

INTERPRETING ANY CONDITION TO CAPTION FOR CONTROLLABLE VIDEO GENERATION

Anonymous authors

Paper under double-blind review

ABSTRACT

To address the bottleneck of accurate user intent interpretation within current video generation community, we present **Any2Caption**, a novel framework for controllable video generation from any condition. The key idea is decoupling various condition interpretation steps from the video synthesis step. By leveraging modern multimodal large language models (MLLMs), **Any2Caption** interprets diverse inputs—text, images, videos, and specialized cues such as region, motion, and camera poses—into dense, structured captions that offer backbone video generators with better guidance. We also introduce **Any2CapIns**, a large-scale dataset with 337K instances and 407K conditions for any-condition-to-caption instruction tuning. Comprehensive evaluations demonstrate significant improvements in our system’s controllability and video quality compared to existing video generation models. Codes and resources will be released upon acceptance.

1 INTRODUCTION

Video serves as a fundamental medium for capturing real-world dynamics, making diverse and controllable video generation a key capability for modern artificial intelligence (AI) systems. Recently, video generation has gained significant attention, driven by advancements in Diffusion Transformers (DiT) Zheng et al. (2024b); Jiang et al. (2024); Singer et al. (2023); Ma et al. (2024a); kua (2024), which have demonstrated the ability to generate realistic, long-duration videos from text prompts. These advancements have even led to industrial applications, such as filmmaking. However, we observe that a major bottleneck in the current video generation community lies in **accurately interpreting user intention**, so as to produce high-quality, controllable videos.

In text-to-video (T2V) generation, studies Yang et al. (2024c); Ju et al. (2024); Hong et al. (2023) have suggested that detailed prompts, specifying objects, actions, attributes, poses, camera movements, and style, significantly enhance both controllability and video quality. Thus, a series of works have explored video recaption techniques (e.g., ShareGPT4Video Chen et al. (2024a), MiraData Ju et al. (2024), and InstanceCap Fan et al. (2024)) to build dense structural captions for optimizing generation models. While dense captions are used during training, in real-world inference scenarios, users most likely provide concise or straightforward input prompts Fan et al. (2024). Such a gap inevitably weakens instruction following and leads to suboptimal generation due to an incomplete understanding of user intent. To combat this, there are two possible solutions, manual prompt refinement, or automatic prompt enrichment Fan et al. (2024); Yang et al. (2024c) using large language models (LLMs). Yet, these approaches either require substantial human effort or risk introducing noise from incorrect prompt interpretations. As a result, this limitation in precisely interpreting user intent hinders the adoption of controllable video generation for demanding applications such as anime creation and filmmaking.

On the other hand, to achieve more fine-grained controllable video generation, one effective strategy is to provide additional visual conditions besides text prompts—such as reference images Wu et al. (2023); Guo et al. (2024a), identity Yuan et al. (2024); He et al. (2024b); Ma et al. (2024c), style Ye et al. (2024); Liu et al. (2024a), human pose Ma et al. (2024b); Karras et al. (2023), or camera He et al. (2024a); Zheng et al. (2024a)—or even combinations of multiple conditions together Zhao et al. (2023); Lin et al. (2024b); Wang et al. (2023). This multimodal conditioning approach aligns well with real-world scenarios, as users quite prefer interactive ways to articulate their creative intent. Several studies have examined video generation under various conditions, such

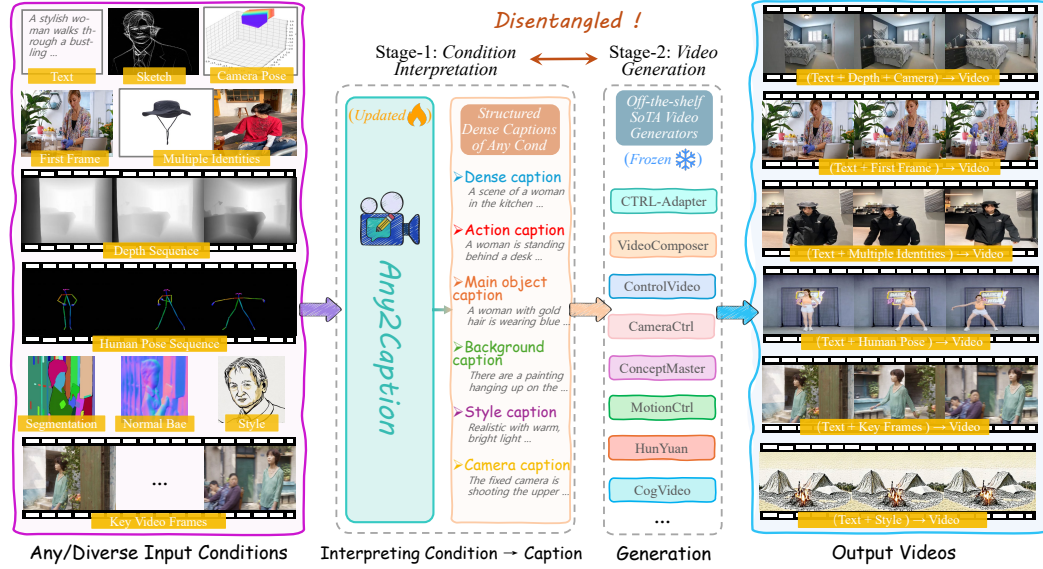


Figure 1: We propose Any2Caption, an efficient and versatile framework for interpreting diverse conditions to structured captions for highly controllable video generation.

as VideoComposer Wang et al. (2023), Ctrl-Adapter Lin et al. (2024b), and ControlVideo Zhao et al. (2023). Unfortunately, these methods tend to rely on the internal encoders of Diffusion/DiT to parse rich heterogeneous input conditions with intricate requirements (e.g., multiple object IDs and complex camera movements). Before generation, the model must accurately interpret the semantics of varied visual conditions in tandem with textual prompts. Yet even state-of-the-art (SoTA) DiT backbones have limited capacity for reasoning across different input modalities, resulting in suboptimal generation quality.

This work is dedicated to addressing these bottlenecks of any-conditioned video generation. Our core idea is to **decouple the first job of interpreting various conditions from the second job of video generation**, motivated by two important observations: **a)** SoTA video generation models (e.g., DiT) already excel at producing high-quality videos when presented with sufficiently rich text captions; **b)** Current MLLMs have demonstrated robust vision-language comprehension. Based on these, we propose **Any2Caption**, an MLLM-based universal condition interpreter designed not only to handle text, image, and video inputs but also equipped with specialized modules for motion and camera pose inputs. As illustrated in Fig. 1, Any2Caption takes as inputs any conditions (or combinations) and produces a densely structured caption, which is then passed on to any backend video generator for controllable, high-quality video production. As Any2Caption disentangles the role of complex interpretation of multimodal inputs from the generator, it advances in seamlessly integrating into a wide range of well-trained video generators without the extra cost of fine-tuning.

To facilitate the any-to-caption instruction tuning for Any2Caption, we construct **Any2CapIns**, a large-scale dataset that converts a concise user prompt and diverse non-text conditions into detailed, structured captions. Concretely, the dataset encompasses four main categories of conditions: depth maps, multiple identities, human poses, and camera motions. Through extensive manual labeling combined with automated annotation by GPT-4V Gpt (2023), followed by rigorous human verification, we curate a total of **337K** high-quality instances, with **407K** condition annotations, with the short prompts and structured captions averaging 55 and 231 words, respectively. Additionally, we develop a comprehensive evaluation strategy to thoroughly assess the model’s ability to interpret user intent under various conditions.

Experimentally, we first validate Any2Caption on our Any2CapIns, where results demonstrate that it achieves an impressive captioning quality that can faithfully reflect the original input conditions. We then experiment with integrating Any2Caption with multiple SoTA video generators, finding that (a) the long-form semantically rich prompts produced by Any2Caption are pivotal for generating high-quality videos under arbitrary conditions, and (b) Any2Caption consistently enhances performance across different backbone models, yielding noticeably improved outputs. Furthermore, Any2Caption demonstrates a pronounced advantage when handling mul-

multiple combined conditions, effectively interpreting and synthesizing intricate user constraints into captions that closely align with user expectations. Our contributions are threefold:

- We for the first time pioneer a novel *any-condition-to-caption* paradigm for video generation, which bridges the gap between user-provided multimodal conditions and structured video generation instructions, leading to highly controllable video generation.
- We propose `Any2Caption` to effectively integrate and comprehend diverse multimodal conditions, producing semantically enriched and structured captions, which consistently improve both condition flexibility and video quality. `Any2Caption` can also be widely integrated as a plug-in module for any existing video generator.
- We introduce `Any2CapIns`, a large-scale, high-quality dataset for any-condition-to-caption task, and establish a suite of evaluation metrics to rigorously assess the quality and fidelity of condition-based caption generation.

2 RELATED WORK

Controllable video generation Sun et al. (2024); Chen et al. (2024d); Fang et al. (2024); He et al. (2024a) has long been a central topic in AI. Recent advanced DiT methods, such as OpenAI’s Sora (2024) and HunyuanVideo Kong et al. (2024), yield photorealistic videos over extended durations. Early work focused on text-controlled video generation Singer et al. (2023); Hong et al. (2023), the prevalent approach. Yet, text prompts alone may insufficiently capture user intent, spurring exploration of additional inputs including static images Wu et al. (2023); Guo et al. (2024a), sketches Zhao et al. (2023); Wang et al. (2023), human poses Zhong et al. (2024); Ma et al. (2024b); Karras et al. (2023), camera views Zheng et al. (2024a); He et al. (2024a), and even extra videos Kara et al. (2024); Deng et al. (2024); Zhang et al. (2023). Thus, unifying these diverse conditions into an “any-condition” framework is highly valuable.

Recent works such as VideoComposer Wang et al. (2023), Ctrl-Adapter Lin et al. (2024b), and ControlVideo Zhao et al. (2023) have explored any-condition video generation. However, they face challenges in controlling multiple modalities due to the limited interpretability of text encoders in Diffusion or DiT. Motivated by existing MLLMs’ multimodal reasoning Liu et al. (2024b); Lin et al. (2024a); Wang et al. (2024a), we propose leveraging an MLLM to consolidate all possible conditions into structured dense captions for better controllable generation. SoTA DiT models already exhibit the capacity to interpret dense textual descriptions as long as the input captions are sufficiently detailed in depicting both the scene and the intended generation goals. Thus, our MLLM-based encoder alleviates the comprehension bottleneck, enabling the generation of higher-quality videos. To our knowledge, this is the first attempt in the field of any-condition video generation. Moreover, as the captioning stage is decoupled from backbone DiT, `Any2Caption` can integrate with existing video generation solutions without additional retraining.

Our approach also relates to video recaptioning, as our system generates dense captions under specific conditions. In text-to-video settings, prior work Fan et al. (2024); Yang et al. (2024c); Nan et al. (2024); Islam et al. (2024) shows that recaptioning yields detailed annotations that improve DiT training. For instance, ShareGPT4Video Chen et al. (2024a) uses GPT-4V Gpt (2023) to reinterpret video content, while MiraData Ju et al. (2024) and InstanceCap Fan et al. (2024) focus on structured and instance-consistent recaptioning. Unlike these methods, we avoid retraining powerful DiT models with dense captions by training an MLLM as an any-condition encoder on pairs of short, dense captions that are easier to obtain. Moreover, recaptioning entire videos can introduce noise or hallucinations that undermine DiT training, whereas our framework sidesteps this risk. Finally, while previous studies rely on dense-caption-trained DiT models, the real-world user concise prompts might create a mismatch that degrades generation quality.

