UniVBench: Towards Unified Evaluation for Video Foundation Models

Supplementary Material

A. Evaluation Cases

In this section, we presents the evaluation resulst in different tasks. In Figure A1 and A2, we present the source video and the reference captions we provided, along with generated video from CogVideoX, OmniVideo and Wan2.2-15B. In Figure A3, we present the results of reference images to video generation by Seedance-Lite.

From the rows of images, we can see that current video generation models still struggle to meet the text requirements. In Figure A1, the two animals enter the frame and walk to the front of the camera and wave hands are not captured by CogVideoX and OmniVideo. In Figure A2, the dinosaur-shaped pet bed opens when the cat enters. CogVideoX and OmniVideo's results didn't conform to it. In Figure A3, the referenced subject has serious identity shift when cut to the next shot. These qualitative results show that current video generation models still have large room for improvements.

B. More Details of UniVBench

B.1. Captioning Meta Data Distribution

In Figure B4, we provided the video content distribution across each sub-dimensions. This indicates that our dataset is semantically rich and diverse.

B.2. Referenced Websites

We referred to the following websites to formalize our camera_info:

- https://www.studiobinder.com/blog/{*film-lighting-techniques*, *how-to-use-color-in-film-50-examples-of-movie-color-palettes*, *ultimate-guide-to-camera-shots/*}
- https://www.epidemicsound.com/blog/different-camera-shots/
- https://boords.com/blog/16-types-of-camera-shots-and-angles-with-gifs

B.3. Captioning Prompt

In this subsection, we release our system prompts to generate dense video captions for our benchmark construction. They are shown in Figure B7 to B13. The prompts are divided into two steps: First, the model is tasked to extract the necessary content attributes from the video. This can include: subjects, actions, background, camera information, color, lighting, video style, etc., Then, the model merges them together to generate a coherent and structured video script, the format is shown in Figure B5. The essence of a video script format is: first describes the fixed, unchanging content, including the overall style and atmosphere of the video. Then, specify the information of the video's first frame, such as the subjects appearing in the first frame, their positions, and initial states. Subsequently, output subject actions, camera movements, and any changing information in chronological order—including adjustments to the relative positions of subjects and camera parameters. If the video is multi-shot, appending the keyword: Shot cut, and repeate the first frame description, subject actions, camera movements in chronological order.

C. Evaluation System Prompt

In this section, we provide a detailed description of the system prompts used in UniV-Eval, organized by task categories. Specifically, we present the system prompts corresponding to the six major tasks: V2T (Figure F14), T2V (Figure F15), R2V (Figure F16), TV2V (Figure F17), RV2V (Figure F18), and V2V (Figure F19).

It is important to note that, for the V2T task, the evaluation prompt must be used together with a predefined template (Figure F20), since the final comparison is conducted between the ground-truth caption and the baseline caption. For the other tasks, the comparison rules for generic objects are illustrated in Figure F21. In practice, these components should be combined to form the complete system prompt used for evaluation.

D. Evaluation Cost

Average cost of running one case is provided in Table D1. The cost of evaluating one task is less than 10 US dollars.

	V2V	TV2V	R2V	RV2V	T2V	V2T
I/O total tokens	25104	25898	16743	27534	19567	1413
Times (s)	45	62	44	55	49	27

Table D1. The cost of evaluation

E. Potential LLM-as-Judge Bias

Self-preference bias exists when the same LLMs act as both evaluatee and evaluator, they can recognize their own outputs and give higher scores, which is well-discussed in existing work [30]. In our settings, evaluatee and evaluators are different. The evaluatee models are video generation models, while the evaluator models are vision-language models. These two models differ significantly in both their architectural designs and training data.

Video Source

Video Captions:

The video is in a 3D cartoon style, presenting an atmosphere that transitions from initially mysterious to friendly and cute. It employs a long and wide-angle shot, a ground-level shot (level with the animals' eyes), and a deep depth of field (keeping everything from the foreground trash can to the background wooden door in sharp focus). The scene opens at night in a narrow, CG-generated alleyway where the sides are gray brick buildings, a wooden door with the words \"HAMCS PAINT\" is at the end of the alley, and an illuminated street lamp lights up part of the lane. The lighting is backlit and hard (the main light source comes from the street lamp behind the characters, casting sharp shadows in front of them), with a neutral color tone (no obvious warm or cool bias; black, white, gray), low saturation (the entire frame is in black and white), medium contrast, and low brightness (it's a night scene). In the left foreground is a silver metal trash can with a lid and four wheels, and on the ground in the center foreground is a rectangular metal drain cover. A German Shepherd with black, gray, and white fur, perked-up ears, a black collar, standing on all fours, and a raccoon with black, gray, and white fur, the signature black mask pattern on its face, standing on all fours, are in the distant center of the frame. The raccoon is to the German Shepherd's right, and they are both facing the camera. First, the camera remains static as the German Shepherd and the raccoon walk slowly side-by-side from the back of the alley towards the camera. The German Shepherd appears alert with its ears perked up, while the raccoon walks with a slightly bouncy gait. Their shadows stretch out long in front of them, and as they approach, the shot size gradually changes to a medium shot, during which the German Shepherd slowly sticks out its tongue. Then, as they stop in front of the camera, the German Shepherd slightly opens its mouth, tilts its head a little, and shows a curious and friendly expression. Finally, the raccoon also stops, raises its right front paw, lifts its eyebrows, and waves at the camera with a pleasant smile.

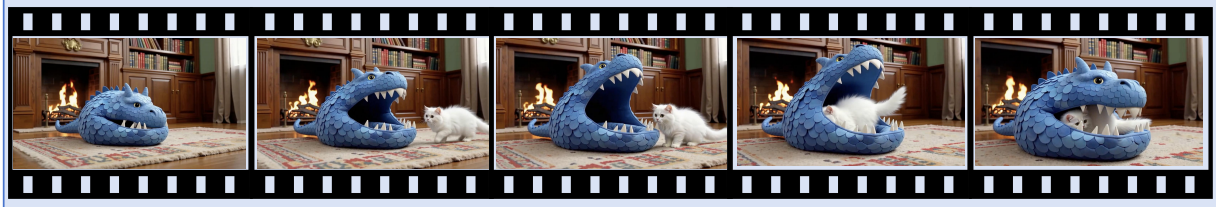
T2V Generated Video by (Wan2.2-14B)

T2V Generated Video by (CogVideoX)

T2V Generated Video by (OmniVideo)

Figure A1. Examples of T2V generation results across different baselines

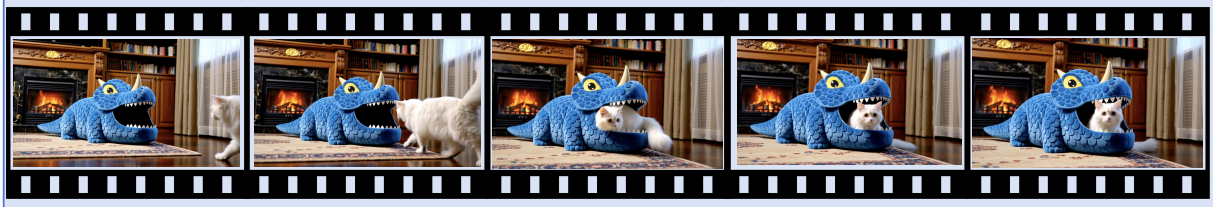
Video Source



Video Captions:

The video is in a realistic style, presenting an overall warm, cozy, and peaceful atmosphere, using a slow zoom-in shot, a medium-long shot, a horizontal angle, a ground-level height, and a shallow depth of field (focusing on the pet bed and the cat, with the background slightly blurred). The scene opens on a warm and cozy indoor study; the environment includes a brown wooden floor and a beige rug with red and blue patterns. Behind the rug is a brown wooden fireplace with exquisite wood carvings (with a golden ornament on the mantelpiece), where an orange flame flickers on the logs. To the right of the fireplace is a brown bookcase filled with various books, and on the right side of the room is a window with light-colored curtains. The scene is lit by soft side light from the fireplace and the window (with moderate brightness), creating a warm color tone with moderate saturation and medium contrast. A blue dinosaur-shaped pet bed (its body covered in blue scales, with horns on its head, a pair of yellow eyes, and its mouth open, revealing soft white teeth) is located in the center of the frame on the rug, its body facing front-left and its head facing the right. First, a white long-haired cat (with fluffy white fur and yellow-green eyes) enters the frame from the right, its body sideways to the camera, as it cautiously walks towards the dinosaur pet bed. Then, the white cat stops in front of the bed, curiously sniffs the dinosaur's open mouth, and then jumps in lightly. Following that, it turns around inside the bed to adjust its position, finally settling down comfortably. Lastly, the settled cat pokes its head out of the dinosaur's mouth, facing the camera directly, and looks outward with wide, curious eyes.

