Multi-Modal Video Generative Foundation Models
多模态视频生成基础模型

叶伟才 快手Kling

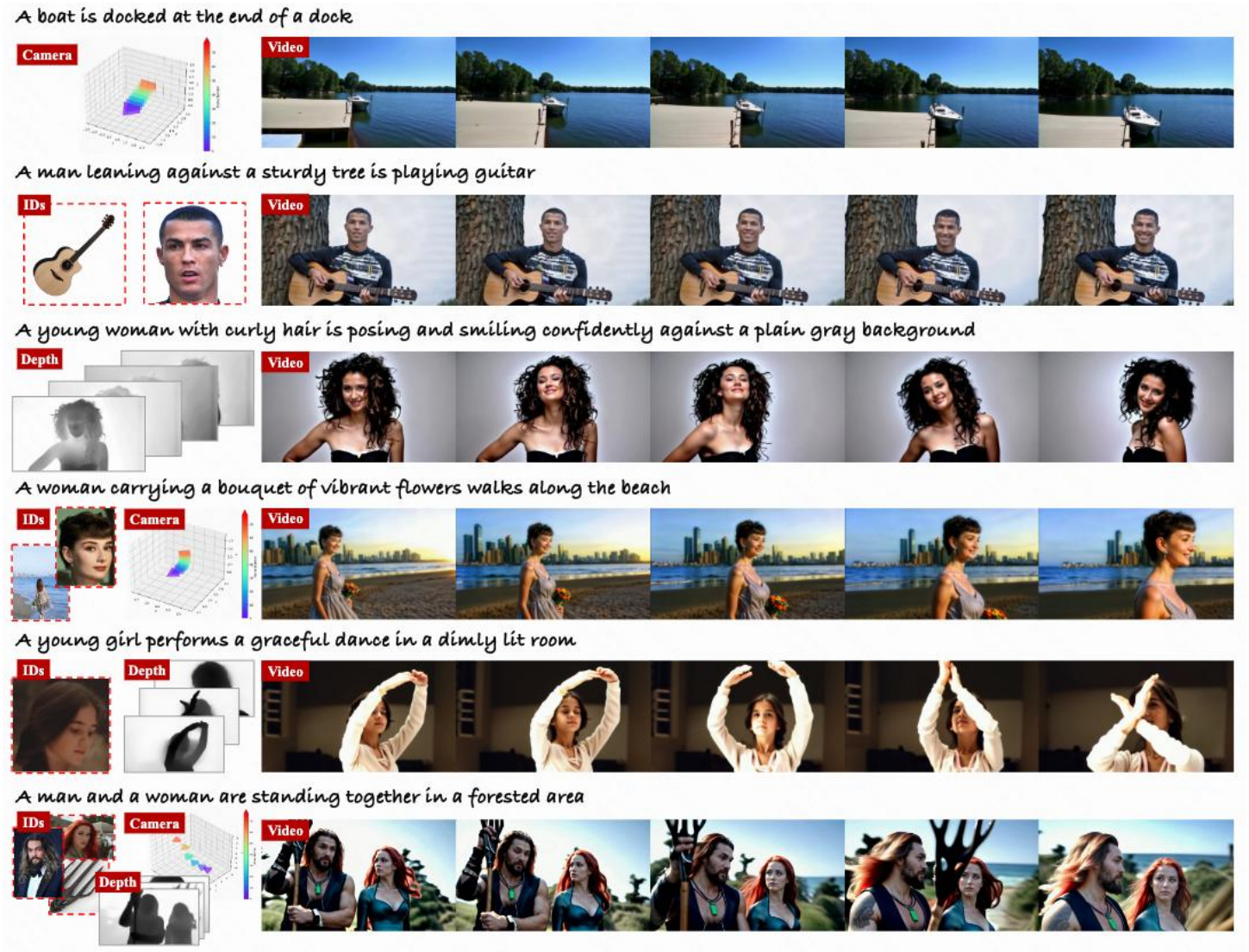
Valse Webinar 2025.03.12

https://ywcmaike.github.io/

Outline

- FullDiT: Video Generative Foundation Models with Multimodal Control via Full Attention (多模态视频生成基础模型）
- Any2Caption: Interpreting Any Condition to Caption for Controllable Video Generation (多模态用户意图理解caption模型）

FullDiT: Video Generative Foundation Models with Multimodal Control via Full Attention



FullDiT: Video Generative Foundation Models with Multimodal Control via Full Attention

- Motivation
  - **Trend**: T2V,I2V->M2V
  - drawback of adapter-based from single-task to multi-modal:
    - **branch conflicts** between independently trained adapters
    - **parameter redundancy** leading to increased computational cost
    - **suboptimal performance** compared to full fine-tuning
  - Ours：
    - **Unified Framework**
    - **Higher performance**
    - **Scalablity & Emergent Ability**
    - **Effective Training policy**

FullDiT: Video Generative Foundation Models with Multimodal Control via Full Attention

