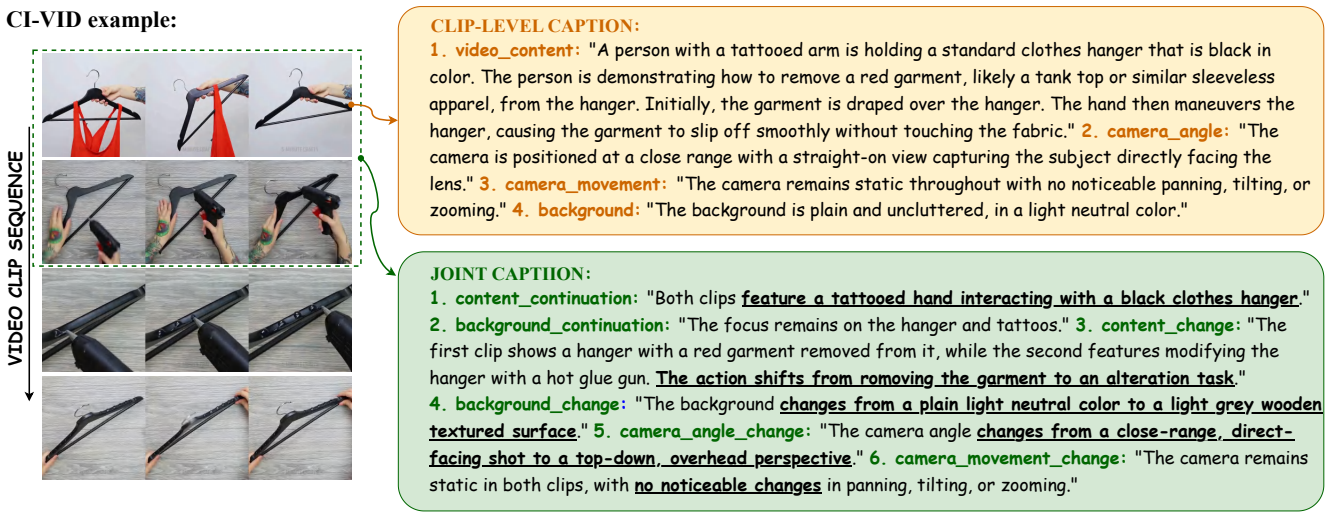


Figure 1. An example from the CI-VID dataset. Each sample consists of a sequence of video clips (shown on the left), with clip-level captions (orange box) describing individual clips and joint captions (green box) capturing continuity and change across adjacent clips.

other high-level scenarios.

To address this gap, we introduce CI-VID (Coherent Interleaved Video Dataset)—a carefully curated dataset designed to model inter-clip relationships beyond clip-level descriptions. The key characteristics of CI-VID are as follows:

- **High-Quality Video Content.** CI-VID sources its videos from over 4,000 carefully curated YouTube channels spanning a wide range of themes. Video clips are rigorously filtered based on on-screen text ratio, motion differences, and visual clarity, with fewer than 20% retained for further processing.
- **Semantically Coherent while Visually Diverse Video Sequences.** As shown in Figure 1, CI-VID video sequences maintain semantic coherence while exhibiting rich visual diversity. Semantic coherence enables earlier clips to provide context for generating subsequent ones, while visual diversity—such as shot transitions, action changes, and new entities—ensures textual captions to offer meaningful guidance rather than merely echoing the visual input.
- **Interleaved, High-Quality Text Captions.** As shown in Figure 1, CI-VID provides detailed and structured captions that go beyond individual clip descriptions by capturing both the continuity and contrast between adjacent clips. These enriched captions enable video generation guided jointly by both visual and textual context.

CI-VID contains over 340,000 samples. To evaluate its effectiveness, we establish a comprehensive benchmark for coherent multi-clip video generation, integrating human evaluation, vision-language model (VLM)-based assessment, and similarity-based metrics. Experimental results demonstrate that models fine-tuned on CI-VID can generate coherent, story-driven videos with smooth transitions and consistent content, significantly outperforming baseline methods. The CI-VID datasets, including both captions and videos, and the accompanying code for data construction and evaluation are released.

Our key contributions are summarized as follows:

1. We identify the limitations of existing isolated T–V pair datasets and highlight the need for resources supporting both T-to-V and T&V-to-V modeling to enable coherent and controllable video sequence generation.
2. We present CI-VID, a large-scale dataset containing coherent text–video sequences with both clip-level content and inter-clip relationships.
3. We establish a benchmark for coherent multi-clip video generation and empirically validate the effectiveness of CI-VID.

2. Related Work

Text-to-Video Datasets. Building powerful T2V models requires high-quality text-video datasets. Existing datasets like LSMDC [30], ActivityNet [5], HowTo100M [25], WebVid-10M [3], HDVILA-100M [39], and Panda-70M [8] provide text-video pairs, but they suffer from limitations such as poor video quality and noisy captions. More recent datasets, including VidGen-1M [32], OpenVid-1M [26], MiraData [19], Vript [40], InternVid [38], Koala-36M [36], and ShareGPT4Video [7], have addressed some limitations by introducing longer videos with richer motion dynamics, along with more detailed, accurate, or structured descriptions. However, they still primarily consist of isolated text–video pairs, lacking modeling of inter-clip relationships.

