

ETVA: Evaluation of Text-to-Video Alignment via Fine-grained Question Generation and Answering

Supplementary Material

7. Details on Evaluation Categories

The prompts were divided into 10 categories: existence, action, material, spatial, number, shape, color, camera, physics, other. Figure 5 shows that ETVABench-2k and ETVA-105 generate questions with similar distributions across these categories, indicating consistent coverage. Table 5 shows question example and prompt example for each category.

8. Detailed results on ETVABench-2k

More details of 10 open-source T2V models is in Table 6.

9. Details of Text-to-Video Models

Sora Sora [3] is a state-of-the-art text-to-video model built upon the DiT [41] architecture. As one of the most advanced and widely discussed video generation models, Sora is developed by OpenAI, though many of its underlying details remain undisclosed. Notably, OpenAI has not released an API for Sora, restricting video generation to browser-based access, which is both resource-intensive and inconvenient. In this study, we adopt a 16:9 aspect ratio at 480p resolution to generate 5-second video samples.

Vidu Vidu [49], a text-to-video (T2V) diffusion model developed by Shengshu Technology, integrates advanced semantic comprehension with dynamic shot composition capabilities to achieve hierarchical video synthesis across resolutions ranging from low-definition to 1080p. Our experimental framework employed Vidu 1.5 to generate 4-second video sequences at 720p resolution (16:9 aspect ratio), quantitatively assessing its dual capacity for contextual fidelity and cinematographic control.

Pika Pika [48], a proprietary video synthesis model developed by Pika Labs, demonstrates versatile capabilities in both video generation and multimodal editing across diverse visual styles. For experimental validation, we utilized the Pika 1.5 implementation to synthesize 5-second video sequences at a 16:9 aspect ratio, systematically evaluating its capacity to preserve temporal consistency and stylistic fidelity under standardized conditions.

Kling Kling [19] is a series of proprietary video generation models developed by Kuaishou. It is built on the Diffu-

sion Transformer (DiT) architecture and has demonstrated exceptional capabilities in generative tasks. For our experiments, we utilized Kling-1.0 and Kling-1.5, both standard models with a 16:9 aspect ratio, to generate 5-second video sequences.

Hunyuan-Video Hunyuan-Video [57], developed by Tencent, is currently the most advanced open-source T2V model. It features a massive 13 billion parameters architecture and is trained using the flow matching [27] method on a hierarchically structured, high-fidelity video dataset. For our experiments, we set the resolution to 544×966, generating a 5-second video consisting of 121 frames with 24 fps.

Mochi Mochi [46], a text-to-video (T2V) diffusion model developed by Genmo, demonstrates significant advancements in video synthesis through its 10-billion-parameter architecture. Preliminary evaluations indicate exceptional performance in motion fidelity and textual prompt alignment, substantially reducing the quality disparity between proprietary and open-source video generation systems. For experimental validation, we generated a 2-second video sequence (61 frames at 24 fps) with 480×848 pixel resolution, effectively demonstrating the model’s temporal coherence and detail preservation capabilities.

CogVideoX CogVideoX [60] is a large-scale open-source T2V model released by Zhipu, available in three versions: CogVideoX-2B, CogVideoX-5B, and CogVideoX-1.5-5B. It incorporates a 3D causal VAE and an expert transformer, enabling the generation of coherent, long-duration, and high-action videos. For CogVideoX-2B and CogVideoX-5B, we set the frame rate to 8 fps, generating a 6-second video with a total of 49 frames at a resolution of 720×480. For CogVideoX-1.5-5B, we used a resolution of 1360×768 at 16 fps, producing a 5-second video with 91 frames.

OpenSora OpenSora [65] is a high-quality DiT-based text-to-video model that introduces the ST-DiT-2 architecture. It supports flexible video generation with varying aspect ratios, resolutions, and durations. The model is trained on a combination of images and videos collected from open-source websites, along with a labeled self-built dataset. We utilized the officially released OpenSora 1.2 code and model, setting the spatial resolution to 720p and the frame rate to 24 fps, producing a 4-second (96-frame) video.