- Framework

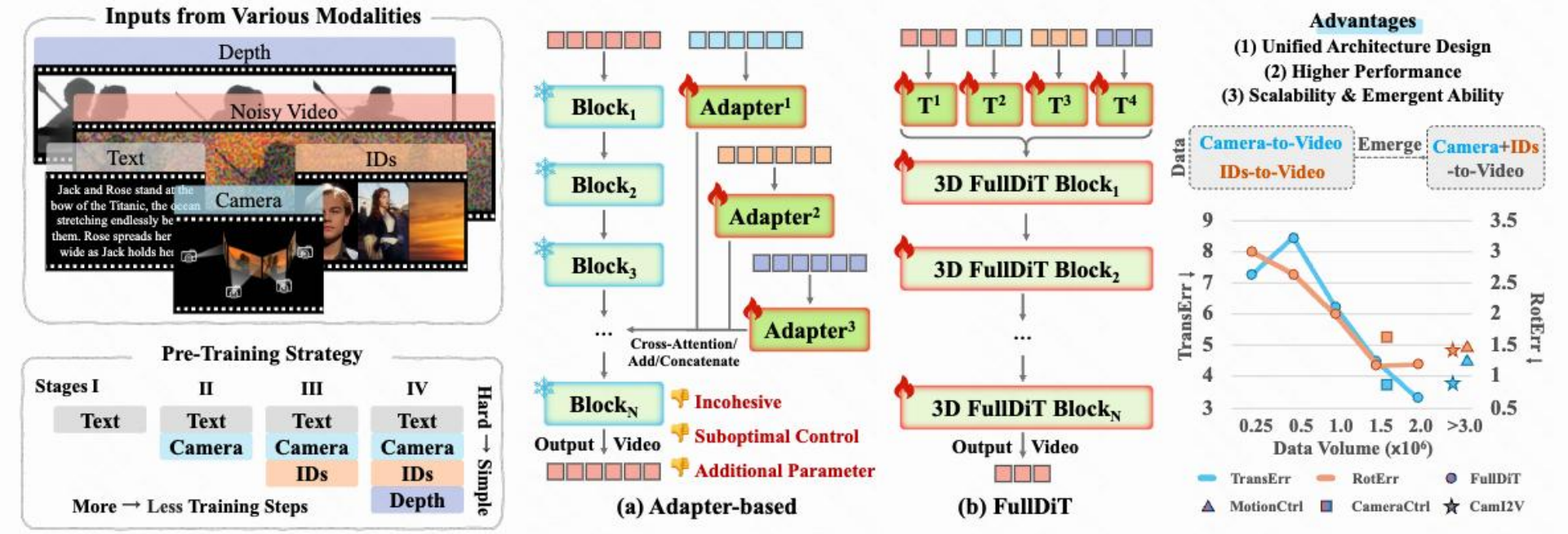


Figure 2. **Overview of *FullDiT* architecture, training strategy, and advantages.** We present input examples from various modalities in the top-left corner and illustrate the *FullDiT* pretraining strategy in the bottom-left corner. The middle part of the figure compares different integration methods for incorporating additional conditions. ❄ represents frozen blocks. 🔥 denotes trainable blocks. The subscript of each block signifies its layer index, while the superscript indicates the index of modality associated with the module. T represents the tokenization block for each modality in *FullDiT*. We show the advantages of *FullDiT* design in the right, especially the scalability (camera control performance as the increase of data volume) and emergent ability (generalizing to new tasks unseen in training data).

FullDiT: Video Generative Foundation Models with Multimodal Control via Full Attention

- Ablation
  - Training Strategy
    - Condition Training Order

Metrics	Camera			Identities		Depth
Stage	RotErr↓	TransErr↓	CamMC↓	DINO-I↑	CLIP-I↑	MAE↓
Depth→Camera→IDs	2.50	6.57	8.17	36.46	64.56	14.76
IDs→Camera→Depth	2.46	7.43	8.52	42.71	65.99	14.94
Camera→IDs→Depth	**1.20**	**3.31**	**3.98**	**46.22**	**68.59**	**14.71**

Table 2. **Ablation on condition training order.**

    - Training Stages

Metrics	Camera			Identities		Depth
Stage	RotErr↓	TransErr↓	CamMC↓	DINO-I↑	CLIP-I↑	MAE↓
I: Camera+ID+Depth	2.69	6.19	8.21	35.42	59.49	32.88
I: Camera	**1.19**	4.49	5.01	-	-	-
II: Camera+ID+Depth	1.23	4.14	4.78	37.21	65.85	15.81
I: Camera	**1.19**	4.49	5.01	-	-	-
II: Camera+ID	1.23	4.20	4.82	42.83	64.99	-
III: Camera+ID+Depth	1.20	**3.31**	**3.98**	**46.22**	**68.59**	**14.71**

Table 3. **Ablation on number of training stages.**

  - Model Architecture

Metrics	Text	Camera			Overall Quality		
Model	Clip Score↑	RotErr↓	TransErr↓	CamMC↓	Smoothness↑	Dynamic↑	Aesthetic↑
Adapter	22.58	1.28	3.35	4.17	**96.41**	28.42	4.88
FullDiT	**22.97**	**1.20**	**3.31**	**3.98**	96.40	**30.53**	**4.95**

Table 4. **Comparing *FullDiT* with adapter-based architecture.**

FullDiT: Video Generative Foundation Models with Multimodal Control via Full Attention

- Comparisons
  - single control

Metrics	Text	Camera			Identities		Depth	Overall Quality		
Model	**Clip Score↑**	**RotErr↓**	**TransErr↓**	**CamMC↓**	**DINO-I↑**	**CLIP-I↑**	**MAE↓**	**Smoothness↑**	**Dynamic↑**	**Aesthetic↑**
Camera to Video										
MotionCtrl [54]	22.27	1.49	4.41	4.84	-	-	-	96.16	11.43	4.71
CameraCtrl [17]	21.36	1.57	3.88	4.77	-	-	-	95.16	13.72	4.66
CamI2V‡ [68]	-	1.43	3.81	4.62	-	-	-	94.50	19.40	-
FullDiT	**22.97**	**1.20**	**3.31**	**3.98**	-	-	-	**96.40**	**30.53**	**4.95**
Identities to Video										
ConceptMaster [22]	18.54	-	-	-	39.97	65.63	-	**95.05**	10.14	5.21
FullDiT	**18.64**	-	-	-	**46.22**	**68.59**	-	94.95	**16.68**	**5.46**
Depth to Video										
Ctrl-Adapter‡ [32]	-	-	-	-	-	-	25.63	94.23	15.47	-
ControlVideo [66]	23.38	-	-	-	-	-	30.10	94.44	18.62	**5.91**
FullDiT	**23.40**	-	-	-	-	-	**14.71**	**95.42**	**23.12**	5.26

‡ Since this method only supports image-to-video generation, frame quality metrics are not reported.

Table 1. **Quantitative comparison of single task video generation.** We compare *FullDiT* with MotionCtrl [54], CameraCtrl [17], and CamI2V [17] on camera-to-video generation. For identity-to-video, due to a lack of open-source multiple identities video generation method, we compare with the internal $1B$ model: ConceptMaster [22]. We compare *FullDiT* with Ctrl-Adapter [32] and ControlVideo [66] for depth-to-video. We follow the default setting of each model for evaluation. Since most of previous methods can generate only 16 frames of video, we uniformly sample 16 frames from methods that generate more than 16 frames for comparison.