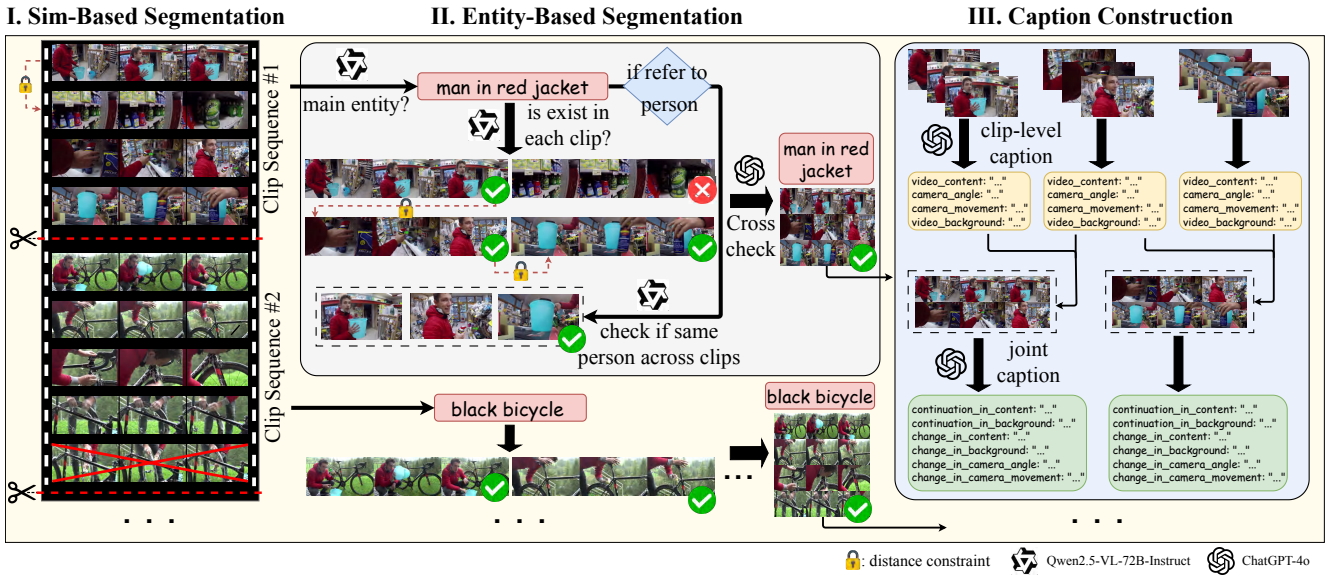


Figure 2. Pipeline for constructing CI-VID samples from raw videos. Modules I and II segment raw videos into coherent multi-clip sequences, while Module III generates clip-level captions and joint captions for adjacent clips.

Interleaved Datasets. The concept of interleaved data, widely explored in the image–text domain, refers to sequences where text and visual content appear alternately. Models such as Flamingo [1] and KOSMOS-1 [15] show that training on interleaved data yields better performance than using isolated image–caption pairs, highlighting the benefits of leveraging correlations in interleaved content. While recent datasets like MMC4 [42], OBELICS [21] and CoMM [9] enable research on interleaved image–text modeling. However, this paradigm remains largely unexplored for video generation. CI-VID fills this gap by introducing the first large-scale interleaved text–video dataset.

3. CI-VID Dataset Construction

This section describes the construction process of CI-VID.

3.1. Raw Video Collection and Preprocessing

CI-VID construction requires complete raw videos rather than pre-segmented clips. Thus, CI-VID collects raw videos directly from YouTube-similar to many existing public datasets [7, 8, 19, 38–40]—rather than relying on existing large-scale video datasets such as Panda-70M or HDVILA-100M.

3.1.1. Raw Video Collection

To ensure raw video quality, collection was first performed at the channel level. Specifically, the training data from Emu3 [37] was used to extract corresponding YouTube channels. From this pool, 4,068 high-quality channels were manually selected. Annotators reviewed candidate channels based on resolution, color fidelity, motion strength, and watermark presence, without imposing content restrictions[1] Each annotator also received daily expert checks; if their agreement rate dropped below 80%, the day's work was redone. All public videos from the selected channels were then downloaded, resulting in a collection of 592,429 raw videos.

3.1.2. Preliminary Segmentation and Filtering

The raw videos were first segmented into clips using content-aware detection of PySceneDetect[2] with a threshold of 3, ensuring that each clip contained a single shot. Long-duration clips were evenly split to ensure that no clip exceeded ten seconds. Moreover, clips shorter than one second are filtered to ensure sufficient duration for model training. Next, optical flow [33] is computed every 0.5 seconds to ensure sufficient motion strength. The average flow magnitude per pixel, normalized by the shorter edge of the frame, is used to filter out clips falling below a threshold of 70. Then, text detection is performed using PaddleOCR [11], and clips with excessive text coverage (more than 10%) are discarded. Overall, these filtering steps eliminate over 80% of the candidate clips.