Prompt Category Analysis

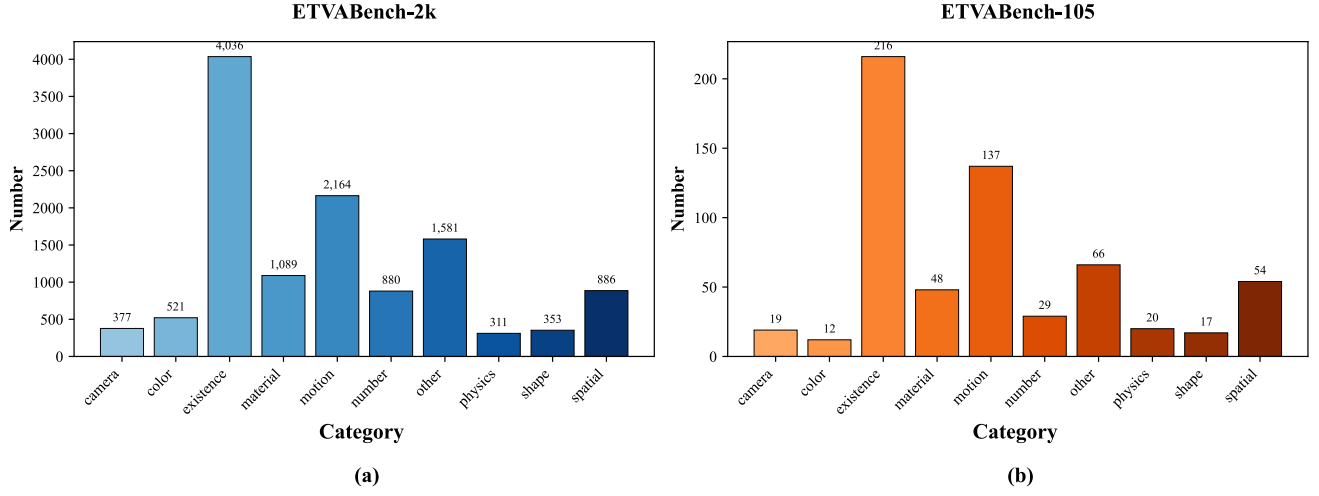


Figure 5. Prompt Category Distribution of ETVABench-2k and ETVABench-105

Category	Question Example	Prompt Example
Existence	<i>Is there a penguin in the video?</i>	<i>A penguin standing on the left side of a cactus in a desert.</i>
Action	<i>Does the player pass the football?</i>	<i>In a crucial game moment, a player passes the football, dodging opponents.</i>
Material	<i>Is the city made of crystals?</i>	<i>A city made entirely of glowing crystals that change colors based on emotions.</i>
Spatial	<i>Does the penguin stand on the left of the cactus?</i>	<i>A penguin standing on the left side of a cactus in a desert.</i>
Number	<i>Are there three owls in the video?</i>	<i>Three owls perch on a branch during a full moon.</i>
Shape	<i>Is the cloud shaped like a hand?</i>	<i>A cloud shaped like a giant hand that picks up people for transportation.</i>
Color	<i>Does the man's hair brown?</i>	<i>There's a person, likely in their mid-twenties, with short brown hair.</i>
Camera	<i>Is the camera pushing in?</i>	<i>A girl is walking forward, camera push in.</i>
Physics	<i>Is the water pouring out in the space station?</i>	<i>Water is slowly pouring out of glass cup in the space station.</i>
Other	<i>Is the video in the Van Gogh style?</i>	<i>A beautiful coastal beach waves lapping on sand, Van Gogh style.</i>

Table 5. Question Example and Prompt Example for Each Category.

Opensora-plan OpenSoraPlan [25] is an advanced video generation model built upon Latte [34]. It replaces the Image VAE [17] with Video VAE [25] (CausalVideoVAE), following a similar approach to Sora [3]. For OpenSoraPlan v1.1, we used the 65-frame version with a spatial resolution of 512×512 at 16 fps, generating a 4-second video. For OpenSoraPlan v1.2, we selected the 93-frame version with a resolution of 720p at 16 fps, producing a 5-second video.