3 ANY2CAPINS DATASET CONSTRUCTION

While relevant studies recapturing target videos for dense captions for enhanced T2V generation Chen et al. (2024a); Fan et al. (2024); Ju et al. (2024), these datasets suffer from two key limitations: 1) the absence of non-text conditions, and 2) short prompts that do not account for interactions among non-text conditions, potentially leading to discrepancies in real-world applications. To address these limitations, we introduce a new dataset, `Any2CapIns`, specifically designed to

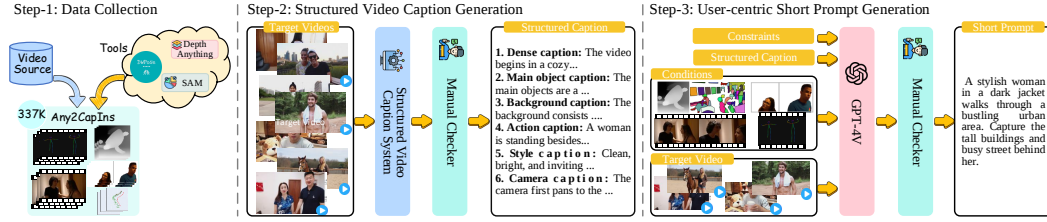


Figure 2: The pipeline for constructing the Any2CapIns dataset involves three key steps: 1) data collection, 2) structured video caption generation, and 3) user-centric short prompt generation.

incorporate diverse multimodal conditions for generating structured video captions. The dataset is constructed through a three-step process (cf. Fig. 2).

Step-1: Data Collection. We begin by systematically categorizing conditions into four primary types: 1) **Spatial-wise conditions**, focus on the structural and spatial properties of the video, e.g., depth maps, sketches, and video frames. 2) **Action-wise conditions**, emphasize motion and human dynamics in the target video, e.g., human pose motion. 3) **Composition-wise conditions**, focus on scene composition, particularly in terms of object interactions and multiple identities in the target video. 4) **Camera-wise conditions**, control video generation from a cinematographic perspective, e.g., movement trajectories. Since it is infeasible to encompass all possible conditions in dataset collection, we curate representative datasets Zhou et al. (2018); Chen et al. (2024c); Wang et al. (2024b); Ma et al. (2024b) under each type, specifically including *depth maps*, *human pose*, *multiple identities*, and *camera motion*. During the data collection process, we leverage SoTA tools to construct conditions. For instance, Depth Anything Yang et al. (2024b) is used to generate depth maps, DW-Pose Yang et al. (2023) provides human pose annotations, and Sam2 Ravi et al. (2024) is utilized for segmentation construction. In total, we collect **337K** video instances and **407K** conditions, with detailed statistics of the dataset presented in Tab. 1.

Category	#Num.	#Condition	#Avg. Len.	#Total Len.
Depth	182,945	182,945	9.87s	501.44h
Human Pose	44,644	44,644	8.38s	108.22h
Multi-Identities	68,255	138,089	13.01s	246.69h
Camera Movement	41,112	41,112	6.89s	78.86h

Table 1: Statistics of the collected dataset across four types of conditions. **#Num.** / **#Condition** means the number of instances / unique conditions. **#Avg.** / **#Total Len.** indicate the average and total video durations, respectively.

Step-2: Structured Video Caption Generation. The granularity of a caption, i.e., the specific elements it encompasses, plays a critical role in guiding the model to produce videos that closely align with desired descriptions while preserving coherence and realism. Drawing inspiration from Ju et al. (2024); Liu et al. (2024d), we design a structured caption format consisting of (1) Dense caption, (2) Main object caption, (3) Background caption, (4) Action caption, (5) Style caption, and (6) Camera caption. Following Wang et al. (2024b), we leverage GPT-4V and a fine-tuned LLaVA to generate the aspects of structured caption separately and merge them into a final structured caption.

Step-3: User-centric Short Prompt Generation. In this step, we construct short prompts from a user-centric perspective, considering how users naturally express their intentions. Firstly, our analysis highlights three key characteristics of users prompts: 1) **Conciseness and Simplicity**: Users favor brief and straightforward wording; 2) **Condition-Dependent Omission**: Users often omit textual descriptions of certain attributes (e.g., camera movement) when such conditions are already specified; and 3) **Implicit instruction of Target Video**: Users convey their intent indirectly (e.g., specifying multiple identities without detailing

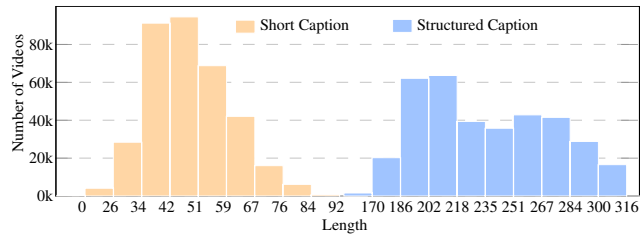


Figure 3: Distribution of the short/structured caption length (in words) in Any2CapIns.

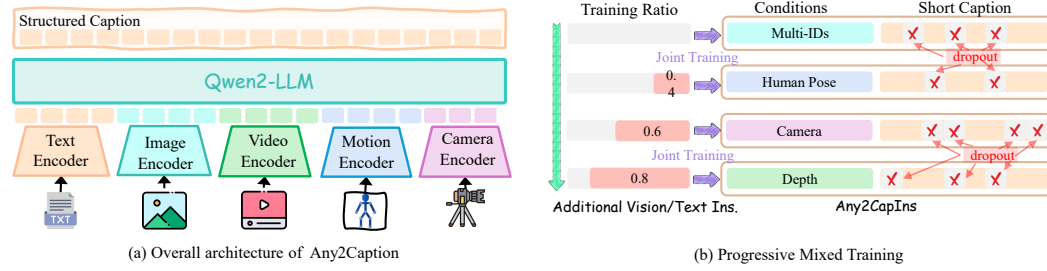


Figure 4: The Architecture of Any2Caption (a) uses Qwen2-LLM as the backbone, paired with text, image, video, motion, and camera encoders to generate structured captions. After alignment learning, a progressive mixed training strategy (b) is employed, progressively adding vision/text instruction datasets for joint training. For input short captions, a sentence-level random dropout mechanism is used to enhance robustness.

interactions). Guided by these observations, we tailor GPT-4V to precisely infer potential user prompts under condition-specific constraints. We also explicitly control the prompt length to maintain conciseness and refine the dataset through manual verification and filtering. Fig. 3 presents the length distribution of the resulting short and structured prompts.

4 ANY2CAPTION MODEL

In this section, we present Any2Caption, a novel MLLM designed to comprehensively interpret arbitrary multimodal conditions into controllable video captions, as illustrated in Fig. 4(a). Formally, given a short text prompt T along with non-text conditions $C = [c_1, \dots, c_n]$, where the non-text conditions can be either none, a single condition, or multiple conditions. The objective of this task is to generate a detailed and structured caption that serves as a control signal for video generation.

Architecture. Similar to existing MLLMs Liu et al. (2024b); Wu et al. (2024); Lin et al. (2024a), Any2Caption incorporates an image encoder F_I , a video encoder F_V , a motion encoder F_M and a camera encoder F_C to process non-text conditions. These encoders are then integrated into an LLM backbone F_{LLM} (i.e., Qwen2-LLM) to facilitate structured video captioning. Specifically, we leverage a ViT-based visual encoder from Qwen2-VL as F_I and F_V for the unified modeling of images and videos, achieving effective interpretation of input conditions represented in image or video formats, such as depth maps and multiple identities. To enable human pose understanding, we represent the extracted human pose trajectories as $H = \{(x_n^k, y_n^k) | k = 1, \dots, K, n = 1, \dots, N\}$, where N denotes the number of video frames and K is the number of keypoints. These trajectories are then visualized within video frames to enable further processing by the motion encoder, which shares the same architectural structure and initialization as the vision encoder. For camera motion understanding, inspired by He et al. (2024a), we introduce a camera encoder that processes a plücker embedding sequence $P \in \mathbb{R}^{N \times 6 \times H \times W}$, where H, W are the height and width of the video. This embedding accurately captures camera pose information, enabling precise modeling of camera trajectories. Finally, in line with Qwen2-VL, we employ special tokens to distinguish non-text conditions from texts. Besides the existing tokens, we introduce $\langle |motion_start| \rangle$, $\langle |motion_end| \rangle$, $\langle |camera_start| \rangle$, $\langle |camera_end| \rangle$, to demarcate the start and end of human&camera pose features.

Training Recipes. To accurately interpret user generation intent under arbitrary conditions and yield structured target video captions, large-scale pretraining and instruction tuning are required. To this end, we adopt a two-stage training procedure: **Stage-I: Alignment Learning.** In this stage, as image and video encoders have been well-trained in Qwen2-VL, we only focus on aligning human pose features from the motion encoder and camera movement features with the word embeddings of the LLM backbone. To achieve this, we freeze the LLM and vision encoder, while keeping the motion encoder trainable and optimize it on a human pose description task. Similarly, for camera movement alignment, we unfreeze the camera encoder and train it on a camera movement description task, ensuring that camera-related conditions are embedded into the model’s latent space. This alignment phase establishes a strong foundation for effective representation learning for these conditions. **Stage II: Condition-Interpreting Learning** Building upon the aligned encoders and

pretrained Qwen2-VL weights, we fine-tune the model on Any2CapIns for multimodal condition interpretation. However, direct fine-tuning leads to catastrophic forgetting due to the fixed output structure. To address this, we propose a progressive mixed training strategy. Specifically, the model is first trained on a single condition to establish a strong condition-specific understanding. As new conditions are introduced, we gradually incorporate vision-language instructions such as LLaVA-instruction Liu et al. (2024b) and Alpaca-52K Taori et al. (2023). This stepwise strategy ensures robust multimodal condition interpretation while preventing knowledge degradation.

5 EVALUATION SUITE

In this section, we introduce the evaluation suite for comprehensively assessing the capability of Any2Caption in interpreting user intent and generating structured captions.

Lexical Matching Score. We employ standard evaluation metrics commonly used in image/video captioning tasks, including BLEU Papineni et al. (2002), ROUGE Lin (2004), and METEOR Banerjee & Lavie (2005). We also introduce a Structural Integrity score to verify whether the generated captions adhere to the required six-component format, thereby ensuring completeness.

Semantic Matching Score. To evaluate the semantic alignment of generated captions, we employ BERTSCORE Zhang et al. (2020), which computes similarity by summing the cosine similarities between token embeddings, effectively capturing both lexical and compositional meaning preservation. Additionally, we utilize CLIP Score Hessel et al. (2021) to assess the semantic consistency between the input visual condition and the generated videos.

Intent Reasoning Score. Inspired by Chai et al. (2025), we introduce the Intent Reasoning Score (IRSCORE) to assess the fidelity of generated captions that capture user intents. The IRSCORE evaluation framework involves four steps: **(1) User Intention Extraction:** Categorize user intent into six aspects: subject, background, movement, camera, interaction, and style. **(2) Gold QA Pair Construction:** Create aspect-specific QA pairs with defined requirements (e.g., object count, appearance). **(3) Answer Prediction:** Prompt GPT-4V to answer the questions solely based on the predicted caption. **(4) Answer Evaluation:** GPT-4V scores each answer for correctness and quality, averaged across all QA pairs. More details are in Appendix §G.

Video Generation Quality Score. We also employ several metrics to evaluate the quality of videos generated from structured captions. Following Huang et al. (2024); Ju et al. (2024), we evaluate video generation quality across four key dimensions: motion smoothness, dynamic degree, aesthetic quality, and image integrity. To further verify adherence to specific non-text conditions, we use specialized metrics: RotErr, TransErr, and CamMC He et al. (2024a) for camera motion accuracy; Mean Absolute Error (MAE) for depth consistency Guo et al. (2024b); DINO-I Ruiz et al. (2023), CLIP-I Ruiz et al. (2023) Score to evaluate identity preservation under multiple identities, and Pose Accuracy (Pose Acc.) Ma et al. (2024b) to assess the alignment in the generated videos.