3.2. Video Sequence Construction and Caption Generation

Constructing semantically coherent while visually diverse video sequences is central to CI-VID. To support T&V2V

[1]The annotation team consisted of six professional annotators, each holding a bachelor's degree. They underwent training with 200 sample cases and were required to review at least three videos per channel.

[2]github.com/Breakthrough/PySceneDetect

generation, sequences must exhibit semantic continuity, allowing earlier clips to serve as a foundation for generating subsequent ones—while also incorporating sufficient variation to enable meaningful textual guidance (e.g., shot transitions, new entities, and action shifts). Simply extracting consecutive clips from raw videos fails to meet these requirements.

To address this, we introduce a two-stage segmentation pipeline—Similarity-Based Segmentation followed by Entity-Based Segmentation—as illustrated in Figure 2 (Module I and Module II). Moreover, CI-VID generates not only individual captions for each clip, but also joint captions that describe relationships between adjacent clip (Module III).

3.2.1. Module I: Similarity-Based Segmentation

This module segments raw videos into distinct sequences by measuring embedding-level visual similarity between adjacent clips, providing a coarse partition that forms the initial pool of candidate sequences for CI-VID.

- Strategy. As illustrated in Figure 2, if the similarity between two adjacent clips is below a lower threshold, the clips are considered to indicate a scene transition and are segmented into different sequences (red dashed lines). Conversely, clips that are overly similar to their preceding clip are considered to lack visual diversity and are thus excluded from the sequence (red cross marks).

- Implementation. As shown in Figure 2, three frames are uniformly sampled from each clip and concatenated horizontally to form a visual representation. We found that widely used intermediate/key-frame-based methods were significantly less effective for detecting scene transitions than using concatenated multi-frame representations. This is likely because the latter provides richer temporal and contextual information through spatial encoding.

The ImageBind model [14] is then used to extract embeddings, with cosine similarity employed as the similarity metric. The similarity thresholds are empirically set to $(T_l, T_h) = (0.6, 0.8)$ to ensure high-quality segmentation, prioritizing quality over quantity. Sequences containing only a single clip are discarded.

- Distance Constraint. Notably, due to clip-level filtering during both raw video preprocessing and segmentation, the remaining "adjacent" clips in a sequence may no longer be contiguous in the original video. As a larger index gap between clips increases the risk of semantic discontinuity, a distance constraint (indicated by the lock icon in Figure 2) is imposed: adjacent clips must be no more than three indices apart in the original video; otherwise, a segmentation break is introduced.

Table 1. Prompt used for main entity extraction.

In this figure, each column contains an image. Can you identify the most common entity (objects/people/goals) among these images? Note: *1) Only return the most common entity.* *2) The entity must be the same one.* *3) The entity must be the main entity, not the background or edge entity.* *4) The entity must appear in more than 60% of the images. Return* `'none'` *if there are none.* *5) Return the entity name directly, with its characteristics.* *6) The same person is also an entity, return person's characteristics (e.g., hair, dress), do not guess person's name.*

3.2.2. Module II: Entity-Based Segmentation

Given the inherent diversity and complexity of video content, embedding-level visual similarity alone is insufficient to ensure semantic coherence. To construct truly semantically coherent sequences, this module leverages the reasoning capabilities of VLMs for entity-based segmentation, serving as a crucial step in CI-VID construction.

- Strategy. A sequence in which all clips share a common main entity is considered likely to exhibit semantic coherence, even in the presence of visual diversity and temporal discontinuities. Accordingly, this module uses the presence of a shared main entity as a proxy for semantic coherence, and refines initial clip sequences by enforcing entity-level consistency.

- Implementation. As illustrated in Figure 2, Entity-Based Segmentation focuses on identifying and verifying the main entity within a video sequence via interaction with state-of-the-art vision-language models (Qwen2.5-VL-72B-Instruct [2] and GPT-4o [17]). The process comprises four main steps:

1. **main entity extraction:** The Qwen model is employed to identify the main entity within each sequence. The input is a $3 \times n$ grid image, where n is the sequence length; each row contains three uniformly sampled frames from a single clip. The prompt is shown in Table 1. Sequences with no identifiable main entity are discarded.
2. **clip entity examination.** The Qwen model is queried to verify whether the extracted main entity appears in each clip. Three uniformly sampled frames from each clip are individually examined, and a clip is considered to pass if the entity is detected in at least one frame. Clips that fail this examination are removed from the sequence. If fewer than 70% of the clips pass, the entire sequence is discarded.
3. **same-person verification.** Failure cases may occur when different individuals across clips share similar vi-

sual features.[3] To address this, one representative frame from each clip is selected and concatenated into a single image. The Qwen model is then queried to verify whether the same individual appears across all clips. If not, the entire sequence is discarded.