Vchitect Vchitect [10], an open-source text-to-video (T2V) generative model developed by Shanghai AI Lab, is built upon the Diffusion Transformer (DiT) [41] architecture. With 2 billion parameters, the model demonstrates robust capabilities in generating high-quality video content. For experimental validation, we synthesized a 5-second video sequence comprising 40 frames at 8 frames per second (fps), with a resolution of 768×432 pixels.

Latte Latte [34] is an early open-source DiT-based text-to-video model, built upon PixArt-Alpha with extended

spatiotemporal modules and further training. We employed the officially released LatteT2V code and model, preserving the original parameter settings. For video generation, we used a spatial resolution of 512×512, a frame rate of 8 fps, and a duration of 2 seconds (16 frames).

10. Details of Human annotation

Figure 6 shows detailed instructions for human annotation.

11. Prompts of ETVA

Figure 7 - Figure 11 present the detailed prompts in ETVA.

12. More Cases about ETVA

Figure 12 and Figure 13 show the detailed comparison results between ETVA and conventional evaluation metrics.

Model	Existence	Motion	Material	Spatial	Number	Shape	Color	Camera	Physics	Other	Avg
<i>Open-Source T2V Models</i>											
Latte	0.505	0.504	0.538	0.558	0.495	0.567	0.563	0.536	0.490	0.529	0.519
OpenSora-plan-1.1	0.534	0.526	0.496	0.568	0.522	0.504	0.522	0.488	0.397	0.568	0.529
OpenSora-plan-1.2	0.590	0.585	0.576	0.635	0.590	0.611	0.595	0.582	0.545	0.610	0.601
OpenSora-1.2	0.666	0.685	0.626	0.688	0.637	0.655	0.623	0.666	0.526	0.667	0.660
Cogvideox-2B	0.679	0.670	0.615	0.713	0.658	0.655	0.627	0.629	0.600	0.691	0.668
Vchitect-2.0	0.686	0.691	0.635	0.731	0.681	0.712	0.652	0.668	0.545	0.700	0.682
CogvideoX-5B	0.700	0.700	0.648	0.769	0.710	0.670	0.664	0.671	0.584	0.698	0.694
CogvideoX-1.5-5B	0.704	0.714	0.686	0.716	0.722	0.670	0.637	0.674	0.606	0.727	0.702
Mochi-1-preview	0.736	0.735	0.666	0.771	0.705	0.726	0.656	0.668	0.548	0.745	0.720
Hunyuan	0.753	0.756	0.699	0.810	0.747	0.761	0.726	0.695	0.632	0.762	0.748

Table 6. ETVA-Bench-2k evaluation results with 10 open-source T2V models and apple_video. A higher score indicates better performance for a dimension. **Bold** stands for the best score.

Instruction for Human Annotation

Task1:

This project aims to evaluate the text-video consistency of existing text-to-video generation models. Given a textual description and its corresponding generated video, annotators are required to assess the alignment between the textual input and the visual output. The evaluation task comprises 105 samples, each containing 14 video outputs generated by different models, stored in individual subdirectories. The videos are named sequentially as 1.mp4, 2.mp4, ..., 14.mp4, with the correspondence between model identifiers and video names randomized to ensure unbiased evaluation.

Evaluation Metric: The text-video consistency is scored on a scale from 0 to 5, with increments of 0.5. A score of 5 indicates perfect alignment between the generated video and the textual description, while a score of 0 denotes complete inconsistency. It is crucial to note that the assessment should focus exclusively on the semantic consistency between text and video, disregarding other factors such as video quality, resolution, or visual clarity.

Evaluation Procedure: As a demonstration, Sample 1 is provided as an illustrative example. Annotators are instructed to evaluate subsequent samples following the same scoring protocol as demonstrated in Sample 1.