FullDiT: Video Generative Foundation Models with Multimodal Control via Full Attention

- Comparisons
  - single control

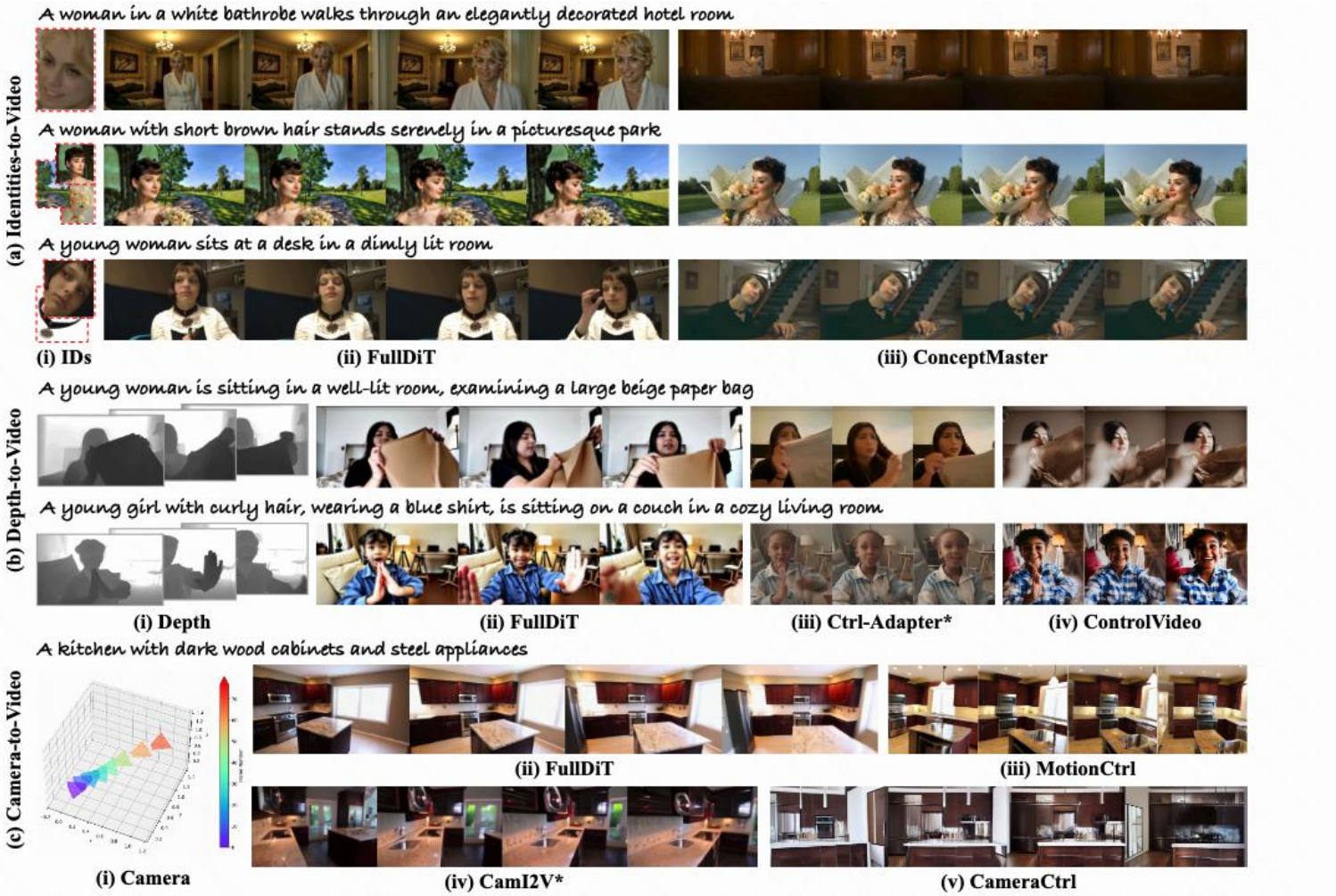


Figure 4. **Qualitative comparison of *FullDiT* and previous single control video generation methods.** We present identity-to-video results compared with ConceptMaster [22], depth-to-video results compared with Ctrl-Adapter [32] and ControlVideo [66], and camera-to-video results compared with MotionCtrl [54], CamI2V [68], and CameraCtrl [17]. Results denoted with * are image-to-video methods.

FullDiT: Video Generative Foundation Models with Multimodal Control via Full Attention

- Qualitative Results
  - multi-control

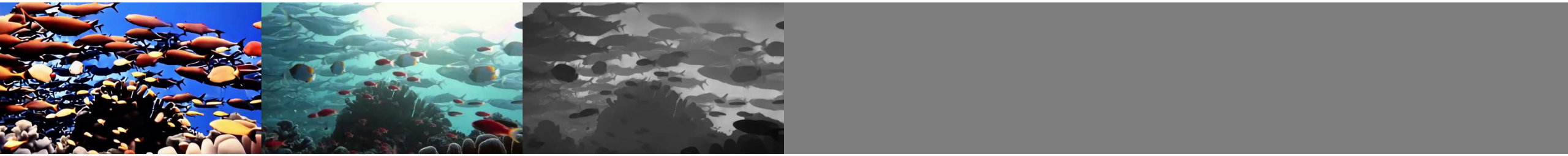
Figure 5. **Qualitative results of *FullDiT* with multiple control signals.** We show camera+identity+depth-to-video in (a) and (b), camera+identity-to-video in (c), identity+depth-to-video in (d), and camera+depth-to-video in (e).