6 EXPERIMENTS

6.1 SETUPS

Dataset. We manually construct 200 test cases for each type of condition (i.e., *depth*, *human pose*, *multiple identities*, *camera*, and *compositional conditions*) to evaluate the model’s performance. Additionally, we assess the model on publicly available benchmarks (e.g., Ye et al. (2024); Lin et al. (2024b)). For further details, please refer to the Appendix §H.

Implementation Details. We leverage Qwen2VL-7B Wang et al. (2024a) as the backbone of our model, which supports both image and video understanding. The human pose in the input conditions is encoded and processed in the video format. The camera encoder adopts the vision encoder architecture, with the following settings: in channels set to 96, patch size of 16, depth of 8, and 8 attention heads. During training, to simulate the brevity and randomness of user inputs, we randomly drop sentences from the short caption with a dropout rate of 0.6; a similar dropout strategy is applied to non-textual conditions. We conducted the training on 8×A800 GPUs. For further details on the training parameters for each stage, please refer to the Appendix §H.

Category	Structural Integrity	Lexical Matching			Semantic Matching		Intent Reasoning	
		B-2	R-L	METER	BERTSCORE	Accuracy	Quality	
Entire Structured Caption	91.25	54.99	48.63	52.47	91.95	68.15	3.43	
Dense Caption	-	44.24	42.89	49.51	92.42	78.47	3.47	
Main Object Caption	-	38.54	47.46	52.48	92.02	56.28	2.74	
Background Caption	-	44.65	46.73	48.87	92.90	69.37	2.69	
Action Caption	-	31.91	39.83	45.25	91.44	57.98	2.13	
Style Caption	-	41.71	47.70	55.9	93.48	63.75	3.05	
Camera Caption	-	60.21	96.10	94.32	99.31	66.31	3.75	

Table 4: Quantitative results of structured caption generation quality under four aspects: *structural integrity*, *lexical matching*, *semantic matching*, and *intent reasoning*. We demonstrate the overall caption generation capability and the performance of individual components within the structure. “B-2” and “R-L” denotes BLEU-2 and ROUGE-L, respectively.

6.2 EXPERIMENTAL RESULTS AND ANALYSES

In this section, we present the experimental results and provide in-depth analyses to answer the following six key research questions, revealing how the system advances.

RQ-1: Is the structured caption necessary?

We compare our structured caption approach with a simpler method, where we first caption the input condition (e.g., multiple identity images) and then concatenate that caption with the original short prompt, as shown in Tab. 2. Our results indicate that merely appending the condition’s caption to the short prompt can reduce video smoothness and image quality. One likely reason is that the identity images may contain extraneous details beyond the target subject, potentially conflicting with the original prompt and causing inconsistencies. Consequently, controllability in the final output is compromised. In contrast, our structured caption method accurately identifies the target subject and augments the prompt with relevant information, yielding more controllable video generation (cf. §I.3).

Prompt Enrichment	Text		Video Generation	
	DINO-I↑	Smoo.↑	Aest.↑	Inte.↑
Multi-IDs+Prompt.	43.72	93.46	5.32	55.39
Multi-IDs+Prompt+IDs Cap.	43.96	93.41	5.41	54.91
Multi-IDs+Structured Cap.	45.77	94.38	5.46	57.47
Only Structured Cap.	38.45	94.43	5.19	56.09

Table 2: Quantitative results comparing multi-identity video generation using different prompt enrichment methods.

RQ-2: How effective is the training strategy?

Next, we investigate the contribution of the training mechanism, and the results are shown in Tab. 3. During training, we employ a two-stage training approach, consisting of alignment learning followed by instruction-tuning. When alignment learning is omitted, and the model proceeds directly to instruction tuning, both captioning and video generation performance degrade significantly. A possible explanation is that bypassing alignment learning disrupts the encoder’s adaptation process which has been aligned to the LLM backbone, leading to suboptimal results in subsequent stages. Additionally, we compare the performance of the model without the dropout mechanism. Although removing dropout yields a marked improvement in captioning quality, the benefit to video generation is marginal. This suggests that without dropout, the model may rely on shortcuts from the input captions rather than fully understanding the underlying intent, thereby increasing the risk of overfitting.

Training Strategy	Caption		Vieo Generation		
	B-2↑	Accuracy↑	Smoo.↑	Dyna.↑	Aest.↑
Any2Caption	47.69	67.35	94.60	17.67	5.53
w/o Two-Stage	33.70	51.79	93.31	16.36	5.10
w/o Dropout	49.24	69.51	94.16	14.54	5.51

Table 3: Ablation study on training strategy. “w/o Two-stage” means no alignment learning, and “w/o Dropout” denotes short captions are not randomly dropped.

RQ-3: How well is the structured caption generation quality?

We first evaluate whether our proposed model could accurately interpret user intent and generate high-quality structured captions. From a caption-generation perspective, we compare the predicted captions with gold-standard captions across various metrics (see Tab. 4). We observe that our model successfully produces the desired structured content, achieving 91.25% in structural integrity. Moreover, it effectively captures

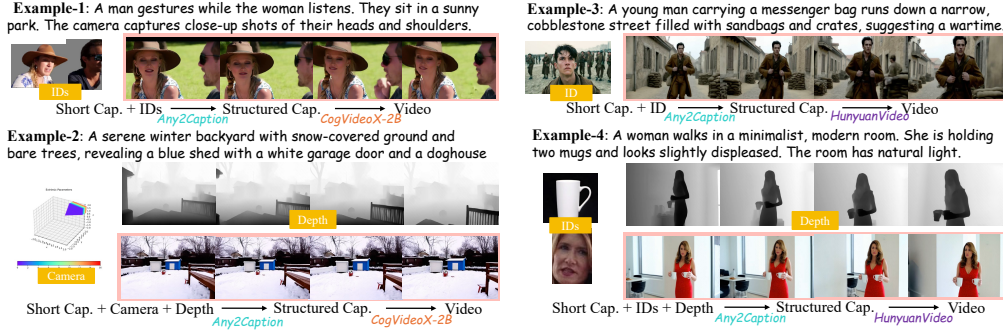


Figure 5: Illustrations of generated videos where only the structured captions yielded by *Any2Caption* are fed into the *CogVideoX-2B* (Left), and *HunyuanVideo* (Right). We can observe that some key features of the input identity images, such as the background and main object, can be accurately visualized in the generated videos.

Model	Text	Camera			Identities		Depth	Human Pose	Overall Quality		
	CLIP-T↑	RotErr↓	TransErr↓	CamMC↓	DINO-I↑	CLIP-I↑	MAE↓	Pose Acc.↑	Smoo.↑	Dyna.↑	Inte.↑
● Camera to Video											
MotionCtrl	19.67	1.54	4.49	4.80	-	-	-	-	96.13	9.75	73.69
+ Structured Cap.	20.16	1.45	4.37	4.78	-	-	-	-	96.16	11.43	74.63
CameraCtrl	18.89	1.37	3.51	4.78	-	-	-	-	94.11	12.59	71.84
+ Structured Cap.	21.70	0.94	2.97	4.37	-	-	-	-	95.16	13.72	72.47
● Depth to Video											
Ctrl-Adapter	20.37	-	-	-	-	-	25.63	-	94.53	20.73	46.98
+ Structured Cap.	23.30	-	-	-	-	-	21.87	-	95.54	15.14	54.20
ControlVideo	22.17	-	-	-	-	-	30.11	-	92.88	5.94	63.85
+ Structured Cap.	24.18	-	-	-	-	-	23.92	-	94.47	18.27	66.28
● Identities to Video											
ConceptMaster	16.04	-	-	-	36.37	65.31	-	-	94.71	8.18	43.68
+ Structured Cap.	17.15	-	-	-	39.42	66.74	-	-	95.05	10.14	49.73
● Human Pose to Video											
FollowYourPose	21.11	-	-	-	-	-	-	30.47	91.71	14.29	58.84
+ Structured Cap.	21.39	-	-	-	-	-	-	31.59	92.87	16.47	56.30

Table 5: Performance on *camera*, *depth*, *identities*, and *human pose* conditions for using short vs. structured captions, evaluated using various video quality metrics. Better results are in **bold**.

the key elements of the gold captions, attaining a ROUGE-L score of 48.63 and a BERTSCORE of 91.95. Notably, the model demonstrates the strongest performance in interpreting camera-related details compared to other aspects. Finally, regarding user intent analysis, we found that the model reliably incorporated user preferences into its structured outputs.

To further showcase the model’s capacity to understand and leverage input conditions, we directly feed the structured captions—derived from our model’s interpretation—into downstream text-to-video generation systems (e.g., CogvideoX Yang et al. (2024c) and Hunyuan Kong et al. (2024)), as illustrated in Fig. 5. Even without explicit visual conditions (e.g., identities), the resulting videos align well with the input prompts, such as hair color and clothing style in Example 1, indicating that our captions successfully capture intricate visual details. In particular, the model is able to accurately grasp dense conditions, such as depth sequences, or compositional requirements in Example 4, ultimately enabling controllable video generation. Although some subtle attributes may be under-specified in text, occasionally leading to mismatches, overall controllability remains robust.

RQ-4: Is the video generation quality enhanced with a structured caption? Here, we investigate whether integrating structured captions consistently improves controllable video generation in multiple methods. We investigate the impact of incorporating camera, depth, identities, and human pose conditions into various controllable video generation methods. As shown in Table 5, all tested models exhibit consistent gains in overall video quality, including smoothness and frame fidelity, after incorporating structured captions, without requiring any changes to the model architectures or additional training. Moreover, these models show enhanced adherence to the specified conditions, suggesting that our generated captions precisely capture user requirements and lead to more accurate, visually coherent video outputs. Additional examples can be found in Appendix §I.3.

Conditions	Text	Camera			Identities		Depth	Overall Quality			
	CLIP-T $\uparrow$	RotErr $\downarrow$	TransErr $\downarrow$	CamMC $\downarrow$	DINO-I $\uparrow$	CLIP-I $\uparrow$	MAE $\downarrow$	Smoo $\uparrow$	Dyna $\uparrow$	Aest $\uparrow$	Inte $\uparrow$
C+I	14.81	1.37	4.04	4.24	25.63	64.14	-	94.43	28.87	4.99	59.81
+ Structured Cap.	19.03	1.30	4.36	4.03	26.75	68.45	-	94.38	34.99	5.25	63.02
C+D	20.80	1.57	3.88	4.77	-	-	32.15	95.36	30.12	4.82	63.90
+ Structured Cap.	21.19	1.49	4.41	4.84	-	-	25.37	95.40	30.10	4.96	65.05
D+I	20.01	-	-	-	35.24	57.82	23.00	93.15	32.21	4.96	61.21
+ Structured Cap.	20.76	-	-	-	36.25	63.48	24.78	92.50	36.43	5.18	60.81
C+I+D	18.49	2.05	7.74	8.47	35.86	64.25	18.37	92.02	30.09	3.91	60.62
+ Structured Cap.	19.52	1.57	7.74	8.20	38.74	64.37	17.41	93.03	32.81	4.99	61.22

Table 6: Quantitative comparison across compositional conditions on FullDiT Ju et al. (2025). C, D, and I denote camera, depth and multiple identities conditions, respectively

RQ-5: How well does the model perform on compositional conditions? We examine the impact of structured captions under compositional conditions. As shown in Tab. 6, we compare the combined camera, identities, and depth on our customized model and observe that structured captions consistently enhance its performance. Moreover, from Examples 2 and 4 in Fig. 5, our model demonstrates a thorough understanding of the interactions among various conditions, for instance, capturing a woman’s hair color and the position of a mug, accurately guiding the production of videos that align with the specified requirements. This finding further highlights that our approach can automatically equip existing T2V models with the ability to handle compositional conditions without requiring additional training.

RQ-6: How well is the generalization capability of Any2Caption?