Task2:

Following the assessment of text-video consistency scores, you are now required to evaluate the accuracy of specific questions derived from the textual descriptions. These questions are generated based on the input text, and your task is to determine whether the video content correctly answers each question. Please respond with "1" for correct and "0" for incorrect, strictly adhering to the evaluation protocol established in Sample 1.

Figure 6. Instruction for Human Annotation

Prompt for *Element Extractor*

System Prompt: You are a professional video question generation assistant specializing in structured data extraction. Your task is to meticulously analyze input prompts according to the following three-phase protocol:

User Prompt:

Task: given input prompts, extract the prompt background, camera and entities from the prompts.

Do not generate same entities multiple times. Do not generate entities that are not present in the prompts. Just output the entities and do not output other things.

output format: Background | "background"

Camera | "camera"

id | entity

Task: given input prompts, extract the attributes from the prompts.

Attributes are intrinsic characteristics of an entity and should not contain external entities that can be divided. Do not generate same attributes multiple times. Do not generate attributes that are not present in the prompts. Do not generate other entities as attributes. If no attribute is present, output "no mention".

output format: attribute | value

Task: given input prompts and entity, extract the relations between entities from the prompts. Notice that the relations are at least between two entities and if there is only one entity, output "no mention".

Do not generate same relations multiple times. Do not generate relations that are not present in the prompts.

output format: id | relation

Figure 7. Prompt for *Element Extractor*.

Prompt for *Graph Builder*

System Prompt: As a graph builder, your task is to construct a structured and semantically rich scene graph by systematically organizing the extracted entities, attributes, and relations. Your goal is to transform these components into a coherent and hierarchical representation that captures the underlying structure and relationships within the given data.

User Prompt:

You are tasked with constructing a structured mapping between entities and their corresponding attributes based on the provided input. Your goal is to analyze the given prompt , entities , and attributes , and accurately match each attribute to its associated entity. Follow these steps precisely: Carefully examine the prompt to understand the context and relationships between the entities and attributes. Identify the specific entity that each attribute describes or modifies. . Ensure that all mappings are accurate and contextually relevant, avoiding any mismatches or omissions. Output the results in the following strict format:

entity | attribute

Your task is to analyze the provided prompt , entities , and relations to construct a structured representation of the relationships between entities. Follow these steps precisely: Carefully examine the prompt to understand the context and identify how the entities are connected through the given relations. For each relation , determine the two entities it connects, ensuring the pairing is contextually accurate and meaningful. Ensure that all relationships are logically consistent and directly derived from the input data.

Output the results in the following strict format:

entity1 | relation | entity2 .

Figure 8. Prompt for *Graph Builder*.

Prompt for *Graph Traverser*

System Prompt: You are a research worker with excellent paper reading skills.

User Prompt:

Task: You are a helpful question generator for video. You are asked to generate questions based on the input video prompts and related entities, attributes and relations. Please ask questions as the format of examples. All the questions may can be answered by yes or no.

output format: question

Example 1:

prompt: During harvest, a bear rampages through a cornfield, stalks collapsing in waves. Film a group of skateboarders tearing through an urban skatepark, performing flips, grinds, and tricks with lightning-fast agility.

question_type : content: Background | Harvest

question: Is the video background in the scene of Harvest?

Example 2:

prompt: : A young man is riding a bicycle. He is wearing a blue hoodie and black jeans. His hair is brown and styled messily. He has a small scar above his left eye.

question_type : attribute

content: man | hair color brown

question: Is the hair color of the man brown?

Example 3:

prompt: A young man is riding a bicycle. He is wearing a blue hoodie and black jeans. His hair is brown and styled messily. He has a small scar above his left eye.

question_type : relation (entity, attribute, relation)

content: man | riding | bicycle

question: Is the man riding a bicycle?

Example 4:

prompt: {prompt}

question_type: {type}

content: {content}

question:

Figure 9. Prompt for *Graph Traverser*.