FullDiT: Video Generative Foundation Models with Multimodal Control via Full Attention

- Qualitative Results
  - multiid -> video

FullDiT: Video Generative Foundation Models with Multimodal Control via Full Attention

- Qualitative Results
  - generated video | GT Video | depth



a vibrant underwater scene featuring a diverse school of fish swimming around a coral reef. The fish, predominantly red and white, move gracefully through the water, creating a mesmerizing dance of colors and shapes. The coral reef, with its in

a serene rural landscape during sunset, featuring a large, solitary electricity pylon standing prominently in the foreground. The pylon is surrounded by a vast, open field with dry, brown grass, and a few scattered trees. The sky is painted with war

FullDiT: Video Generative Foundation Models with Multimodal Control via Full Attention

- Qualitative Results
  - generated video | reference camera

a hallway leading to a living room with a couch. The video showcases a serene and well-lit hallway in a residential home. The hallway is spacious and features a large, open doorway leading to another room. The walls are painted white, and the f

an aerial view of a house on a hillside. The video showcases a serene and picturesque residential area with a large, modern house surrounded by lush greenery and well-maintained landscaping. The house features a combination of dark and light

FullDiT: Video Generative Foundation Models with Multimodal Control via Full Attention

- Qualitative Results (Emergent Ability)
  - generated video | reference camara | multiid

A young woman, adorned with a gold and diamond necklace and earrings, confidently strides down a charming street lined with elegant buildings featuring white facades and black windows, exuding an air of sophistication and poise. Her long, wav

A woman with short brown hair and a radiant smile stands in a serene park, surrounded by lush green trees and a clear blue sky. She is dressed in a beautiful green dress with a sheer, puffed-sleeve top and a solid skirt, her outfit perfectly co

FullDiT: Video Generative Foundation Models with Multimodal Control via Full Attention

- Qualitative Results (Emergent Ability)
  - generated video | GT Video | depth | multiid

A young girl with blonde hair styled in pigtails and wearing a pink sweater and white shorts is energetically dancing in a playroom filled with colorful furniture and toys. She is accompanied by a small, fluffy dog with a white and brown coat. The g

A woman enters a caf?? through a glass door, adjusting her scarf as she steps inside. She looks around, seemingly taking in the surroundings. As she moves further into the caf??, she glances out the window, observing the busy street outside.

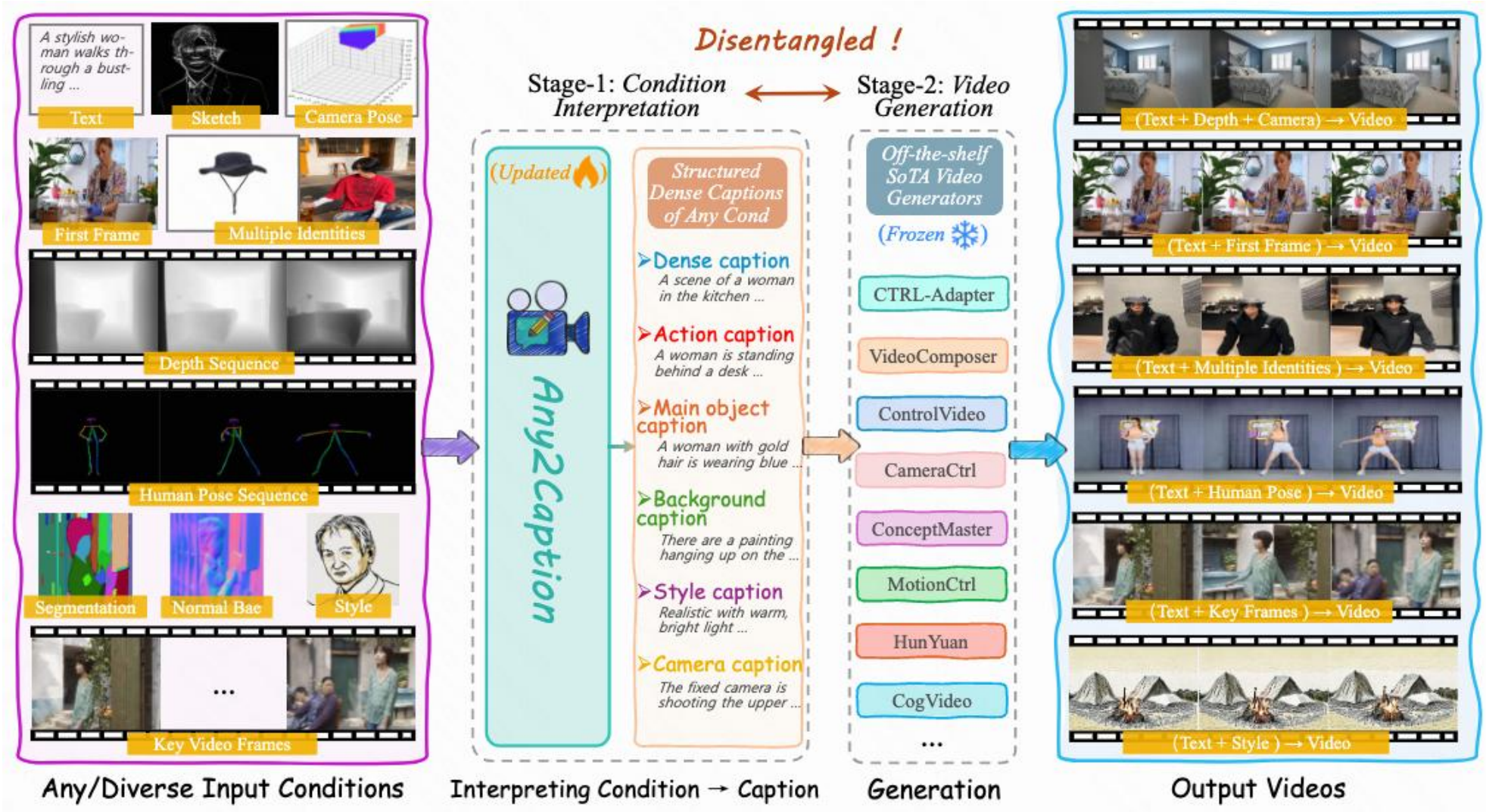
FullDiT: Video Generative Foundation Models with Multimodal Control via Full Attention