4. **cross-validation**: The previous three steps rely on the Qwen model. To avoid potential errors caused by model-specific biases or limitations, GPT-4o is used to detect and filter out unqualified sequences. The visual input remains identical to that used in main entity extraction, and GPT-4o is prompted to verify whether the sequence and the extracted main entity meet the requirements in Table 1.

3.2.3. Module III: Caption Generation

This module employs GPT-4o to generate clip-level and joint captions. The two-stage design allows CI-VID to capture both fine-grained clip details and high-level inter-clip relationships, supporting coherent video generation.

- Strategy. We observe that different input strategies offer complementary strengths: the *sequential-frame input strategy*—feeding frames into the model sequentially—produces more detailed descriptions, such as intricate background compositions and fine-grained object features. In contrast, the *joint-frame input strategy*—combining multiple frames into a single image—better captures high-level scene relationships, such as character transitions and shifts in perspective. Thus, as illustrated in Module III of Figure 2, clip-level captions are first generated using sequential frames to capture visual details, followed by joint captions generated using joint frames to model inter-clip relationships, with the clip-level captions serving as textual guidance.

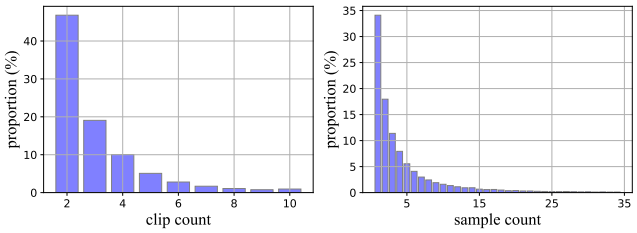
- Implementation. For clip-level caption generation, 4–8 frames are sampled from each clip at even intervals based on its duration. These frames are then sequentially fed into GPT-4o to generate a structured caption covering four key aspects: *video content*, *camera angle*, *camera movement*, and *video background*. For joint caption generation, an $x \times 2$ grid image is constructed as input, where each row contains x frames sampled evenly from a clip.[4] Since video content and background are often complex and exhibit both continuity and variation across clips, the model is prompted to generate descriptions that explicitly address both aspects. Each joint caption covers six aspects: *continuation in video content*; *change in video content*; *continuation in video background*; *change in video background*; *change in camera angle*; and *change in camera movement*.

[3] For example, if the main entity is "a person with a black T-shirt," Clip 1 may feature Person A, while Clip 2 features Person B in similar attire. Although A and B are different individuals, the sequence may still erroneously pass the entity verification step.

[4] x ranges from 3 to 5, depending on the longer clip duration.

Table 2. Comparison of CI-VID with existing large-scale text-video datasets. "VD" denotes average video duration, and "TL" denotes average text length. CI-VID$_{\text{clips}}$ refers to treating each clip and its clip-level caption as an independent sample. CI-VID refers to the proposed dataset, which consists of interleaved video–text samples.

Dataset	sample num	VD (s)	TL (words)	structure caption	inter-clip realtion
LSMDC	118K	4.8	7.0	✗	✗
ActivityNet	100K	36.0	13.5	✗	✗
HowTo100M	136M	3.6	4.0	✗	✗
WebVid-10M	10M	18.0	12.0	✗	✗
HD-VG-130M	130M	5.1	9.6	✗	✗
VidGen-1M	1M	7.4	89.3	✗	✗
InternVid	234M	13.4	32.5	✗	✗
Panda-70M	70M	8.5	13.2	✗	✗
Koala-36M	36M	3.8	202.1	✓	✗
OPENVID-1M	1M	7.2	98.3	✗	✗
MiraData	798K	72.1	318.0	✓	✗
Vript	420K	11.1	145.0	✗	✗
CI-VID$_{clips}$	1M	4.7s	218.6	✓	✗
CI-VID	342K	14.0s	1071.6	✓	✓



(a) Video Sequence Length (b) Sample Generated per Raw Video

Figure 3. The analysis of sample characteristics: (a) the number of samples generated per raw video. (b) The distribution of video sequence length.

3.3. Dataset Statistics and Analysis

3.3.1. Basic Characteristics

Table 2 compares CI-VID with existing large-scale text–video datasets.[5] CI-VID comprises a total of 341,550 samples, with over 98% of videos in 1080p or higher resolution. The dataset includes 1M T–V pairs, offering a sufficient scale for fine-tuning T2V generation models. A strict PySceneDetect threshold of 3 is used during preprocessing to ensure that each clip remains within a single visual shot, resulting in an average clip duration of 4.7 seconds. In contrast, existing datasets adopt higher thresholds—e.g., MiraData (26), InternVid (27), and Panda-70M (25)—which can produce longer clips but may introduce shot transitions

[5] CI-VID$_{\text{clips}}$ refers to treating each clip and its clip-level caption as an independent sample.