Finally, we investigate the model’s generalization ability by evaluating its performance on “unseen” conditions, including *style*, *segmentation*, *sketch*, and *masked images*. As demonstrated in Fig. 6, the structured captions generated by our model consistently enhance existing T2V frameworks, offering benefits such as increased motion smoothness, aesthetics quality, and more accurate generation control. We attribute these advantages to two primary factors: the strong reasoning capabilities of our MLLM backbone and our training strategy, i.e., progressive mixed training, which leverages existing vision and text instructions for fine-tuning while mitigating knowledge forgetting, thereby ensuring robust generalization.

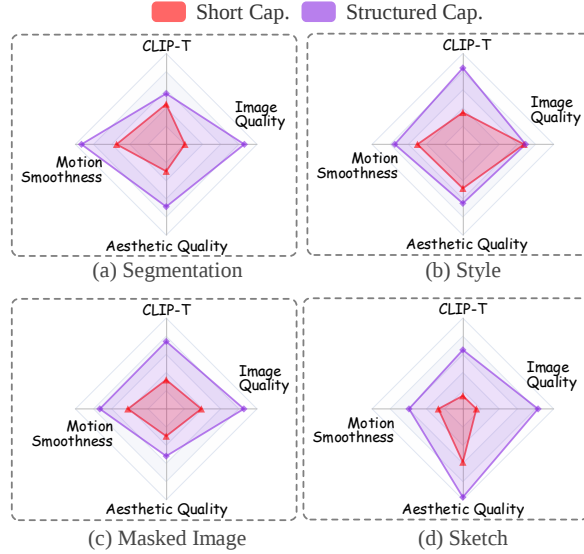


Figure 6: Quantitative results on unseen conditions (i.e., *segmentation* Wang et al. (2023), *style* Ye et al. (2024), *masked image* Wang et al. (2023), and *sketch* Wang et al. (2023)) when using short and structured captions, respectively.

7 CONCLUSION

In this work, we focus on addressing the challenge of accurately interpreting user generation intentions under various conditions for controllable video generation. We introduce Any2Caption, a framework that decouples the interpretation of multimodal conditions from video synthesis. Built based on an MLLM, Any2Caption converts diverse inputs into dense captions that drive high-quality video generation. We further present Any2CapIns, a large-scale dataset for effective instruction tuning. Experiments show that our method improves controllability and video quality across various backbones.

REFERENCES

- Gpt-4v(ision) system card, 2023. URL https://cdn.openai.com/papers/GPTV_System_Card.pdf.
- Kuaishou. kling, 2024. URL <https://kling.kuaishou.com>.
- Sora, 2024. URL <https://openai.com/sora/>.
- Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL*, pp. 65–72, 2005.
- Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *CoRR*, abs/2311.15127, 2023a.
- 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*, pp. 22563–22575, 2023b.
- Tim Brooks, Janne Hellsten, Miika Aittala, Ting-Chun Wang, Timo Aila, Jaakko Lehtinen, Ming-Yu Liu, Alexei A. Efros, and Tero Karras. Generating long videos of dynamic scenes. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *NeurIPS*, 2022.
- Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jenq-Neng Hwang, Saining Xie, and Christopher D. Manning. Auroracap: Efficient, performant video detailed captioning and a new benchmark. In *ICLR*, 2025.
- Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Lin Bin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, and Jiaqi Wang. Sharegpt4video: Improving video understanding and generation with better captions. In *NeurIPS*, 2024a.
- Ling-Hao Chen, Shunlin Lu, Ailing Zeng, Hao Zhang, Benyou Wang, Ruimao Zhang, and Lei Zhang. Motionllm: Understanding human behaviors from human motions and videos. *CoRR*, abs/2405.20340, 2024b.
- 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, and Sergey Tulyakov. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *CVPR*, pp. 13320–13331, 2024c.
- Yiran Chen, Anyi Rao, Xuekun Jiang, Shishi Xiao, Ruiqing Ma, Zeyu Wang, Hui Xiong, and Bo Dai. Cinepregen: Camera controllable video previsualization via engine-powered diffusion. *CoRR*, abs/2408.17424, 2024d.
- Mengyu Chu, You Xie, Jonas Mayer, Laura Leal-Taixé, and Nils Thuerey. Learning temporal coherence via self-supervision for gan-based video generation. *ACM Trans. Graph.*, 39(4):75, 2020.
- Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-audio technical report. *CoRR*, abs/2407.10759, 2024.
- Yufan Deng, Ruida Wang, Yuhao Zhang, Yu-Wing Tai, and Chi-Keung Tang. Dragvideo: Interactive drag-style video editing. In *ECCV*, pp. 183–199, 2024.
- Tiehan Fan, Kepan Nan, Rui Xie, Penghao Zhou, Zhenheng Yang, Chaoyou Fu, Xiang Li, Jian Yang, and Ying Tai. Instancecap: Improving text-to-video generation via instance-aware structured caption. *CoRR*, abs/2412.09283, 2024.
- Haopeng Fang, Di Qiu, Binjie Mao, Pengfei Yan, and He Tang. Motioncharacter: Identity-preserving and motion controllable human video generation. *CoRR*, abs/2411.18281, 2024.

- Xun Guo, Mingwu Zheng, Liang Hou, Yuan Gao, Yufan Deng, Pengfei Wan, Di Zhang, Yufan Liu, Weiming Hu, Zhengjun Zha, Haibin Huang, and Chongyang Ma. I2v-adapter: A general image-to-video adapter for diffusion models. In *ACM SIGGRAPH*, pp. 112, 2024a.
- Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models. In *ECCV*, pp. 330–348, 2024b.
- Shir Gur, Sagie Benaim, and Lior Wolf. Hierarchical patch VAE-GAN: generating diverse videos from a single sample. In *NeurIPS*, 2020.
- Niv Haim, Ben Feinstein, Niv Granot, Assaf Shocher, Shai Bagon, Tali Dekel, and Michal Irani. Diverse generation from a single video made possible. In *ECCV*, pp. 491–509, 2022.
- Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *CoRR*, abs/2404.02101, 2024a.
- Xuanhua He, Quande Liu, Shengju Qian, Xin Wang, Tao Hu, Ke Cao, Keyu Yan, Man Zhou, and Jie Zhang. Id-animator: Zero-shot identity-preserving human video generation. *CoRR*, abs/2404.15275, 2024b.
- Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP*, pp. 7514–7528, 2021.
- Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pre-training for text-to-video generation via transformers. In *ICLR*, 2023.
- Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video generative models. In *CVPR*, pp. 21807–21818, 2024.
- Md Mohaiminul Islam, Ngan Ho, Xitong Yang, Tushar Nagarajan, Lorenzo Torresani, and Gedas Bertasius. Video recap: Recursive captioning of hour-long videos. In *CVPR*, pp. 18198–18208, 2024.
- Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. In *NeurIPS*, 2023.
- Junpeng Jiang, Gangyi Hong, Lijun Zhou, Enhui Ma, Hengtong Hu, Xia Zhou, Jie Xiang, Fan Liu, Kaicheng Yu, Haiyang Sun, Kun Zhan, Peng Jia, and Miao Zhang. Dive: Dit-based video generation with enhanced control. *CoRR*, abs/2409.01595, 2024.
- 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. In *NeurIPS*, 2024.
- Xuan Ju, Weicai Ye, Quande Liu, Qiulin Wang, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, and Qiang Xu. Fulldit: Multi-task video generative foundation model with full attention. *CoRR*, abs/2503.19907, 2025.
- Ozgur Kara, Bariscan Kurtkaya, Hidir Yesiltepe, James M. Rehg, and Pinar Yanardag. RAVE: randomized noise shuffling for fast and consistent video editing with diffusion models. In *CVPR*, pp. 6507–6516, 2024.
- Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dream-pose: Fashion image-to-video synthesis via stable diffusion. In *ICCV*, pp. 22623–22633, 2023.
- Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojuan Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchu Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhen-tao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Daquan Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar

- Zhong. Hunyuanvideo: A systematic framework for large video generative models. *CoRR*, abs/2412.03603, 2024.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *Trans. Mach. Learn. Res.*, 2025, 2025.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICM*, pp. 19730–19742, 2023.
- Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *ECCV*, pp. 323–340, 2024a.
- Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *CoRR*, abs/2403.18814, 2024b.
- Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *EMNLP*, pp. 5971–5984, 2024a.
- Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, 2004.
- Han Lin, Jaemin Cho, Abhay Zala, and Mohit Bansal. Ctrl-adapter: An efficient and versatile framework for adapting diverse controls to any diffusion model. *CoRR*, abs/2404.09967, 2024b.
- Gongye Liu, Menghan Xia, Yong Zhang, Haoxin Chen, Jinbo Xing, Yibo Wang, Xintao Wang, Ying Shan, and Yujiu Yang. Stylecrafter: Taming artistic video diffusion with reference-augmented adapter learning. *ACM Trans. Graph.*, 43(6):251:1–251:10, 2024a.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pp. 26286–26296, 2024b.
- Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024c. URL <https://llava-vl.github.io/blog/2024-01-30-llava-next/>.
- Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. In *CVPR*, pp. 22139–22149, 2024d.
- Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *CoRR*, abs/2401.03048, 2024a.
- Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *AAAI*, pp. 4117–4125, 2024b.
- Ze Ma, Daquan Zhou, Chun-Hsiao Yeh, Xue-She Wang, Xiuyu Li, Huanrui Yang, Zhen Dong, Kurt Keutzer, and Jiashi Feng. Magic-me: Identity-specific video customized diffusion. *CoRR*, abs/2402.09368, 2024c.
- Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *ACL*, pp. 12585–12602, 2024.
- 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. *CoRR*, abs/2407.02371, 2024.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. Bleu: a method for automatic evaluation of machine translation. pp. 311–318, 2002.

- Zhiwu Qing, Shiwei Zhang, Jiayu Wang, Xiang Wang, Yujie Wei, Yingya Zhang, Changxin Gao, and Nong Sang. Hierarchical spatio-temporal decoupling for text-to- video generation. In *CVPR*, pp. 6635–6645, 2024.
- Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloé Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross B. Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. *CoRR*, abs/2408.00714, 2024.
- Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, pp. 22500–22510, 2023.
- Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. In *ICLR*, 2023.
- Wenqiang Sun, Shuo Chen, Fangfu Liu, Zilong Chen, Yueqi Duan, Jun Zhang, and Yikai Wang. Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion. *CoRR*, abs/2411.04928, 2024.
- Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023.
- Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Iyer, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, Xichen Pan, Rob Fergus, Yann LeCun, and Saining Xie. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *NeurIPS*, 2024.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *CoRR*, abs/2409.12191, 2024a.
- Qiuhe Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, Fei Yang, Pengfei Wan, and Di Zhang. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. *CoRR*, abs/2410.08260, 2024b.
- Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. In *NeurIPS*, 2023.
- Yaohui Wang, Piotr Bilinski, François Brémond, and Antitza Dantcheva. Imaginator: Conditional spatio-temporal GAN for video generation. In *WACV*, pp. 1149–1158, 2020.
- Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *ICCV*, pp. 7589–7599, 2023.
- Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal LLM. In *ICML*. OpenReview.net, 2024.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yeqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report. *CoRR*, abs/2407.10671, 2024a.

- Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything V2. In *NeurIPS*, 2024b.
- Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *ICCV*, pp. 4212–4222, 2023.
- Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Xiaotao Gu, Yuxuan Zhang, Weihang Wang, Yean Cheng, Ting Liu, Bin Xu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer, 2024c.
- Zixuan Ye, Huijuan Huang, Xintao Wang, Pengfei Wan, Di Zhang, and Wenhan Luo. Stylemaster: Stylize your video with artistic generation and translation. *CoRR*, abs/2412.07744, 2024.
- Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyuan Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-preserving text-to-video generation by frequency decomposition. *CoRR*, abs/2411.17440, 2024.
- Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *CoRR*, abs/2406.16852, 2024a.
- Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *CoRR*, abs/2311.04145, 2023.
- Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Chen Change Loy, and Shuicheng Yan. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. In *NeurIPS*, 2024b.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with BERT. In *ICLR*, 2020.
- Yiming Zhang, Zhening Xing, Yanhong Zeng, Youqing Fang, and Kai Chen. PIA: your personalized image animator via plug-and-play modules in text-to-image models. In *CVPR*, pp. 7747–7756, 2024c.
- Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. Llava-next: A strong zero-shot video understanding model, April 2024d. URL <https://llava-vl.github.io/blog/2024-04-30-llava-next-video/>.
- Min Zhao, Rongzhen Wang, Fan Bao, Chongxuan Li, and Jun Zhu. Controlvideo: Adding conditional control for one shot text-to-video editing. *CoRR*, abs/2305.17098, 2023.
- Guangcong Zheng, Teng Li, Rui Jiang, Yehao Lu, Tao Wu, and Xi Li. Cami2v: Camera-controlled image-to-video diffusion model. *CoRR*, abs/2410.15957, 2024a.
- Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *CoRR*, abs/2412.20404, 2024b.
- Yong Zhong, Min Zhao, Zebin You, Xiaofeng Yu, Changwang Zhang, and Chongxuan Li. Posecrafter: One-shot personalized video synthesis following flexible pose control. In *ECCV*, pp. 243–260, 2024.
- Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM Trans. Graph.*, 37(4):65, 2018.

APPENDIX OVERVIEW

The appendix presents more details and additional results not included in the main paper due to page limitations. The list of items included is:

- Clarification on the Usage of Large Language Models in Section §A.
- Limitation in Section §B.
- Ethic Statement results in Section §C.
- Reproducibility Statement in Section §D.
- Extended Related Work in Section §E.
- Extended Dataset Construction Details in Section §F.
- More Statistics Information of IRSCORE in Section §G.
- Detailed Setups in Section §H.
- Extended Experiment Results and Analyses in Section §I.

A THE USAGE OF LARGE LANGUAGE MODELS (LLMs)

In this paper, we employed GPT-4o for data construction and, separately, leveraged large language models (LLMs) as auxiliary tools to improve the clarity and readability of the manuscript. Specifically, LLMs were utilized to refine sentence structure, correct grammatical errors, and enhance the overall presentation of the draft. All technical content, including research ideas, algorithm design, experimental setup, analysis, and conclusions, was entirely conceived, implemented, and validated by the authors without reliance on LLMs.

B LIMITATION

Despite the advancement of our proposed framework, several limitations may remain:

Firstly, the diversity of annotated data is constrained by the capabilities of the current annotation tools, which may limit the variety of generated content. Moreover, the scarcity of real-world data introduces potential domain gaps, reducing the model’s generalizability in practical scenarios.

Secondly, due to inherent model limitations, hallucinations may occur, resulting in inaccurate structured captions and consequently degrading the quality of generated videos. A possible direction to mitigate this issue is to develop an end-to-end approach that jointly interprets complex conditions and handles video generation.

Lastly, the additional condition-understanding modules inevitably increase inference time. However, our empirical results suggest that the performance gains from these modules are substantial, and future work may explore more efficient architectures or optimization techniques to balance speed and accuracy.

C ETHIC STATEMENT

This work relies on publicly available datasets and manually constructed datasets, ensuring that all data collection and usage adhere to established privacy standards. We recognize that automatic annotation processes may introduce biases, and we have taken measures to evaluate and mitigate these biases. Nonetheless, we remain committed to ongoing improvements in this area.

By enhancing video generation capabilities, Any2Caption could inadvertently facilitate negative societal impacts, such as the production of deepfakes and misinformation, breaches of privacy, or the creation of harmful content. We, therefore, emphasize the importance of strict ethical guidelines, robust privacy safeguards, and careful dataset curation to minimize these risks and promote responsible research practices.

D REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our work, we have made a concerted effort to provide all necessary details and materials. We provide comprehensive details of the proposed Any2Caption framework, including its definition and input–output formulation (Section §4). The model backbone and training methodology are described in detail in Section §4 and Appendix §H.3. We further report all hyperparameter settings and training configurations in Section §6.1 and Appendix §H, using fixed random seeds to ensure the replicability of the experiments. All datasets used in this study are publicly available open-source resources, and the data construction process, along with the amount of data used at each training stage, is thoroughly documented in Section §3 and Appendix §F. Finally, we will release the full codebase and data processing scripts to the community upon acceptance.

E EXTENDED RELATED WORK

E.1 TEXT-TO-VIDEO GENERATION

The development of video generation models has progressed from early GAN- and VAE-based approaches Brooks et al. (2022); Wang et al. (2020); Haim et al. (2022); Chu et al. (2020); Gur et al. (2020) to the increasingly popular diffusion-based methods Blattmann et al. (2023b;a); Zhang et al. (2024c); Qing et al. (2024). Among these, diffusion-in-transformer (DiT) architectures, such as OpenAI’s Sora (2024) and HunyuanVideo (Kong et al. (2024)), have demonstrated remarkable performance, producing photorealistic videos over extended durations. Controllable video generation (Sun et al. (2024); Chen et al. (2024d); Fang et al. (2024); He et al. (2024a)) has become an essential aspect of this field. Initially, research efforts centered predominantly on text-to-video generation (Singer et al. (2023); Hong et al. (2023)), which remains the most common approach. However, relying solely on text prompts can be insufficient for accurately capturing user intent, prompting exploration into other conditioning inputs such as static images (Wu et al. (2023); Guo et al. (2024a)), user sketches (Zhao et al. (2023); Wang et al. (2023)), human poses (Zhong et al. (2024); Ma et al. (2024b); Karras et al. (2023)), camera perspectives (Zheng et al. (2024a); He et al. (2024a)), and even additional videos (Kara et al. (2024); Deng et al. (2024); Zhang et al. (2023)). Given this diversity of potential conditions, unifying them into a single “any-condition” video generation framework is highly valuable.

E.2 CONTROLLABLE VIDEO GENERATION

Recent methods like VideoComposer (Wang et al. (2023)), Ctrl-Adapter (Lin et al. (2024b)), and ControlVideo (Zhao et al. (2023)) have investigated any-condition video generation. Nevertheless, they still struggle with comprehensive controllability due to the complexity of multiple modalities and the limited capacity of standard diffusion or DiT encoders to interpret them. Inspired by the strong multimodal reasoning capabilities of modern MLLMs (Liu et al. (2024b); Lin et al. (2024a); Wang et al. (2024a)), we propose leveraging an MLLM to consolidate all possible conditions into a unified reasoning process, producing structured dense captions as inputs to a backbone Video DiT. SoTA DiT models already exhibit the capacity to interpret dense textual descriptions, as long as the input captions are sufficiently detailed in depicting both the scene and the intended generation goals. Building on this, our MLLM-based condition encoder directly addresses the comprehension bottleneck, theoretically enabling higher-quality video generation. To our knowledge, this work is the first to develop an MLLM specifically tailored for any-condition video generation. Because the caption-generation mechanism is decoupled from DiT, our proposed Any2Caption can be integrated into existing DiT-based methods without additional retraining.

E.3 VIDEO CAPTIONING

Our approach is closely related to video recaptioning research, as our MLLM must produce dense captions based on the given conditions. In text-to-video settings, prior work (Fan et al. (2024); Yang et al. (2024c); Nan et al. (2024); Islam et al. (2024)) has demonstrated the benefits of recaptioning videos to obtain more detailed textual annotations, thereby improving the training of longer and higher-quality video generation via DiT. ShareGPT4Video (Chen et al. (2024a)), for example, employs GPT-4V (GPT (2023)) to reinterpret video content and produce richer captions. MiraData (Ju

Type	#Inst.	#Condition	Short Caption (#Avg. Len.)	#Structured Caption(#Avg. Len.)
Identities	200	350	65.28	284.97
Camera	200	200	50.25	208.01
Depth	200	200	54.22	225.09
Human Pose	200	200	58.38	259.03
Camera+Identities	200	622	53.41	209.17
Camera+Depth	200	400	51.43	208.81
Identities+Depth	200	555	53.14	286.83
Camera+Identities+Depth	200	756	58.35	289.21

Table 7: Statistics of the constructed test datasets. **#Inst.** denotes the number of instances, and **#Condi.** indicates the number of unique conditions. **Short Cap. #Avg. Len** represents the average caption length of short captions, and **Structured Cap. #Avg. Len.** represents the average caption length of structured captions.

et al. (2024) introduces structured dense recaptioning, while InstanceCap Fan et al. (2024) focuses on instance-consistent dense recaptioning. Although we also pursue structured dense captions to enhance generation quality, our method diverges fundamentally from these previous approaches. First, because DiT models are already sufficiently powerful, we directly adopt an off-the-shelf Video DiT without incurring the substantial cost of training it with dense captions. Instead, we train an MLLM as an any-condition encoder at a comparatively lower cost; in the text-to-video scenario, for instance, we only need to train on pairs of short and dense captions, which are far easier and more abundant to obtain. Second, prior methods that recapturing the entire video risk introducing noise or even hallucinated content due to the current limitations of MLLMs in video understanding, potentially undermining DiT training quality, whereas our framework avoids this issue. Most importantly, while these approaches may rely on dense-caption-trained DiT models, real-world inference often involves very concise user prompts, creating a mismatch that can diminish final generation quality.

E.4 MULTIMODAL LARGE LANGUAGE MODELS

Recent advances in Large Language Models (LLMs) Yang et al. (2024a) have catalyzed a surge of interest in extending their capabilities to multimodal domains Lin et al. (2024a); Li et al. (2024a); Zhang et al. (2024b). A number of works integrate a vision encoder (e.g., CLIP Liu et al. (2024b), DINOv2Tong et al. (2024), OpenCLIPLi et al. (2024b)) with an LLM, often through a lightweight “connector” module (e.g., MLP Liu et al. (2024b), Q-former Li et al. (2023)), enabling the model to process image and video inputs with minimal additional training data and parameters Liu et al. (2024b); Wu et al. (2024); Lin et al. (2024a). These approaches have demonstrated promising performance on tasks such as image captioning, visual question answering, and video understanding. Beyond purely visual data, some researchers have investigated broader modalities, such as 3D motion Chen et al. (2024b); Jiang et al. (2023) or audio Chu et al. (2024); Wu et al. (2024), thereby expanding the application range of multimodal LLMs. Despite these advances, most existing MLLMs focus on a limited set of visual modalities and do not readily accommodate more specialized inputs like human pose or camera motion. This gap restricts their ability to handle more diverse and complex conditions in the controllable video generation field. In contrast, our work targets a broader spectrum of modalities, aiming to build a single model capable of interpreting and unifying image, video, human pose, and camera conditions. Specifically, we augment an existing MLLM with dedicated encoders for motion and camera features, thereby equipping the model to process arbitrary multimodal conditions and facilitate controllable video generation.

F EXTENDED DATASET DETAILS

Visualization of Short and Structured Caption. Here, we visualize the short and structured caption in Fig. 7, 8. Notably, the structured caption captures the video content in greater detail and across multiple aspects. In contrast, the short caption deliberately omits any information already supplied by a non-text condition—for example, camera movement is excluded in Fig. 7’s short caption but included in Fig. 8 because it is not specified by the provided multiple-identity images. Moreover, we visualize the word distribution of the structured captions in Any2CapIns in Fig. 9.

