Any2Caption: Interpreting Any Condition to Caption for Controllable Video Generation

Shengqiong Wu[1,2*] Weicai Ye[1,✉] Jiahao Wang[1] Quande Liu[1] Xintao Wang[1]
Pengfei Wan[1] Di Zhang[1] Kun Gai[1] Shuicheng Yan[2] Hao Fei[2,✉] Tat-Seng Chua[2]
[1]Kuaishou Technology [2]National University of Singapore
https://sqwu.top/Any2Cap/

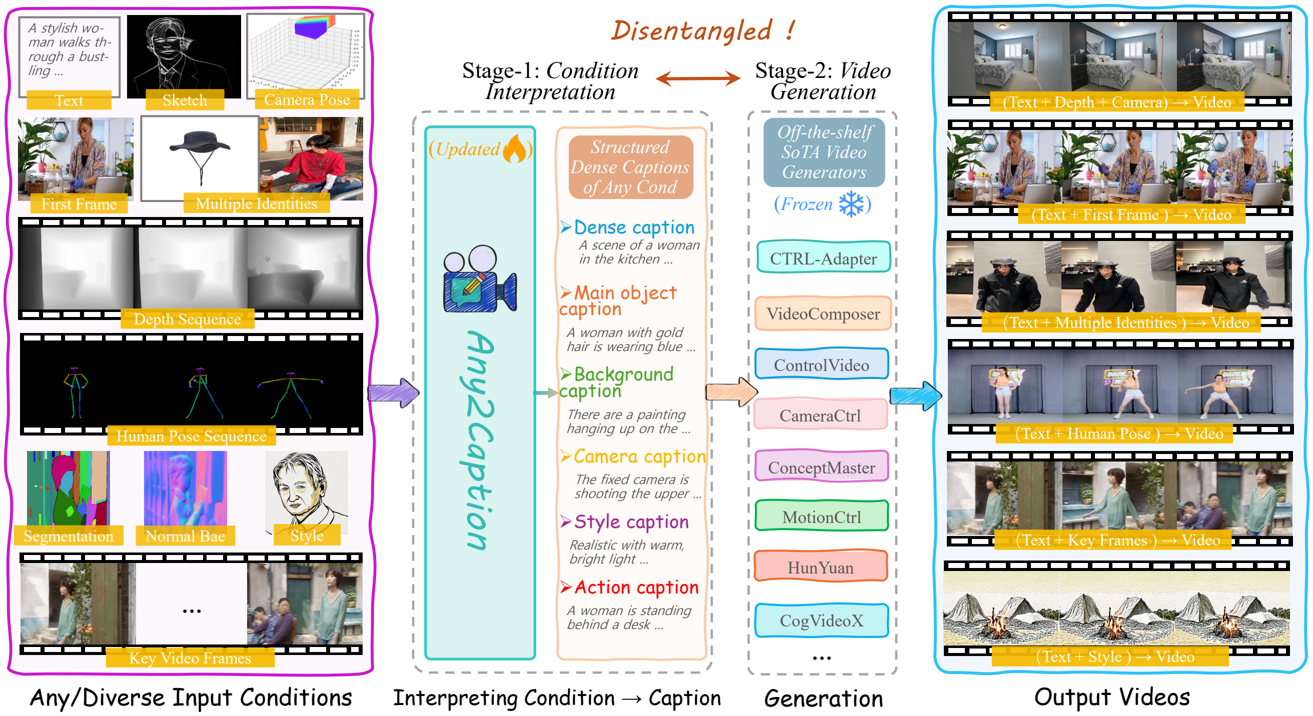
A stylish woman walks through a bustling ...
Text
Sketch
Camera Pose
First Frame
Multiple Identities
Depth Sequence
Human Pose Sequence
Segmentation
Normal Bae
Style
...
Key Video Frames
Any/Diverse Input Conditions
Disentangled !
Stage-1: Condition Interpretation
Stage-2: Video Generation
(Updated 🔥)
Any2Caption
Structured Dense Captions of Any Cond
Dense caption
A scene of a woman in the kitchen ...
Main object caption
A woman with gold hair is wearing blue ...
Background caption
There are a painting hanging up on the ...
Camera caption
The fixed camera is shooting the upper ...
Style caption
Realistic with warm, bright light ...
Action caption
A woman is standing behind a desk ...
Interpreting Condition → Caption
Off-the-shelf SoTA Video Generators
(Frozen ❄)
CTRL-Adapter
VideoComposer
ControlVideo
CameraCtrl
ConceptMaster
MotionCtrl
HunYuan
CogVideoX
...
Generation
(Text + Depth + Camera) → Video
(Text + First Frame) → Video
(Text + Multiple Identities) → Video
(Text + Human Pose) → Video
(Text + Key Frames) → Video
(Text + Style) → Video
Output Videos

Figure 1. We propose Any2Caption, an efficient and versatile framework for interpreting diverse conditions to structured captions, which then can be fed into any video generator to generate highly controllable videos.

Abstract

To address the bottleneck of accurate user intent interpretation within the current video generation community, we present ***Any2Caption****, a novel framework for controllable video generation under any condition. The key idea is to decouple 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 of our system in controllability and video quality across various aspects of existing video generation models.*

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 signif-

[*] Work done during internship at Kuaishou Technology.
[✉] Corresponding Author.

icant attention, driven by advancements in Diffusion Transformers (DiT) [2, 29, 44, 52, 76], 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 [24, 30, 67] 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 [10], MiraData [30], and InstanceCap [15]) to build dense structured captions for optimizing generative models. While dense captions are used during training, in real-world inference scenarios, users most likely provide concise or straightforward input prompts [15]. 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 [15, 67] 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.

In addition, to achieve more fine-grained controllable video generation, one effective strategy is to provide additional visual conditions besides text prompts—such as reference images [17, 62], identity [22, 46, 69], style [42, 68], human pose [33, 45], or camera [21, 75]—or even combinations of multiple conditions together [41, 58, 74]. 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 as VideoComposer [58], Ctrl-Adapter [41], and ControlVideo [74]. 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;

a) 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/diverse condition (or combination), and produces a densely structured caption, which is then passed on to any backbone video generators for controllable, high-quality video production. As Any2Caption disentangles the role of complex interpretation of multimodal inputs from the backbone 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 poses. Through extensive manual labeling combined with automated annotation by GPT-4V [1], 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. In addition, we devise a comprehensive evaluation strategy to thoroughly measure the model's capacity for interpreting user intent under these 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 shows a pronounced advantage when handling multiple combined conditions, effectively interpreting and synthesizing intricate user constraints into captions that align closely with user expectations. Our contributions are threefold:

- We for the first time pioneer a novel *any-condition-to-caption* paradigm of video generation, which bridges the gap between user-provided multimodal conditions and structured video generation instructions, leading to