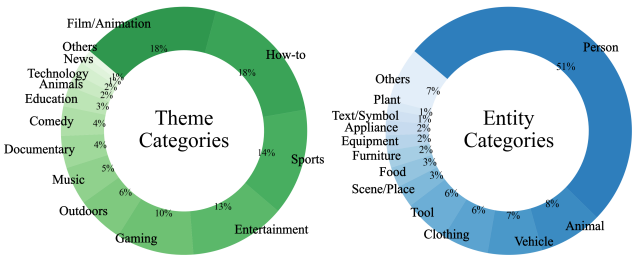


Figure 4. Theme and entity distributions of CI-VID.

(rapid optical changes) that serve as noise during model training.

As shown in Table 2, CI-VID provides structured captions with an average length exceeding 200 words, offering rich and detailed textual supervision. Importantly, unlike existing datasets limited to independent T–V pairs, CI-VID features interleaved T–V sequences that capture inter-clip relationships. On average, each sample contains 3.1 video clips. As shown in Figure 3 (a), over 30% of CI-VID samples (more than 100K) contain four or more clips, making the dataset suitable not only for pairwise learning but also for sequence-level learning.

3.3.2. Video Content Distribution

CI-VID samples are derived from 63,807 original YouTube videos. As shown in Figure 3 (b), most source videos contribute fewer than five samples, indicating that CI-VID avoids overrepresentation from a small subset of videos and maintains source diversity. To analyze thematic coverage, we utilize video metadata (titles, tags, channels, and descriptions) and classify each sample into an official YouTube category using Qwen2.5-72B-Instruct [28], followed by clustering related categories into broader themes.

As illustrated in Theme Categories of Figure 4, CI-VID covers a wide range of open-domain scenarios, including film and animation, how-to content, entertainment, gaming, outdoor scenes, and more, demonstrating its diverse thematic composition. Furthermore, since CI-VID sequences are built around main entities, we analyze the distribution of main entity types in the dataset. As shown in Entity Categories of Figure 4, most samples feature persons as the main entity, accounting for approximately half of all samples. The remaining non-human entities span a diverse set of object categories, including animals, vehicles, tools, scenes, and more, highlighting the range of entity types represented in the dataset.

4. Experiment

To validate the effectiveness of CI-VID, we train a video generation model and evaluate it using a multi-dimensional benchmark specifically designed for coherent multi-clip generation.

Table 3. An example test prompt used for evaluation.

Scene #1: *"A curious little girl wearing a bright yellow dress tiptoes into a lush botanical garden, wide-eyed as she takes in the vibrant flowers and towering trees surrounded by crystal-clear ponds."*
Scene #2: *"She spots a giant butterfly with shimmering blue wings fluttering over a bed of purple orchids and begins to follow it, her footsteps light and careful."*
Scene #3: *"The butterfly leads her to a magnificent greenhouse, its glass walls reflecting the green world outside. Inside, tropical plants with oversized leaves spiral toward the ceiling."*
Scene #4: *"Suddenly, the girl comes across an ancient, worn bench beneath a sprawling tree. She settles down and notices a squirrel nibbling on a tiny nut, staring curiously at her."*
Scene #5: *"After feeding the squirrel a crumb from her pocket, the girl notices brilliant golden rays of sunlight breaking through the glass ceiling, lighting up the garden like a magical wonderland."*
Scene #6: *"The butterfly lands gently on her shoulder, and she laughs in delight as the camera pans out, showing her peacefully seated amidst the blooming paradise."*

4.1. Model Setting and Implementation

We primarily follow the approach of the T2V model NOVA-0.6B [12], which sequentially predicts temporal frames, to process the interleaved text–video data in CI-VID. The model comprises a temporal encoder, a spatial encoder, and a decoder—each with 16 layers and a hidden dimension of 1024, resulting in 0.6 billion parameters. The denoising multi-layer perceptron (MLP) consists of three blocks, each with a dimension of 1280. For spatial modeling, we use the encoder-decoder architecture from MAR [22]. Following Lin et al. [23], we leverage a pre-trained and frozen variational autoencoder (VAE) as an image encoder to achieve spatio-temporal compression of the video, achieving 4×4 compression in the temporal dimension and 8×8 compression in the spatial dimension. During training, we apply the masking and diffusion schedulers from Nichol and Dhariwal [27], using a masking ratio ranging from 0.7 to 1.0. In the inference phase, the ratio is gradually decreased from 1.0 to 0 according to a cosine schedule [6].