Prompt for *Knowledge Augmentation*

System Prompt: You are a helpful assistant tasked with analyzing video text prompts.

User Prompt:

Upon receiving a text prompt, your objective is to extract and identify implicit knowledge—such as common sense or relevant physical principles—that is not explicitly stated in the prompt but is necessary for generating an accurate and realistic video. You should present this knowledge as examples below:

Example 1:

prompt: A cyclist rides through a bustling city street during rush hour, weaving between pedestrians and navigating around parked cars.

Implicit knowledge: 1.Kinetic Friction and Balance:

Video Representation: Show the cyclist maintaining balance by subtly shifting their body weight and adjusting the handlebars, especially when navigating tight spaces or making sharp turns.

2.Reaction Time and Decision Making:

Video Representation: Depict the cyclist making quick decisions, such as braking suddenly to avoid a pedestrian or swerving to bypass an obstacle, illustrating the importance of reaction time in traffic.

3.Traffic Flow and Human Behavior:

Video Representation: Illustrate the flow of traffic and pedestrian movement, showing how the cyclist anticipates and responds to the actions of others, such as stopping at a crosswalk or merging into traffic lanes.

4.Sound and Environmental Cues:

Video Representation: Incorporate ambient city sounds like honking horns, footsteps, and the cyclist's breathing or gear shifting to convey the dynamic and potentially chaotic environment.

Example 2:

prompt: A cup of water is slowly poured out in the space station, releasing the liquid into the surrounding area.

Implicit knowledge: 1.Microgravity Behavior of Liquids:

Video Representation: Show water separating from the cup into spherical droplets that float freely in the cabin instead of falling to the floor.

2.Surface Tension Effects

Video Representation: Highlight close-up shots of water droplets retaining their round shape as they detach from the stream.

3.Absence of Air Resistance

Video Representation: Depict droplets drifting gently across the cabin, moving in straight paths without slowing down quickly.

Example 3:

prompt: {prompt}

Implicit knowledge:

Figure 10. Prompt for *Knowledge Augmentation*.

Prompt for *Multi-stage Reasoning*

System Prompt: You are a helpful video assistant. Now you are watching a video and given an answer to the question.

User Prompt:

You are a multimodal understanding assistant. You have access to the following:

- 1.Text (Input Description): A textual prompt that was used to generate the video.
- 2.Additional Common Sense Knowledge: Information that is not directly in the text but should be implicitly present in the video (e.g., logical or contextual details).
- 3.Video Clip: A generated video that you should analyze based on the text prompt and common sense.

Your task is to answer a video-related question, but first, engage in thorough reflection by considering the following steps:

- 1.Video Understanding: Analyze how the video matches the provided text and common sense. Does the video show what is expected based on the text? Are there any discrepancies or missing elements that should logically be present in the video?
- 2.Critical Reflection: Think through the implications of the video content in light of the text and common-sense knowledge. Does the video fully align with what was described, or are there gaps?
- 3.Conclusion: You should answer [YES] or [NO] special answer token to the question based on your analysis and provide a brief explanation to support your answer.

Here, the text prompt is: {prompt}

The common sense knowledge is: {common sense}

And the question is: {question}

Please finish the Video Understanding, Critical Reflection, and Conclusion stages with [YES] or [NO] special answer token.

Figure 11. Prompt for *Multi-stage Reasoning*.

Text input: Water is slowly pouring out of glass cup in the **space station**.