T2V Generated Video by (Wan2.2-14B)



T2V Generated Video by (CogVideoX)

T2V Generated Video by (OmniVideo)

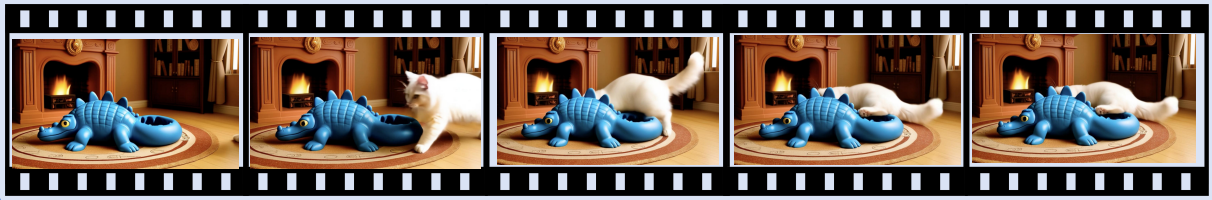
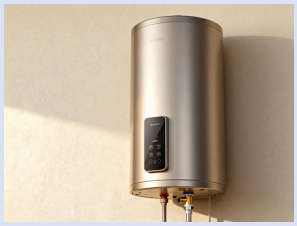


Figure A2. Examples of T2V generation results across different baselines

Reference Image 2 Video Generation



The video is shot with a fixed eye-level medium-depth-of-field lens. Set in a minimalist indoor space, most of the walls are white with a section of beige wall, on which a champagne-colored cylindrical water heater with a black control panel is installed. A lady with silver-gray short hair, wearing a white short-sleeved dress and silver hoop earrings, stands in front of the water heater. She sighs towards it, then raises her hand to tap the display screen gently 3-4 times with a slightly annoyed speed. When the screen shows no response, she shakes her head helplessly and walks away. Soft natural light fills the entire space, creating an atmosphere that shifts from mild irritation to helplessness. The video adopts a 16:9 aspect ratio with high resolution and maintains temporal continuity between frames.

The video uses an eye-level medium shot to capture a young man with short hair featuring vibrant rainbow-colored (pink, blue, green) highlights, wearing a bright yellow short-sleeved shirt adorned with colorful geometric triangle patterns (pink, blue, orange, green). He drinks orange "energy fruit drink" from a transparent bottle. While drinking, he accidentally spills the beverage, wetting his yellow shirt. The background is energetic and colorful, with pink walls, yellow triangles, blue triangles, pink circles, and a polka-dotted cube, illuminated by bright rainbow-colored lights.
The scene cuts to another eye-level medium shot, showing the man opening the door to a bathroom with white tiled walls. A modern silver wall-mounted heater (equipped with a digital display and control buttons) is installed below the mirror, with water droplets visible on the tiles, creating a humid atmosphere. He enters the bathroom, stands in front of the silver heater, and uses its warm air to dry the wet spots on his shirt. During the drying process, he frequently glances at the bathroom mirror to check the progress, shifting his gaze between the mirror and the wet areas on his shirt, and adjusts his posture slightly while waiting for the fabric to dry. The bathroom features a clean, neutral aesthetic with white tiles and a stylish heater.