The captions and videos are first tokenized into text and visual tokens using a pre-trained language model [18] and an image encoder. These tokenized elements are then sequentially arranged into an input sequence that preserves their interleaved structure. For example, the input sequence is structured as follows: [$caption_{individual\#1} \rightarrow video_{clip\#1}$ + ($caption_{individual\#2}$, $caption_{joint\#1}$) $\rightarrow video_{clip\#2}$ ...], and so on. Supervision is applied exclusively to the visual tokens through a diffusion loss [22]. For optimization, we use the AdamW optimizer [24] with $\beta_1 = 0.9$ and $\beta_2 = 0.95$, a weight decay of 0.02, and a

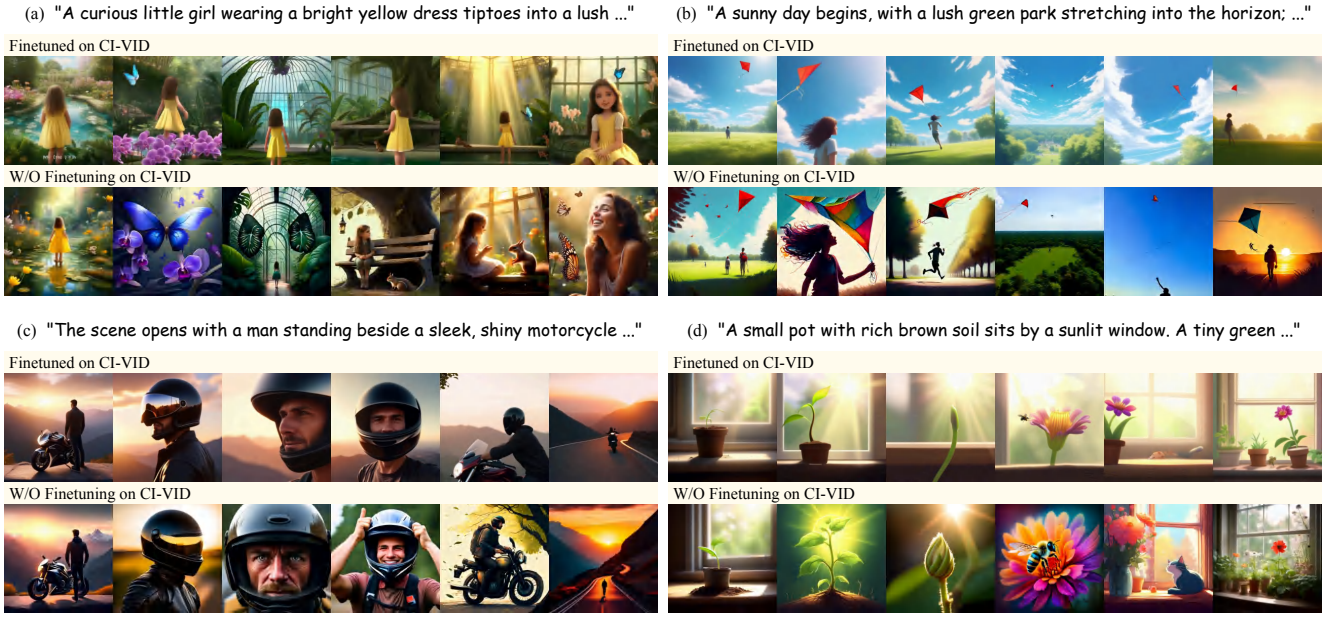


Figure 5. Comparison between generated results with and without finetuning on CI-VID, where (a) shows the result generated from the test prompt in Table 3.

base learning rate of 1×10^{-4} in all experiments. We initialize the model weights using the T2V model NOVA [12] to accelerate convergence. All experiments are conducted on NVIDIA A100 40GB GPUs

4.2. Test Prompt Generation

To support evaluation, we construct a set of 1,000 test prompts derived from seed keywords in VBench [16], with each keyword expanded into a multi-scene instruction. As illustrated in Table 3, each prompt contains six semantically connected scenes designed to form a coherent and engaging narrative, enabling a controlled assessment of both scene-level generation quality and multi-scene temporal reasoning capabilities.

4.3. Qualitative Experimental Results

A video generation model pretrained on the large-scale Emu3 dataset is used as the baseline and further fine-tuned on CI-VID. Figure 5 presents qualitative comparisons between outputs with and without CI-VID fine-tuning, based on the generated test prompts. For instance, sample (a) corresponds to the test prompt shown in Table 3.

As illustrated, fine-tuning on CI-VID significantly enhances the model's ability to generate coherent multi-clip videos. The fine-tuned model produces outputs with consistent style, color, texture, and layout, maintaining character identity and environmental coherence across clips. In contrast, the baseline model struggles to establish meaningful relationships between scenes and often fails to maintain temporal and semantic continuity. It may also introduce inconsistencies due to a lack of contextual awareness. Additionally, the fine-tuned model achieves high-quality camera transitions based on text prompts, resulting in video sequences with improved narrative flow and storytelling quality.

Table 4. Human evaluation results (Win/Tie/Loss) comparing the fine-tuned model against the baseline model.

Metric	Win	Tie	Loss
consistency	90.0%	6.5%	3.6%
narrativity	80.9%	15.0%	4.1%
correctness	78.3%	9.8%	11.9%

4.4. Quantitative Benchmark Results

4.4.1. Human Evaluation

We conduct a comprehensive human evaluation on all model outputs generated from the test prompts. Each output is presented as a row of merged keyframes, with one representative frame per scene, as illustrated in Figure 5. To ensure fairness, model identities are anonymized and their vertical order is randomized for every comparison.