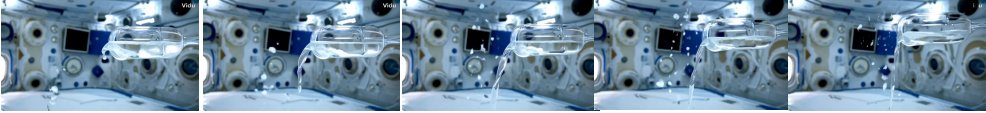




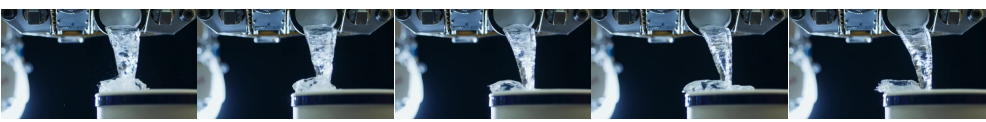



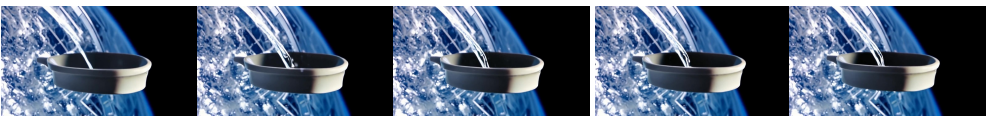
Vidu		BLIP_BLEU: 0.264 CLIPScore : 0.387 VideoScore : 2.110 ETVA : 0.750
Kling-v1.5		BLIP_BLEU: 0.163 CLIPScore : 0.384 VideoScore : 1.960 ETVA : 0.375
Sora		BLIP_BLEU: 0.081 CLIPScore : 0.374 VideoScore : 1.690 ETVA : 1.000
Pika		BLIP_BLEU: 0.194 CLIPScore : 0.361 VideoScore : 1.870 ETVA : 0.875
Klingv1		BLIP_BLEU: 0.176 CLIPScore : 0.373 VideoScore : 2.004 ETVA : 0.500
Hunyuan		BLIP_BLEU: 0.183 CLIPScore : 0.320 VideoScore : 2.006 ETVA : 0.375
CogVidex-5B		BLIP_BLEU: 0.132 CLIPScore : 0.366 VideoScore : 2.330 ETVA : 0.375
CogVidex-2B		BLIP_BLEU: 0.203 CLIPScore : 0.321 VideoScore : 2.010 ETVA : 0.375
Mochi		BLIP_BLEU: 0.139 CLIPScore : 0.323 VideoScore : 2.213 ETVA : 0.625
OpenSora		BLIP_BLEU: 0.264 CLIPScore : 0.318 VideoScore : 2.271 ETVA : 0.250

Figure 12. Illustration of a comparative analysis between ETVA and conventional evaluation metrics, based on the text prompt: “Water is slowly pouring out of glass cup in the space station”. We compare our ETVA score with conventional text-to-video alignment metrics.

Text input: A leaf turns from green to red.











Vidu		BLIP_BLEU: 0.177 CLIPScore : 0.343 VideoScore : 2.434 ETVA : 1.000
Kling-v1.5		BLIP_BLEU: 0.268 CLIPScore : 0.347 VideoScore : 2.477 ETVA : 0.500
Sora		BLIP_BLEU: 0.201 CLIPScore : 0.378 VideoScore : 1.946 ETVA : 0.250
Pika		BLIP_BLEU: 0.212 CLIPScore : 0.344 VideoScore : 1.767 ETVA. : 0.500
Klingv1		BLIP_BLEU: 0.157 CLIPScore : 0.331 VideoScore : 2.354 ETVA. : 0.250
Hunyuan		BLIP_BLEU: 0.177 CLIPScore : 0.351 VideoScore : 2.545 ETVA. : 1.000
CogVidex-5B		BLIP_BLEU: 0.128 CLIPScore : 0.324 VideoScore : 2.713 ETVA. : 0.250
CogVidex-2B		BLIP_BLEU: 0.277 CLIPScore : 0.321 VideoScore : 2.707 ETVA. : 0.500
Mochi		BLIP_BLEU: 0.277 CLIPScore : 0.384 VideoScore : 2.590 ETVA. : 0.250
Opensora		BLIP_BLEU: 0.264 CLIPScore : 0.339 VideoScore : 2.666 ETVA : 0.250

Figure 13. Illustration of a comparative analysis between ETVA and conventional evaluation metrics, based on the text prompt: "A leaf turns from green to red". We compare our ETVA score with conventional text-to-video alignment metrics.