- Qualitative Results (Emergent Ability)
  - generated video | reference camara | depth

a kitchen with a stainless steel refrigerator and a microwave oven. The video showcases a smooth, continuous pan across a cozy, well-furnished living room. The camera moves from the kitchen area, through a doorway, and into the living room,

a staircase in a home with carpet on the steps. The video captures a serene and well-lit staircase in a modern home. The staircase is made of light-colored wood with a carpet runner, and it is flanked by wooden railings on both sides. The vide

Any2Caption: Interpreting Any Condition to Caption for Controllable Video Generation

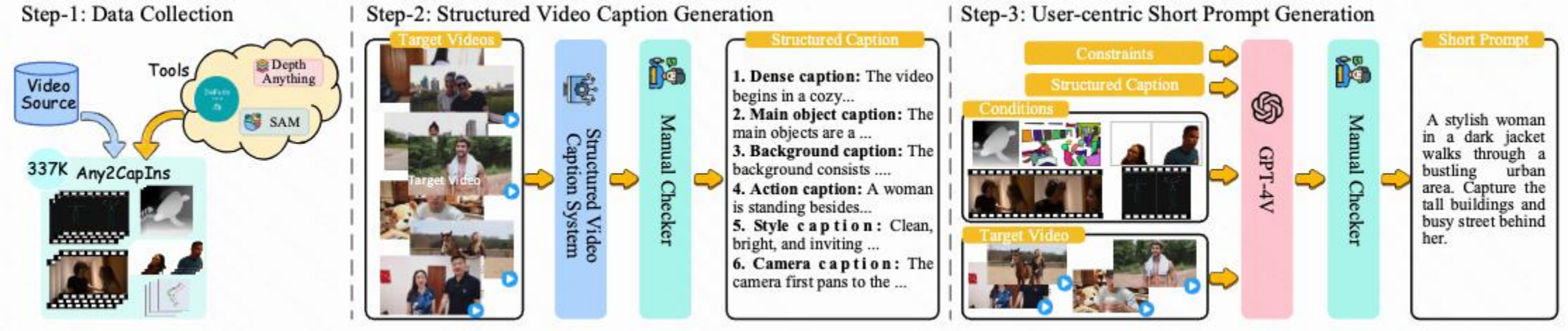


Any2Caption: Interpreting Any Condition to Caption for Controllable Video Generation

- Motivation
  - Video Generation Models: high-quality videos with rich text captions;
  - MLLMs: robust vision-language comprehension.
  - **Challenge: accurately interpreting user intention**
    - short user prompt
    - non-text conditions
  - Core Idea: decouple interpreting various conditions->Video Generation
    - New Dataset
    - New Model
    - New Benchmark
    - Sota performance

Any2Caption: Interpreting Any Condition to Caption for Controllable Video Generation

- Dataset Construction Pipeline
  - Step-1: Data Collection:
    - Spatial-wise, Action-wise, Composition-wise, Camera-wise conditions
  - Step-2: Structured Video Caption Generation
  - Step-3: User-centric Short Prompt Generation
    - Conciseness and Simplicity
    - Condition-Dependent Omission
    - Implicit instruction of Target Video



Any2Caption: Interpreting Any Condition to Caption for Controllable Video Generation

- Dataset Construction Pipeline
  - prompt GPT-4V to generate short prompt

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 maps 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 8. Demonstration of the prompt used for GPT-4V to generate the short prompt when the input condition is the depth.

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

Any2Caption: Interpreting Any Condition to Caption for Controllable Video Generation

- Dataset Construction Pipeline
  - caption statistics

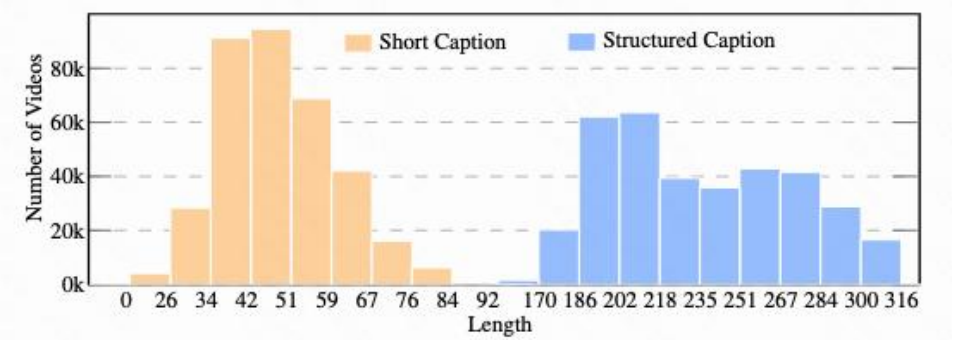


Figure 3. **Distribution of the short/structured caption length** (in words) in `Any2CapIns`.

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.** means the number of instances, and **#Condition** denotes the number of unique conditions. **#Avg. / #Total Len.** indicate the average and total video durations, respectively.

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

Any2Caption: Interpreting Any Condition to Caption for Controllable Video Generation

- Dataset Construction Pipeline
  - Visualization of Short and Structured Caption

Short Caption + Camera → Video

➢ **Short Caption**: A cozy and well-lit home. Start by showing a dining table with chairs and a chandelier, then capture the living room with a sofa and fireplace. Move towards the large windows to reveal the deck outside and the grassy area. Emphasize the warm and inviting atmosphere.

➢ **Structured Caption**:

1. Dense caption: The video showcases a cozy, well-lit dining and living room area. Initially, the camera captures a dining table with chairs, a chandelier, and a view of the living room with a sofa and fireplace. Gradually, the camera moves towards the large windows, revealing a deck outside. The scene transitions smoothly, emphasizing the warm, inviting atmosphere of the interior and the serene outdoor view.
2. Main object caption: The main objects include a dining table with chairs, a chandelier, a sofa, a fireplace, and large windows. The camera moves from the dining area to the windows.
3. Background caption: The background features a cozy living room with a sofa, fireplace, and a clock on the wall. The windows reveal a deck and a grassy outdoor area. The weather appears clear, and the scene transitions from an indoor view to an outdoor view through the windows.
4. Action caption: The camera moves to the left, shooting the interior of the house, and then the camera moves forward to photograph the outside of the house.
5. Style caption: Warm, inviting, and homely with soft lighting and rich colors.
6. Camera caption: The camera first pans to the left, then moves forward while panning to the left. The camera is roughly at the same height as the subject, maintaining a wide shot of the interior.