Three full-time evaluators assess the outputs in a pairwise manner across three dimensions: *Consistency* (object, background, and visual style consistency across scenes), *Narrativity* (sequence coherence and storytelling quality), and *Factual Correctness* (alignment with the text prompt, visual accuracy, and absence of distortions). For each dimension, evaluators label the comparison as a win, tie, or

Table 5. VLM-based evaluation results. "Cons." refers to Consistency.

Model/ Dimension	Stylistic Cons.	Entity Cons.	Background Cons.	Perspective Transition	Prompt Alignment	Visual Plausibility
Baseline	2.93	2.84	2.80	3.02	3.99	3.25
+CI-VID	**3.83**	**3.73**	**3.75**	**3.81**	**4.07**	**3.62**

loss. Inter-rater agreement reaches 91% including ties and 97% excluding them, indicating strong scoring reliability. Aggregated results are reported in Table 4. The model fine-tuned on CI-VID achieves significantly more wins than the baseline across all three dimensions, demonstrating clear improvements in temporal consistency, narrative flow, and factual faithfulness.

4.4.2. VLM-based Evaluation

Following VBench [16], we employ VLMs to evaluate model outputs across six dimensions, as listed in Table 5. These dimensions cover both inter-clip coherence (first four dimensions) and standard video-generation qualities such as text alignment and visual plausibility. We use Qwen2.5-VL-72B-Instruct as the evaluator, which assigns a score from 0 to 5 (very poor to very excellent) based on the input prompt and the generated video frames.

To obtain reliable assessments, we evaluate each sample on the full generated sequence as well as its five adjacent clip pairs, enabling the VLM to judge both global and local temporal consistency. The final score for each dimension is computed as the average over these six evaluations (1 full video + 5 pairwise evaluations). To reduce evaluator drift, we also calibrate the VLM using a fixed reference example before all scoring.

As shown in Table 5, the model fine-tuned on CI-VID achieves substantial improvements on the first four coherence-oriented dimensions, demonstrating that CI-VID effectively enhances multi-clip temporal reasoning. It also attains slightly higher visual plausibility (dimension 6) and competitive performance on text–prompt alignment (dimension 5), indicating that our fine-tuned model improves coherence without sacrificing prompt faithfulness or visual quality.

4.4.3. Similarity-based Evaluation

We construct a similarity evaluation set of 1,103 samples from CI-VID. To avoid data leakage, all samples are selected from raw videos that contribute to only a single sample in the dataset. As illustrated in Figure 6, given an initial clip, the model generates continuation clips conditioned on CI-VID captions. We then evaluate both overall similarity—using the middle frame of each clip—and entity-level similarity to the ground-truth continuation.

For entity-level evaluation, YOLO-World-L [10] is used to detect objects in each frame, and the primary semantically relevant entity is manually annotated as ground truth. As shown in Figure 6, the ground-truth entity (blue box) corresponds to the cyclist in white. For each generated–ground-truth clip pair, we compute the similarity between the annotated entity and all detected candidates (green boxes) and take the highest value as the entity-level similarity for that clip. This procedure evaluates whether the model preserves the identity and appearance of the key entity across the generated continuation.

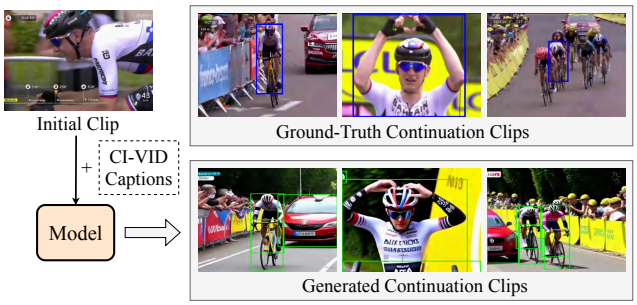


Figure 6. Example of similarity-based evaluation setup.

Table 6. Similarity evaluation results. Metrics include CLIP similarity, inverse LPIPS, and SSIM. Higher is better (↑).

Model/	Overall			Entity		
Metric	CLIP	1 - LPIPS	SSIM	CLIP	1 - LPIPS	SSIM
Baseline	0.512	0.309	0.199	0.601	0.360	0.278
+CI-VID	**0.670**	**0.381**	**0.272**	**0.702**	**0.412**	**0.391**

We adopt three complementary similarity metrics: (1) CLIP similarity [29], computed with ViT-H/14 pretrained on LAION-2B [31], to assess semantic consistency; (2) 1–LPIPS, to measure perceptual similarity; and (3) SSIM, to assess structural fidelity. As shown in Table 6, the CI-VID fine-tuned model consistently outperforms the baseline across all three metrics, at both the holistic and entity levels, demonstrating that CI-VID effectively enhances continuity and semantic preservation in video generation.

5. Conclusion

We present **CI-VID**, a dataset that moves beyond isolated text–video pairs toward coherent, interleaved text–video sequences. By modeling inter-clip relationships, CI-VID enables T&V2V generation and supports multi-clip video generation with semantic coherence. We further introduce a multi-dimensional benchmark for evaluating CI-VID from both human and automated perspectives. Experimental results underscore the value of CI-VID as a foundation for advancing research in controllable and coherent video generation.