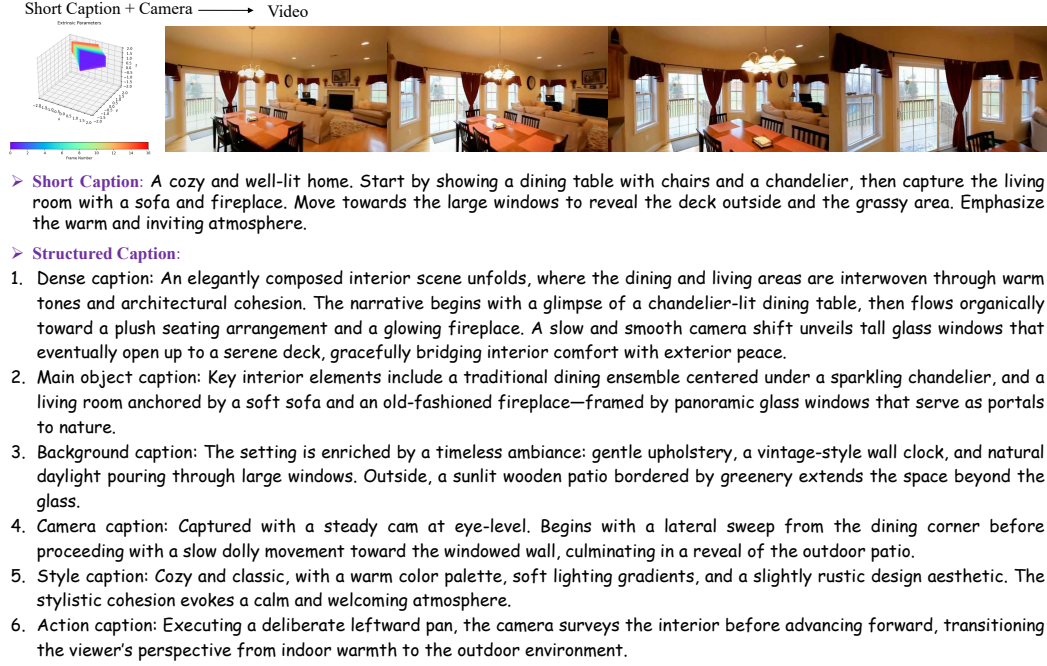


Figure 7: Illustrations of constructed short and structured captions under the camera-to-video generation.

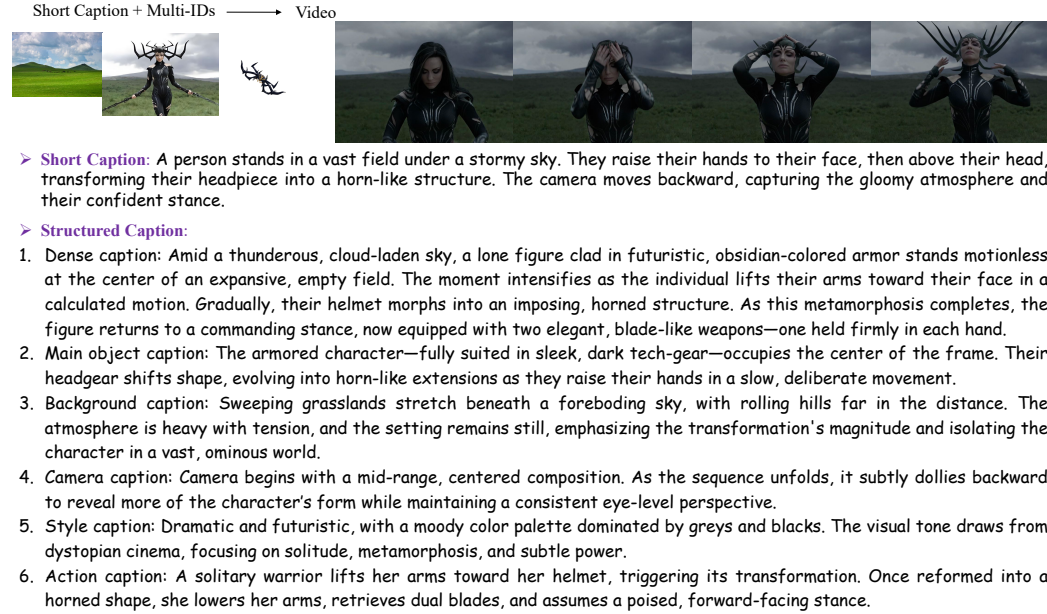


Figure 8: Illustrations of constructed short and structured captions under the multiIDs-to-video generation.

Prompt Visualization for Short Caption Construction. In Tab. 8 and 9, we show the system prompts used by GPT-4V to generate short captions. The prompt explicitly instructs GPT-4V to consider the given conditions comprehensively and produce short prompts that focus on information not covered by the non-textual inputs. For instance, when multiple identities are specified, the short prompt should avoid repeating their appearance attributes and instead highlight interactions among the identities. Conversely, when depth is the input condition, the short prompt should include more detailed appearance-related information.



Figure 9: Word cloud of different structured captions in Any2CapIns dataset, showing the diversity.

Multi-IDs Here is the scenario: We have an MLLM model that supports a text image-conditioned intrinsic-video-caption generation task. The system input consists of:

1. A reference image composed of 2-3 horizontally stitched images (provided by the user), each stitched image containing one or several target objects for reference); and
2. A concise textual prompt (referred to as text B, the user’s instruction).

The model’s output is a detailed descriptive caption (**text A**) that thoroughly describes the video corresponding to the user’s input prompt (**text B**) in great detail. Your task is to perform a reverse engineering. Based on the given reference image (the target objects) and the detailed target video caption (text A), you need to generate a **reasonable and concise user prompt (text B)** through your understanding, analysis, and imagination. You must adhere to the following rules:

1. Text A is a dense caption of a video, including all the key objects, their attributes, relationships, background, camera movements, style, and more. Carefully analyze this caption for all relevant details.
2. Analyze the provided reference images in detail to identify the differences or missing details compared to the target video description. These may include environment details, the interaction between objects, the progression of actions, camera movements, style, or any elements not covered by the reference image. Based on these analyses, generate the user’s instructions.
3. The user’s prompt must include the following aspects: first, an overall description of where the target objects are and what they are doing, along with the temporal progression of their actions. Then, it should describe the background, style, and camera movements.
4. If the target video introduces new objects not present in the reference images, the user’s prompt should describe the attributes of the new target objects and their interactions with the other target objects.
5. If the video’s style differs from the reference, briefly describe the style in a few words.
6. When the background needs to be described, include details about people, settings, and styles present in the background.
7. Avoid repeating information that can be inferred from the reference images, and eliminate redundant descriptions in the user prompt.
8. The user prompt (text B) must be written in simple wording, maintaining a concise style with short sentences.
9. The user’s instructions should vary in expression; For example, prompts do not always need to start with the main subject. They can begin with environmental details, camera movements, or other contextual aspects.

Here are three examples representing the desired pattern:

=====

[In-context Examples]

=====

[Input]

Table 8: Demonstration of the prompt used for GPT-4V to generate the short prompt when the input condition is the multi-IDs.

G MORE STATISTICS INFORMATION OF IRSCORE

We generate a total of 15,378 question-answer (QA) pairs, averaging 19.2 pairs per structured caption. Fig. 10 presents the distribution of constructed questions across different aspects in the structured caption, and Tab. 10 shows representative QA pairs for each aspect. Notably, questions under the *main object* category emphasize fine-grained details such as clothing color or hairstyles, while *action* questions focus on object interactions and movements. This level of specificity allows us to rigorously assess whether the generated captions are both complete and precise.

Depth Here is the scenario: We have an MLLM model that supports a text & image-conditioned intrinsic-video-caption generation task. The system input consists of:

1. A reference image composed of 3-5 horizontally stitched depth maps in temporal sequence (provided by the user, each map containing depth information for reference); and

2. A concise textual prompt (referred to as text B, the user's instruction).

The model's output is a detailed descriptive caption (text A) that thoroughly describes the video corresponding to the user's input prompt (text B) in great detail. Now, I need you to perform a reverse engineering task. Based on the given reference image (the depths) and the detailed target video caption (text A), you must generate a reasonable and concise user prompt (text B) through your understanding, analysis, and imagination. To ensure accurate and effective outputs, follow these rules strictly:

1. Text A is a dense caption of a video, including all the key objects, their attributes, relationships, background, camera movements, style, and more. Carefully analyze this caption to extract the necessary details.

2. Since the depth information already provides the necessary geometric outlines and layout details. Do not repeat this information in the user prompt. Instead, focus on the aspects not covered by the depth maps.

3. The user's instruction should highlight details not included in the depth map, such as environmental details, the appearance of the subjects, interactions between subjects, the progression of actions, relationships between the subjects and the environment, camera movements, and overall style.

4. For dense depth maps (more than 5 maps), assume the maps provide the camera movements and actions between objects, focusing on describing the appearance of the subjects and environment, the atmosphere, and subtle interactions between subjects and their environment.

5. For sparse depth maps (5 maps or fewer), assume the maps only provide scene outlines. Emphasize details about the subjects' appearance, environment, interactions between subjects, relationships between subjects and the environment, and camera movements.

6. The user prompt (text B) must be written in simple wording, maintaining a concise style with short sentences, with a total word count not exceeding 100.

7. Your output should be a continuous series of sentences, not a list or bullet points.

8. The user's instructions should vary in expression; they don't always need to begin with a description of the main subject. They could also start with environmental details or camera movements.

Here are three examples representing the desired pattern:

=====

[In-context Examples]

=====

[Input]

=====

Table 9: Demonstration of the prompt used for GPT-4V to generate the short prompt when the input condition is the depth.

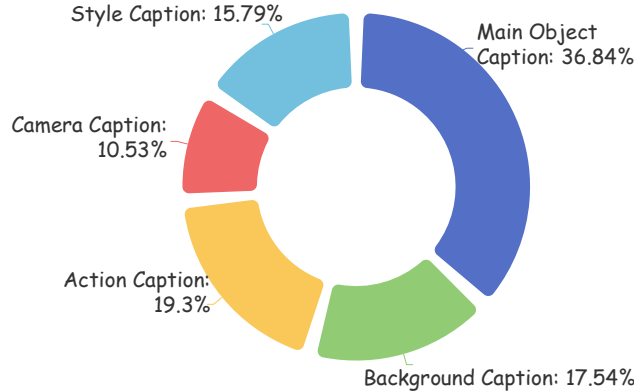


Figure 10: QA pairs proportion in structured captions.

H DETAILED SETUPS

H.1 DETAILED TESTING DATASET

Here, we present the statistics of the test in Tab. 7, which covers four types of single conditions (e.g., *Depth*, *Camera*, *Identities*, and *Human pose*), and four types of compositional conditions (e.g., *Camera+Identities*, *Camera+Depth*, *Identities+Depth* and *Camera+Identities+Depth*). Each category contains 200 instances.

Aspect	QA Pairs
Main Object	What is the young woman adjusting as she walks down the corridor? Her wide-brimmed hat.
	What color is the young woman's T-shirt? Light blue.
	How does the young woman feel as she walks down the corridor? Happy and carefree.
	What is the young woman wearing? Light blue t-shirt with pink lettering, blue jeans, and a wide-brimmed hat.
	What is the young woman's hair length? Long.
	What is the position of the young woman in the frame? In the center of the frame.
	What is the main object in the video? A large shark.
	What is the color of the underwater scene? Blue.
	What are the two scientists wearing? White lab coats and gloves.
	What is the first scientist using? A microscope.
Background	Where is the young woman walking? Down a corridor.
	What time of day does the scene appear to be set? Daytime.
	What can be seen in the background of the corridor? Beige walls and large windows.
	What is the weather like in the video? Clear.
	Where is the shark located? On the ocean floor.
	What surrounds the shark in the video? Smaller fish.
	Where is the laboratory setting? In a brightly lit environment with shelves filled with bottles.
Camera	What detail does the background highlight? The scientific setting with static emphasis.
	How does the camera follow the young woman? Moving backward
	What is the camera's height relative to the person? Roughly the same height as the person.
	What shot type does the camera maintain? Medium close-up shot of the upper body.
	How does the camera position itself to capture the subject? At a higher angle, shooting downward.
	How does the camera capture the environment? From a medium distance.
	How is the camera positioned? At approximately the same eye level as the subjects, maintaining a close-up shot.
Style	How does the camera move in the video? It pans to the right.
	What is the style of the video? Casual and candid.
	What kind of design does the corridor have? Modern and clean design.
	What style does the video portray? Naturalistic style with clear, vivid visuals.
	What does the video style emphasize? Clinical, high-tech, and scientific precision.
	What is the color theme of the lighting? Bright and cool.
Action	What kind of atmosphere does the laboratory have? Professional and scientific.
	What does the young woman do with both hands occasionally? Adjusts her hat.
	What is the young woman doing as she moves? Walking forward with her hands on her hat.
	What is the main action of the shark in the video? Lying motionless.
	What is the movement of the fish like? Calm and occasionally darting.
	What is the movement of the first scientist at the beginning? Examines a microscope.
	What task is the second scientist engaged in? Handling a pipette and a beaker filled with green liquid.
	How does the second scientist transfer the liquid? Carefully using a pipette into the beaker.
	Are there any noticeable movements in the background? Occasional small particles floating.