Short Caption + Multi-IDs → Video

➢ **Short Caption**: A person stands in a vast field under a stormy sky. They raise their hands to their face, then above their head, transforming their headpiece into a horn-like structure. The camera moves backward, capturing the gloomy atmosphere and their confident stance.

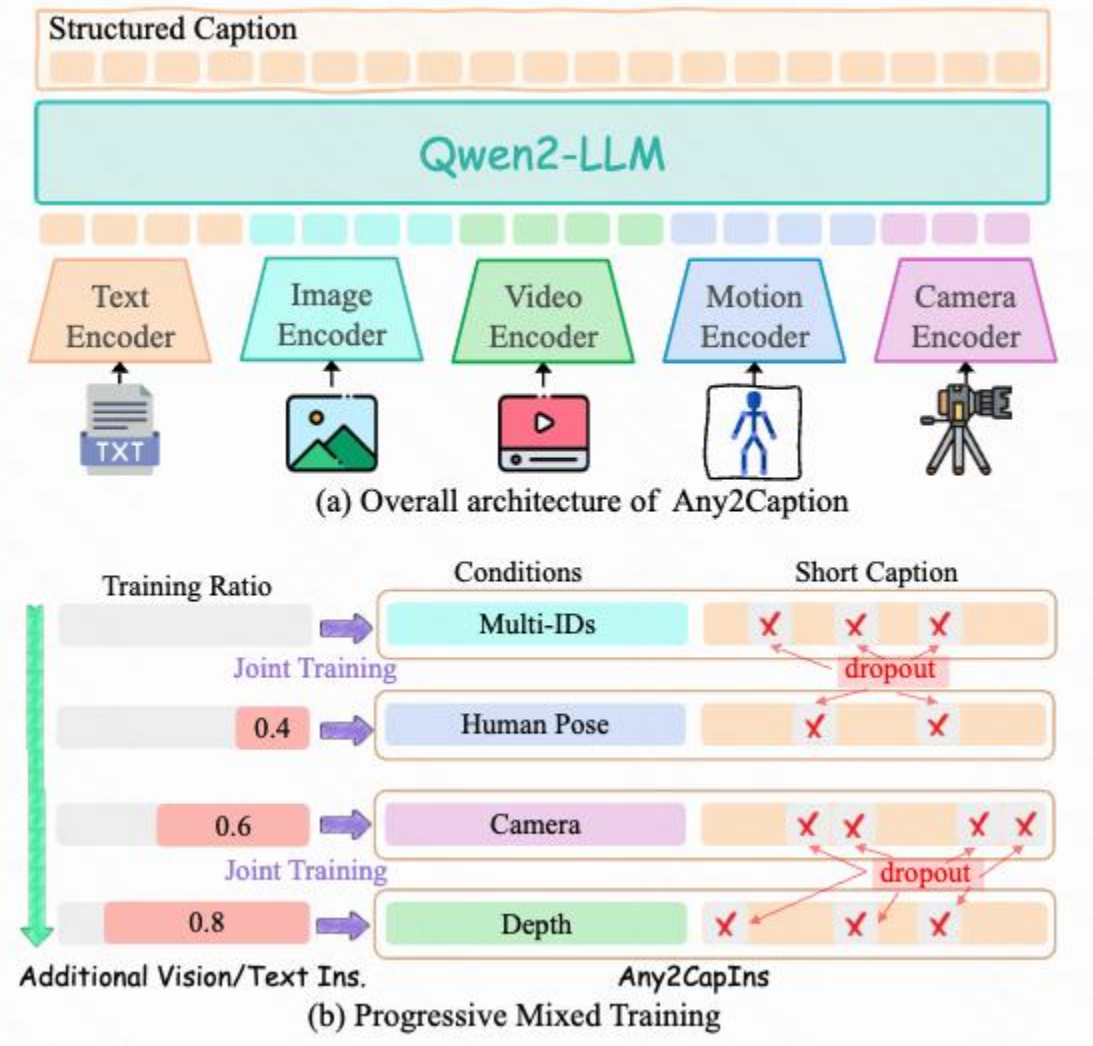
➢ **Structured Caption**:

1. Dense caption: A figure in a futuristic, armored suit stands in a vast, open field under a dark, stormy sky. Initially, the figure raises their hands to their face, then slowly raises them above their head. As the sequence progresses, the figure's headpiece transforms into a large, intricate, horn-like structure. The figure then lowers their hands and adopts a confident stance, holding two long, thin weapons in each hand.
2. Main object caption: The main subject is a person in a black, armored suit with intricate designs. They stand in the center of the frame, raising their hands and transforming their headpiece into a horn-like structure.
3. Background caption: The background is a wide, open field with green grass, surrounded by distant hills under a dark, stormy sky. The weather appears gloomy and foreboding, with heavy clouds looming overhead. The scene remains static, emphasizing the dramatic atmosphere.
4. Action caption: A woman in black clothes raised her hands and put them on her head, then put them down and walked forward with two swords in her hands.
5. Style caption: Dark, dramatic, and cinematic with a focus on the character's transformation and the stormy, desolate landscape.
6. Camera caption: The camera moves backward. The camera is roughly at the same height as the character, maintaining a medium shot of the character's full body. The character is facing the camera and is positioned in the center of the frame..

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

Any2Caption: Interpreting Any Condition to Caption for Controllable Video Generation

- Framework
  - Architecture: Qwen2VL
    - <|motion start|>, <|motion end|>,
    - <|camera start|>, <|camera end|>
  - Training Recipes
    - Alignment Learning
      - freeze llm and vision encoder
      - train camera coder with camera description
    - Condition-Interpreting Learning
      - progressive mixed training strategy
        - single condition->multi condition
        - new condition: incorporate vl instruction



Any2Caption: Interpreting Any Condition to Caption for Controllable Video Generation

- Evaluation Suite
  - Lexical Matching Score
    - BLEU, ROUGE, METEOR
    - Structural Integrity
  - Semantic Matching Score
    - BERTSCORE
    - CLIP Score
  - Intent Reasoning Score
    - User Intention Extraction
    - Ground-Truth QA Pair Construction
    - Predicted Answer Generation
    - Answer Evaluation
  - Video Generation Quality Score
    - Overall Quality
    - Camera: RotErr,TransErr, and CamMC
    - Depth: MAE
    - Identity: DINO-I, CLIP-I
    - Human Pose: Pose Acc.