Acknowledgement

We thank the support of the Youth Fund of the National Natural Science Foundation of China (62406040).

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022. 3

[2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *ArXiv preprint*, abs/2502.13923, 2025. 4

[3] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1728–1738, 2021. 2

[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, 2024. 1

[5] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 2

[6] Huiwen Chang, Han Zhang, Jarred Barber, AJ Maschinot, Jose Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T Freeman, Michael Rubinstein, et al. Muse: Text-to-image generation via masked generative transformers. *ArXiv preprint*, abs/2301.00704, 2023. 6

[7] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pages 370–387. Springer, 2024. 1, 2, 3

[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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13320–13331, 2024. 2, 3

[9] Wei Chen, Lin Li, Yongqi Yang, Bin Wen, Fan Yang, Tingting Gao, Yu Wu, and Long Chen. Comm: A coherent interleaved image-text dataset for multimodal understanding and generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8073–8082, 2025. 3

[10] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16901–16911, 2024. 8

[11] Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, et al. Paddleocr 3.0 technical report. *arXiv preprint arXiv:2507.05595*, 2025. 3

[12] Haoge Deng, Ting Pan, Haiwen Diao, Zhengxiong Luo, Yufeng Cui, Huchuan Lu, Shiguang Shan, Yonggang Qi, and Xinlong Wang. Autoregressive video generation without vector quantization. *ArXiv preprint*, abs/2412.14169, 2024. 6, 7

[13] Songwei Ge, Thomas Hayes, Harry Yang, Xi Yin, Guan Pang, David Jacobs, Jia-Bin Huang, and Devi Parikh. Long video generation with time-agnostic vqgan and time-sensitive transformer. In *European Conference on Computer Vision*, pages 102–118. Springer, 2022. 1

[14] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15180–15190, 2023. 4

[15] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36:72096–72109, 2023. 3

[16] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 7, 8

[17] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *ArXiv preprint*, abs/2410.21276, 2024. 4

[18] Mojan Javaheripi, Sébastien Bubeck, Marah Abdin, Jyoti Aneja, Sebastien Bubeck, Caio César Teodoro Mendes, Weizhu Chen, Allie Del Giorno, Ronen Eldan, Sivakanth Gopi, et al. Phi-2: The surprising power of small language models. *Microsoft Research Blog*, 1(3):3, 2023. 6

[19] Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. *Advances in Neural Information Processing Systems*, 37:48955–48970, 2025. 1, 2, 3

[20] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, et al. Videopoet: A large language model for zero-shot video generation. *ArXiv preprint*, abs/2312.14125, 2023. 1

[21] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, et al. Obelics: An open web-scale filtered dataset of interleaved image-text documents. *Advances in Neural Information Processing Systems*, 36:71683–71702, 2023. 3

[22] Tianhong Li, Yonglong Tian, He Li, Mingyang Deng, and Kaiming He. Autoregressive image generation without vec-

tor quantization. *Advances in Neural Information Processing Systems*, 37:56424–56445, 2025. 6

[23] 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*, abs/2412.00131, 2024. 6

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. 6

[25] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640, 2019. 2

[26] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *ArXiv preprint*, abs/2407.02371, 2024. 1, 2

[27] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International conference on machine learning*, pages 8162–8171. PMLR, 2021. 6

[28] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran, et al. Qwen2.5 technical report, 2025. 6

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 8

[30] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212, 2015. 2

[31] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 8

[32] Zhiyu Tan, Xiaomeng Yang, Luozheng Qin, and Hao Li. Vidgen-1m: A large-scale dataset for text-to-video generation. *arXiv preprint arXiv:2408.02629*, 2024. 2

[33] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 3

[34] Ye Tian, Ling Yang, Haotian Yang, Yuan Gao, Yufan Deng, Jingmin Chen, Xintao Wang, Zhaochen Yu, Xin Tao, Pengfei Wan, et al. Videotetris: Towards compositional text-to-video generation. *ArXiv preprint*, abs/2406.04277, 2024. 1

[35] Vikram Voleti, Alexia Jolicoeur-Martineau, and Chris Pal. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. *Advances in neural information processing systems*, 35:23371–23385, 2022. 1

[36] Qiuheng 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. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8428–8437, 2025. 1, 2

[37] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *ArXiv preprint*, abs/2409.18869, 2024. 1, 3

[38] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *ArXiv preprint*, abs/2307.06942, 2023. 1, 2, 3

[39] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5036–5045, 2022. 2

[40] Dongjie Yang, Suyuan Huang, Chengqiang Lu, Xiaodong Han, Haoxin Zhang, Yan Gao, Yao Hu, and Hai Zhao. Vript: A video is worth thousands of words. *ArXiv preprint*, abs/2406.06040, 2024. 1, 2, 3

[41] 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*, abs/2408.06072, 2024. 1

[42] Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *Advances in Neural Information Processing Systems*, 36:8958–8974, 2023. 3