Table 10: Demonstration of generated question-answer pairs utilized in IRSCORE calculation.

Configuration	Stage-1: Alignment Learning		Stage-2: Condition-Interpreting Learning			
	Camera	Motion	Identities	Human pose	Camera	Depth
Optimizer	AdamW	AdamW	AdamW	AdamW	AdamW	AdamW
Precision	bfloat16	bfloat16	bfloat16	bfloat16	bfloat16	bfloat16
Learning Rate	5e5	5e5	5e5	5e5	5e5	1e5
Weight Decay	0.01	0.01	0.01	0.01	0.01	0.01
Joint Train Ratio	0.0	0.0	0.0	0.4	0.6	0.8
Betas	(0.9, 0.99)	(0.9, 0.99)	(0.9, 0.99)	(0.9, 0.99)	(0.9, 0.99)	
Dropout Possibility	0.0	0.0	0.4	0.6	0.6	0.6
Dropout (Short Cap.)	0.0	0.0	0.6	0.6	0.6	0.6
Batch Size Per GPU	4	4	4	4	4	4
Training Data	Camera Movement Description		MultiIDs	Human Pose		
	Dataset (Manual)	Dataset (Manual)		LLaVA-150K	LLaVA-150K	LLaVA-150K Alpaca-50K

Table 11: Training recipes for Any2Caption.

H.2 IMPLEMENTATION DETAILS

We leverage Qwen2VL-7B Wang et al. (2024a) as the backbone of our model, which supports both image and video understanding. The human pose in the input conditions is encoded and processed

in the video format. The camera encoder adopts the vision encoder architecture, with the following settings: in channels set to 96, patch size of 16, depth of 8, and 8 attention heads. During training, to simulate the brevity and randomness of user inputs, we randomly drop sentences from the short caption with a dropout rate of 0.6; a similar dropout strategy is applied to non-textual conditions. We conducted the training on $8 \times A800$ GPUs.

H.3 DETAILED TRAINING PROCEDURE

We employ a two-stage training process to enhance the alignment and interpretability of multimodal conditions in Any2Caption.

Stage-1: Alignment learning. This stage focuses on aligning features extracted by the camera encoder with the LLM feature space. To achieve this, we first extract camera movement descriptions (e.g., *fixed*, *backward*, *pan to the right*) from the camera captions in Any2CapIns to construct a camera movement description dataset. We then introduce two specialized tokens, `<|camera_start|>` and `<|camera_end|>`, at the beginning and end of the camera feature embeddings. During training, only the camera encoder is optimized, while all other parameters in Any2Caption remain frozen. Similarly, for motion alignment, we construct a motion description dataset by extracting action descriptions (e.g., *walking*, *dancing*, *holding*) from the action captions in Any2CapIns. We then freeze all model parameters except those in the motion encoder to ensure the LLM effectively understands motion-related conditions.

Stage-2: Condition-Interpreting Learning. After alignment learning, we initialize Any2Caption with the pre-trained Qwen2-VL, motion encoder, and camera encoder weights. We then employ a progressive mixed training strategy, updating only the `lm_head` while keeping the multimodal encoders frozen. The training follows a sequential order based on condition complexity: identities $\Rightarrow$ human pose $\Rightarrow$ camera $\Rightarrow$ depth. Correspondingly, the integration ratio of additional vision/text instruction datasets is progressively increased, set at 0.0, 0.4, 0.6, and 0.8, ensuring a balanced learning process between condition-specific specialization and generalization.

H.4 DETAILED IMPLEMENTATIONS

In Tab. 11, we list the detailed hyperparameter settings in two stages. All the training is conducted on $8 \times A800$ (80G) GPUs.

I EXTENDED EXPERIMENT RESULTS AND ANALYSES

I.1 THE EFFECTIVENESS OF THE ANYCAPTION.

We evaluate the performance of our proposed model, AnyCaption, against various baselines on the detailed video captioning benchmark VDC Chai et al. (2025). This evaluation is designed to inversely assess whether the model can accurately generate fine-grained content aligned with the target video. As shown in Table 12, our model consistently outperforms all baseline video-language models across multiple categories. In particular, it achieves the best results in detail-sensitive aspects such as camera perspective and main object descriptions. These results highlight the strong capability of AnyCaption in capturing fine-grained visual semantics and generating detailed, faithful video descriptions which are an essential prerequisite for high-fidelity video generation. Furthermore, we conduct additional comparisons on our proposed benchmark, with results presented in Table 13. AnyCaption also outperforms all baselines across all evaluated dimensions, demonstrating its effectiveness in capturing users’ core intentions and generating detailed, goal-consistent descriptions of target videos.

I.2 THE CAPABILITY FOR UNDERSTANDING COMPLEX INSTRUCTION.

We further examine Any2Caption’s ability to handle complex user instructions, particularly regarding whether it accurately captures the user’s intended generation targets. From Fig. 11, we observe that the model focuses precisely on the user-specified main objects, such as a “woman warrior” or a background “filled with chaos and destruction”—when producing structured captions.

Model	Camera		Short		Background		Main Object		Detailed		Overall	
	Acc	Score	Acc	Score	Acc	Score	Acc	Score	Acc	Score	Acc	Score
LLaMA-VID Li et al. (2024a)	39.47	2.10	29.92	1.56	38.01	1.45	31.24	1.59	25.67	1.38	32.86	1.62
Video-ChatGPT-7B Maaz et al. (2024)	37.46	2.00	28.92	1.51	35.89	1.37	30.13	1.50	25.02	1.34	31.48	1.54
Video-LLaVA-7B Lin et al. (2024a)	37.48	1.97	30.74	1.58	35.80	1.38	30.19	1.46	25.22	1.33	31.89	1.54
LongVA-7B Zhang et al. (2024a)	35.32	1.94	29.35	1.53	33.91	1.33	30.34	1.49	25.40	1.33	30.86	1.52
LLaVA-NeXT-V7B Zhang et al. (2024d)	36.93	1.96	30.59	1.56	36.10	1.43	31.79	1.50	26.01	1.35	32.28	1.56
LLaVA-1.6-7B Liu et al. (2024c)	36.56	1.91	30.94	1.59	35.62	1.38	32.06	1.51	26.31	1.34	32.29	1.55
ShareGPTVideo-8B Chen et al. (2024a)	39.16	2.06	32.61	1.70	38.90	1.99	34.89	1.68	29.81	1.52	35.07	1.79
LLaVA-OV-7B Li et al. (2025)	39.02	2.04	31.84	1.64	36.52	1.84	33.24	1.65	28.41	1.48	33.81	1.73
Qwen2-VL-7B Wang et al. (2024a)	38.25	1.98	30.58	1.61	36.79	1.86	37.49	1.96	32.23	1.61	35.07	1.80
Any2Caption	40.26	2.19	31.88	1.64	38.80	1.96	41.50	2.06	35.13	1.87	37.51	1.94

Table 12: Quantitative comparison on the VDC video captioning benchmark Chai et al. (2025). We report VDCscore across five caption aspects: camera captions, short captions, background captions, main object captions, and detailed captions. The final two columns show the overall performance.

Model	Dense		Main Object		Background		Action		Style		Camera	
	Acc	Qual.	Acc	Qual.	Acc	Qual.	Acc	Qual.	Acc	Qual.	Acc	Qual.
LLaVA-NeXT-V7B Zhang et al. (2024d)	65.23	2.81	43.57	2.13	55.41	2.16	40.87	1.72	52.64	2.43	35.97	1.90
LLaVA-1.6-7B Liu et al. (2024c)	66.01	2.83	44.02	2.17	54.98	2.14	39.11	1.75	52.91	2.41	36.13	1.92
ShareGPTVideo-8B Chen et al. (2024a)	75.36	3.35	54.90	2.68	66.77	2.65	51.92	2.05	62.48	2.94	55.29	2.68
Qwen2-VL-7B Wang et al. (2024a)	71.42	3.14	51.31	2.41	63.51	2.45	50.77	1.93	59.18	2.75	51.22	2.41
Any2Caption	79.45	3.81	56.29	2.78	70.07	2.71	56.78	2.14	65.74	3.15	67.25	3.79

Table 13: Comparison between Video-LLMs baselines and Any2Caption on our proposed benchmark. Since the baseline models are limited to handling only video and image inputs, we conduct experiments on the *Depth* and *Multiple Identities* benchmarks. For a fair comparison, we adopt a one-shot setting for the baseline models, enabling them to generate detailed descriptions and structured outputs for the target videos. We report the average scores of intention reasoning accuracy (Acc) and quality (Qual.) across all tasks.

In contrast, a short caption combined with condition captions often includes extraneous objects or background details present in the identity images, which distract from the user’s intended targets in the final video generation.

Additionally, we assess the model’s performance on instructions containing implicit objects or actions, as shown in Fig. 12 and 13. In these examples, the model correctly interprets phrases like “the most right person” as “a young Black woman with long, curly brown hair, wearing a black and white outfit” and similarly associates implicitly specified objects with the provided conditions, generating structured captions that align with the user’s goals.

Lastly, Fig. 16 compares videos generated using different captions. The results indicate that structured captions significantly improve both the smoothness of motion and the overall consistency of the generated videos.

I.3 MORE VIDEO VISUALIZATION

Here, we compare the results of different customized models Ju et al. (2025) after integrating structured captions. Fig. 14 shows camera-controlled video generation results, and Fig. 15 illustrates depth-controlled outcomes. We observe that structured captions improve image quality and motion smoothness by providing richer scene details.

For multi-identity-controlled video generation, as depicted in Fig. 16, 17, and 18, structured captions lead to more expressive and realistic representations of the referenced identities, with more accurate colors and smoother actions.

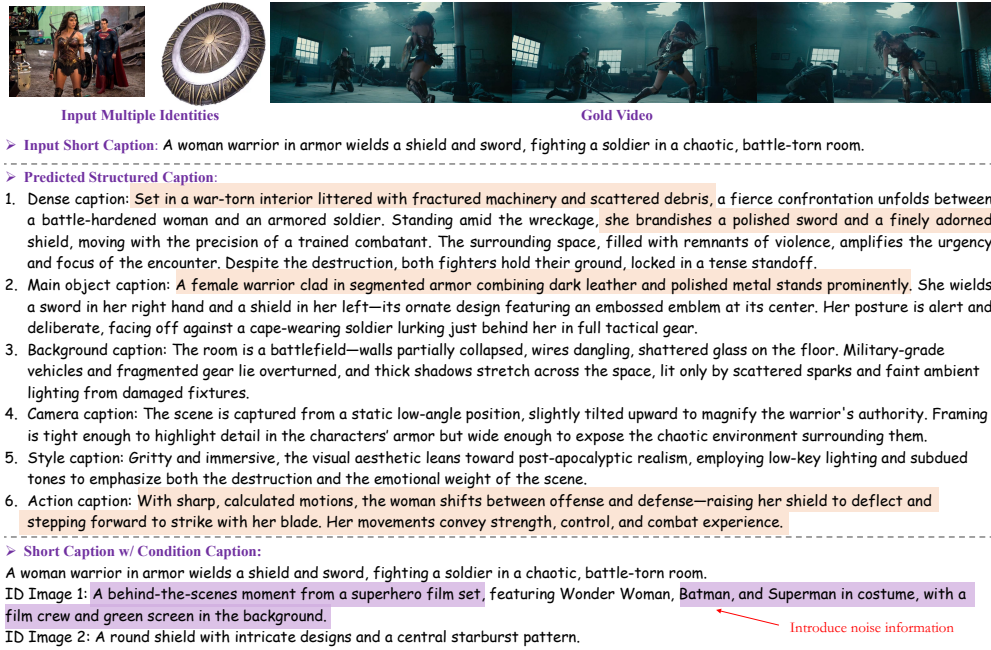


Figure 11: Illustrations of input short caption, predicted structured caption, and short caption combined with condition captions. The condition captions may introduce noisy information due to the focus distraction of the background of input identities, for example, the *Batman*, *Superman*, *film crew*.