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.
	What detail does the background highlight? The scientific setting with static emphasis.
Action	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.
Camera	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.
	How does the camera move in the video? It pans to the right.
Style	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.
	What kind of atmosphere does the laboratory have? Professional and scientific.

Table 9. Demonstration of generated question-answer pairs utilized in IRSCORE calculation.

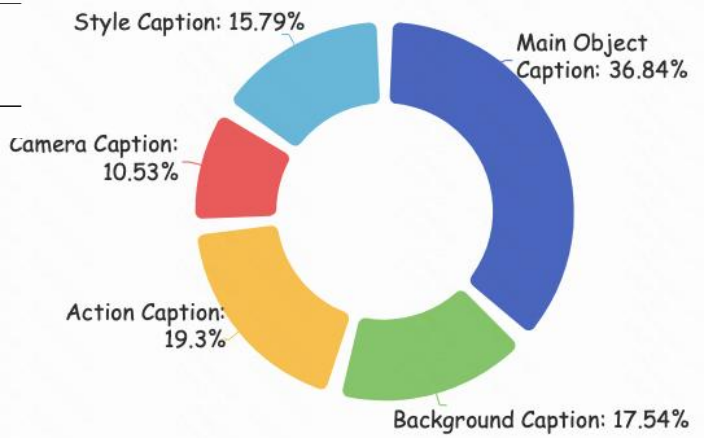


Figure 10. Question-answer pairs proportion in structured captions.

Any2Caption: Interpreting Any Condition to Caption for Controllable Video Generation

- Evaluation Suite
  - Lexical Matching Score
    - BLEU, ROUGE, METEOR
    - Structural Integrity
  - Semantic Matching Score
    - BERTSCORE
    - CLIP Score
  - Intent Reasoning Score
  - Video Generation Quality Score
    - Overall Quality
    - Camera: RotErr,TransErr, and CamMC
    - Depth: MAE
    - Identity: DINO-I, CLIP-I
    - Human Pose: Pose Acc.

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 2. 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 individual component generation performance within the structure. "B-2" and "R-L" denotes BLEU-2 and ROUGE-L, respectively.

Caption Enrich	Text	Video Generation		
	CLIP-T↑	Smoothness↑	Aesthetic↑	Integrity↑
Short Cap.	18.31	93.46	5.32	55.39
Short Cap. w/ Condition Cap.	19.19	93.41	5.41	54.91
Structured Cap.	19.87	94.38	5.46	57.47

Table 3. Quantitative results comparing short caption, short caption combined with condition caption, and structured caption for multi-identity video generation.

Any2Caption: Interpreting Any Condition to Caption for Controllable Video Generation

- Evaluation Suite
  - Lexical Matching Score
    - BLEU, ROUGE, METEOR
    - Structural Integrity
  - Semantic Matching Score
    - BERTSCORE
    - CLIP Score
  - Intent Reasoning Score
    - User Intention Extraction
    - Ground-Truth QA Pair Construction
    - Predicted Answer Generation
    - Answer Evaluation
  - Video Generation Quality Score
    - Overall Quality
    - Camera: RotErr,TransErr, and CamMC
    - Depth: MAE
    - Identity: DINO-I, CLIP-I
    - Human Pose: Pose Acc.

Model	Text	Camera			Identities		Depth	Human Pose	Overall Quality			
	CLIP-T↑	RotErr↓	TransErr↓	CamMC↓	DINO-I↑	CLIP-I↑	MAE↓	Pose Acc.↑	Smoothness↑	Dynamic↑	Aesthetic↑	Integrity↑
• Camera to Video												
MotionCtrl [60]	19.67	1.54	4.49	4.80	-	-	-	-	96.13	9.75	5.40	73.69
+ Structured Cap.	20.16	1.45	4.37	4.78	-	-	-	-	96.16	11.43	5.71	74.63
CameraCtrl [21]	18.89	1.37	3.51	4.78	-	-	-	-	94.11	12.59	4.26	71.84
+ Structured Cap.	21.70	0.94	2.97	4.37	-	-	-	-	95.16	13.72	4.66	72.47
• Depth to Video												
Ctrl-Adapter [40]	20.37	-	-	-	-	-	25.63	-	94.53	20.73	4.63	46.98
+ Structured Cap.	23.30	-	-	-	-	-	21.87	-	95.54	15.14	5.31	54.20
ControlVideo [74]	22.17	-	-	-	-	-	30.11	-	92.88	5.94	5.29	63.85
+ Structured Cap.	24.18	-	-	-	-	-	23.92	-	94.47	18.27	9.77	66.28
• Identities to Video												
ConceptMaster [25]	16.04	-	-	-	36.37	65.31	-	-	94.71	8.18	5.21	43.68
+ Structured Cap.	17.15	-	-	-	39.42	66.74	-	-	95.05	10.14	5.68	49.73
• Human Pose to Video												
FollowYourPose [45]	21.11	-	-	-	-	-	-	30.47	91.71	14.29	4.95	58.84
+ Structured Cap.	21.39	-	-	-	-	-	-	31.59	92.87	16.47	5.88	56.30

Table 5. Performance comparison on four types of conditions (e.g., *camera*, *depth*, *identities* and *human pose*) between directly using short captions and integrating structured captions under various video quality evaluation metrics.