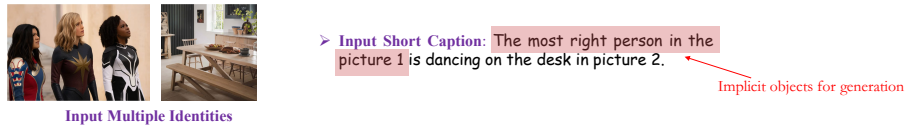


Figure 12: Illustrations of predicted structured captions based on the input multiple identities and the short instruction that expresses the implicit objects for the video generation.

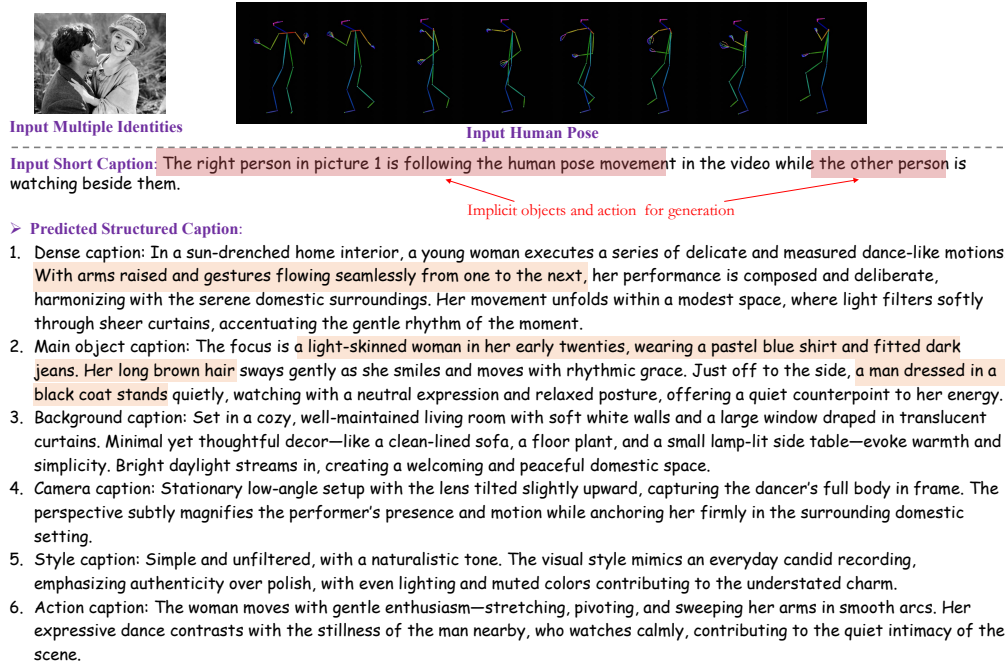
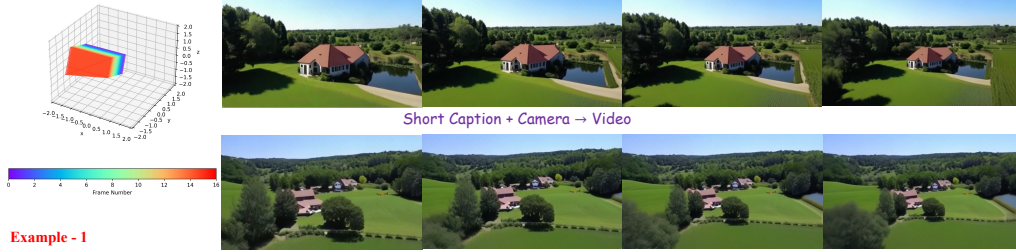


Figure 13: Illustrations of predicted structured captions based on the input multiple identities and the short instruction that expresses the implicit objects and the action for the target video generation.

➤ **Short Caption:** A serene video of a large house with a red roof and a spacious porch, surrounded by lush greenery. A peaceful countryside setting with vibrant colors and a tranquil atmosphere.



➤ **Short Caption:** A well-lit dining and living room with elegant and classic decor. The dining table is surrounded by chairs and has a chandelier above it. There's a wooden cabinet against the wall. The background features a hallway with a staircase and another dining area visible. The decor includes wooden furniture and framed pictures on the walls.



➤ **Short Caption:** The scene is bathed in bright sunlight, emphasizing the warm and inviting atmosphere. A modern house with large windows and a balcony is showcased. Potted plants accent the architectural details. Distant mountains frame the view. Lush greenery surrounds the scene. The sky is a clear blue, dotted with scattered clouds. A dreamy lens flare effect adds to the serene quality. The overall ambience is tranquil and picturesque.



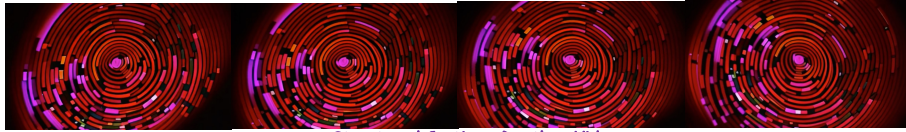
Figure 14: Illustrations of predicted structured captions based on the input multiple identities and the short instruction that expresses the implicit objects and the action for the target video generation.

➤ **Short Caption:** A dark, static background enhances the brightly colored, rotating spiral of small blocks. The camera remains fixed, capturing the mesmerizing effect of the colors shifting subtly. The dynamic movement creates a hypnotic, abstract visual style.

Example - 4



Short Caption + Depth → Video



Structured Caption + Depth → Video

➤ **Short Caption:** The park is sunny and lush with green trees and a small pond. A young couple in their late twenties embraces and shares a kiss. The woman, in a white sleeveless wedding dress and holding a bouquet, playfully touches the man's face. He is dressed in a black suit. The scene is romantic and intimate, with soft, natural lighting. The camera pans gently, capturing their affectionate interaction and the serene environment.

Example - 5



Short Caption + Depth → Video



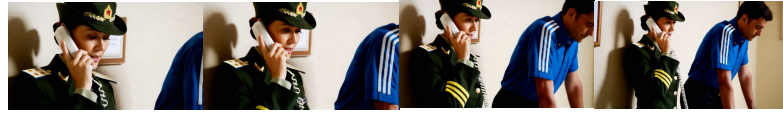
Structured Caption + Depth → Video

Figure 15: Illustrations of predicted structured captions based on the input multiple identities and the short instruction that expresses the implicit objects and the action for the target video generation.

➤ **Short Caption:** A woman in a military uniform talks on the phone while holding a document, standing beside a man in a blue uniform against a wall. The setting is formal and professional, suggesting an official procedure in an institutional environment with light-colored walls and framed documents. The camera captures their upper bodies, moving backward and tilting upward, transitioning from a close-up to a medium close-up shot.



Short Caption + Identities → Video



Short Caption w/ Condition Caption + Identities → Video



Structured Caption + Identities → Video

➤ **Short Caption:** A young woman in a traditional colorful outfit rides a galloping black horse through a lush green landscape. The camera follows her movements, capturing the dynamic and vibrant scene, with her hair flowing in the wind. The background is blurred to emphasize the speed and joy of the rider. The overall feel is natural and bright.



Short Caption + Identities → Video



Short Caption w/ Condition Caption + Identities → Video



Structured Caption + Identities → Video

Figure 16: Illustrations of predicted structured captions based on the input multiple identities and the short instruction that expresses the implicit objects and the action for the target video generation.

➤ **Short Caption:** A martial arts dojo scene where an instructor in black demonstrates techniques, throwing a student in white to the ground. Students sit in a circle on the green mat floor, observing attentively. In the background, banners and signs indicate martial arts training, with a seated audience and standing spectators. The camera moves from a medium to a long shot, capturing the full scene with respect and focus.



Short Caption + Identities → Video



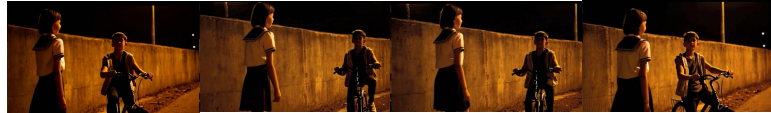
Short Caption w/ Condition Caption + Identities → Video

Example - 8

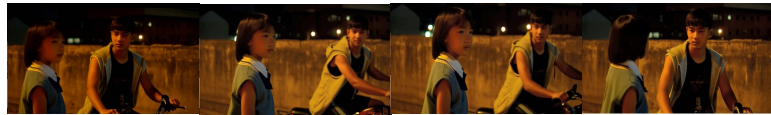


Structured Caption + Identities → Video

➤ **Short Caption:** A young girl wearing a school uniform and a young man in casual clothes are walking side by side along a dimly lit concrete wall at night. The girl walks on the left while the boy rides a bicycle on the right. The background is urban and gritty, with warm, moody lighting. The camera follows them closely, capturing a medium close-up shot of their upper bodies from different angles as they move. The scene has a nostalgic and contemplative atmosphere.



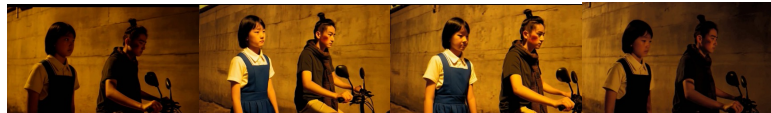
Short Caption + Identities → Video



Short Caption w/ Condition Caption + Identities → Video



Example - 9



Structured Caption + Identities → Video

Figure 17: Illustrations of predicted structured captions based on the input multiple identities and the short instruction that expresses the implicit objects and the action for the target video generation.

➤ **Short Caption:** A woman wearing fashionable clothes stands in the room, smiling and showing the goods in her hand. Then the camera zooms in and focuses on the details of the goods in the person's hand.

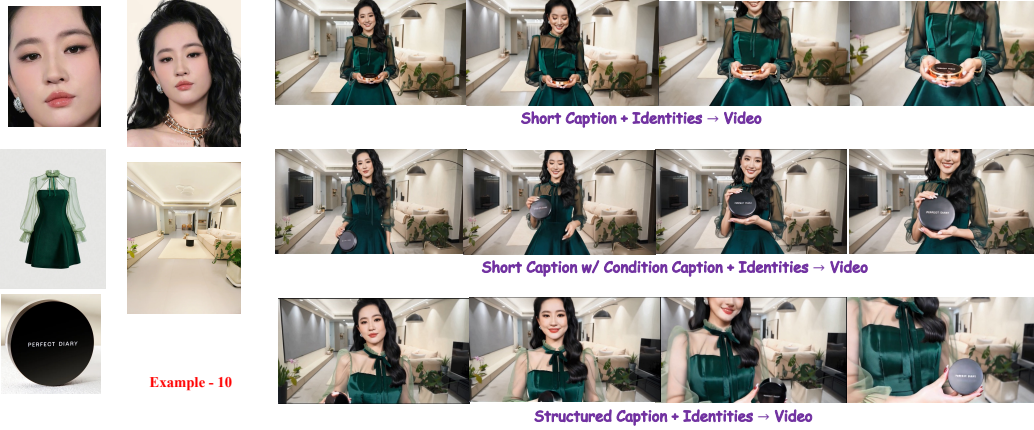


Figure 18: Illustrations of predicted structured captions based on the input multiple identities and the short instruction that expresses the implicit objects and the action for the target video generation.