Compositional Condition	Text	Camera			Identities		Depth	Overall Quality			
	CLIP-T↑	RotErr↓	TransErr↓	CamMC↓	DINO-I↑	CLIP-I↑	MAE↓	Smoothness↑	Dynamic↑	Aesthetic↑	Integrity↑
Camera+Identities	14.81	1.37	4.04	4.24	25.63	64.14	-	94.43	28.87	4.99	59.81
+ Structured Cap.	19.03	1.49	4.41	4.84	26.75	68.45	-	94.38	34.99	5.25	63.02
Camera+Depth	20.80	1.57	3.88	4.77	-	-	25.37	95.36	30.12	4.82	63.90
+ Structured Cap.	21.19	1.49	4.41	4.84	-	-	32.15	95.40	30.10	4.96	65.05
Depth+Identities	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
Camera+Identities+Depth	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 of structured captions when handling compositional conditions.

Any2Caption: Interpreting Any Condition to Caption for Controllable Video Generation

Ablation Study

Training Strategy	Caption		Vieo Generation		
	B-2↑	Accuracy↑	Smoothness↑	Dynamics↑	Aesthetic↑
Any2Caption	47.69	67.35	94.60	17.67	5.53
w/o Two-Stage	33.70	51.79	93.31	16.36	5.50
w/o Dropout	49.24	69.51	94.16	14.54	5.51

Table 4. Ablation study on training strategy. "w/o stage" means alignment learning is not applied during training, while "w/o Dropout" denotes that short captions are not randomly dropped.

Generalization Capability

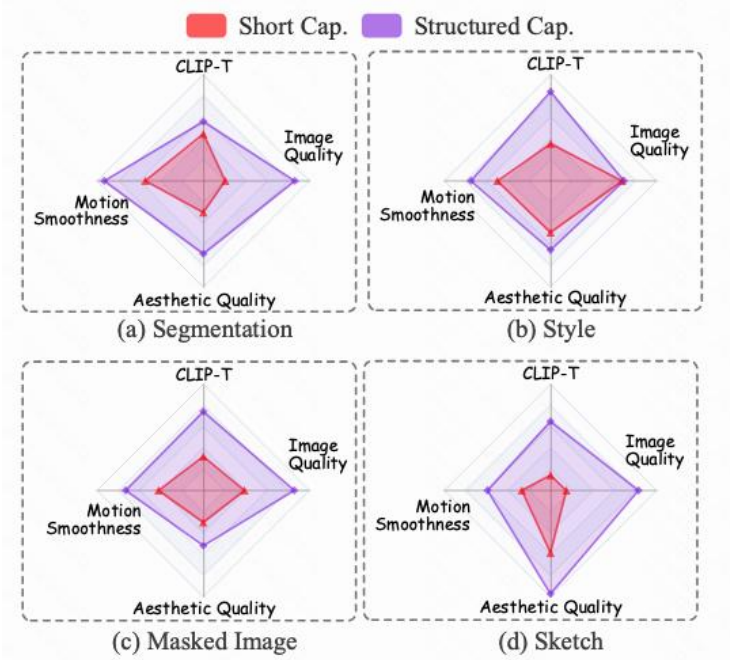
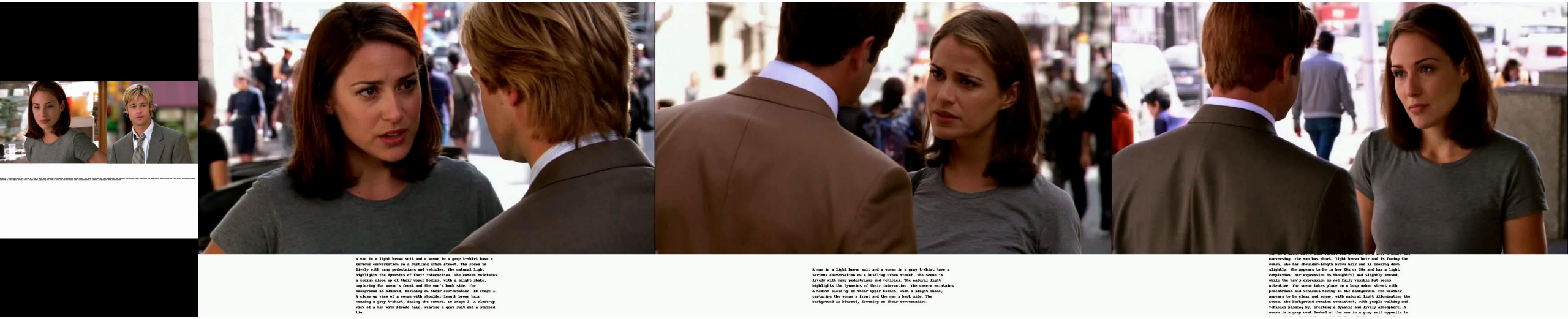


Figure 6. Quantitative results on unseen conditions (i.e., *segmentation* [58], *style* [68], *masked image* [58], and *sketch* [58]) when using short and structured captions, respectively.

Any2Caption: Interpreting Any Condition to Caption for Controllable Video Generation

- Qualitative Results
  - multiid | image-caption+short prompt | short prompt | Any2Caption



Any2Caption: Interpreting Any Condition to Caption for Controllable Video Generation

- Qualitative Results
  - multiid | image-caption+short prompt | short prompt | Any2Caption

Any2Caption: Interpreting Any Condition to Caption for Controllable Video Generation

- Qualitative Results
  - multiid | image-caption+short prompt | short prompt | Any2Caption

Any2Caption: Interpreting Any Condition to Caption for Controllable Video Generation

- Qualitative Results
  - short prompt vs Any2Caption
  - video | depth | multiid

Young woman with cat ears holding a white cat in a sunlit meadow. Blue butterflies flutter around. Gentle caresses and a serene, magical atmosphere. Trees cast a warm glow. Fixed camera at eye level capturing upper body. Whimsical, enchantir

1. Overall description: A serene video depicts a young woman with long, dark hair and cat ears, holding a white cat in a lush, sunlit meadow. Throughout the video, blue butterflies flutter around her, adding a magical touch to the scene. The wom

Summary

- FullDiT:
  - Unified Framework
  - Higher performance
  - Scalablity & Emergent Ability
  - Effective Training policy
- Any2Caption
  - user intent interpretation
  - interprets diverse condition into dense, structured captions
  - improve video quality across various sota video generation models

Q&A