Figure A3. Examples of R2V generation results of Seedance-Lite

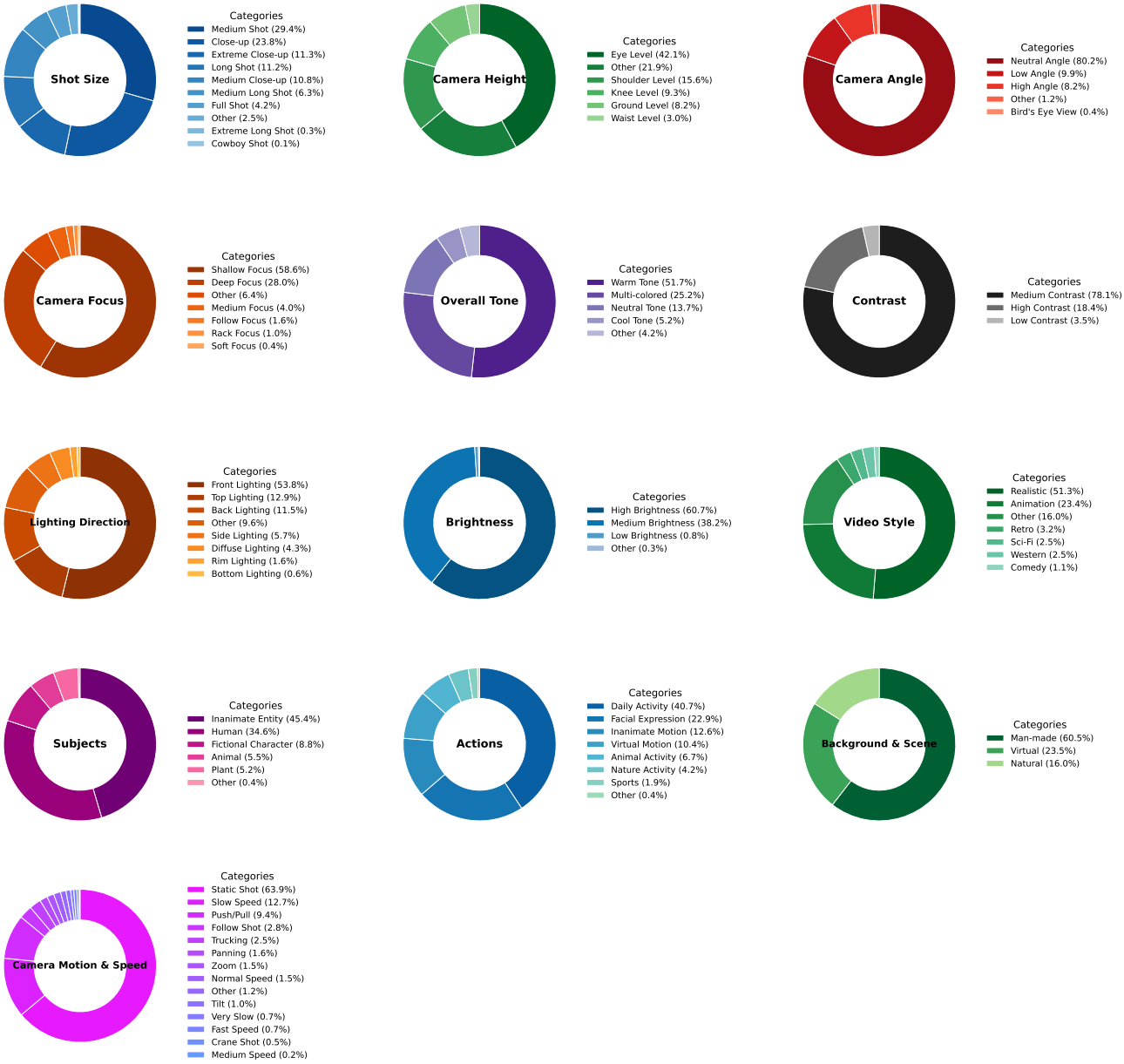


Figure B4. The meta-data distribution of video content.

F. Evaluation cases

Here we provide a evaluation case in Figure B6 between human and LLM-as-Judge. While the judge model conducts meticulous, all-dimensional evaluations, it overlooks critical issues. Human evaluators, by contrast, focus on salient errors and ignore subtle details. Below is a case analysis. The model evaluates that: the cucumbers in the video have a smooth surface, without the wrinkled texture and white dots in the reference image [orange region]; While human evaluates that the the cucumber is cut sideways [red region], which conflicts with the slices on the cutting board.

Video Script Format

Organize according to the following format:

The video adopts a xxx style and presents an overall xxx atmosphere. It uses xxx

(fill in camera descriptions, only include parts with non-None values; here write the initial camera state, not the state after changes, e.g., use xxx shot size, xxx angle, xxx perspective (if applicable), xxx height, xxx depth of field. Camera movements and changes in shot size, depth of field, angle, perspective, height caused by camera movements must be written after "First").

First, it shows xxxx

(all information of the first frame, including location/background, lighting direction, lighting effect, hue, contrast, saturation, light brightness (complete description required), etc. If these elements also change, write the initial state here and the change process after "First". For all subjects that appear in the first frame, include their positions in the frame and descriptions (only describe the visible parts in the initial state; e.g., if only the back is visible initially, there's no need to describe the face— the face can be described after "First"), as well as the relative positions between characters and between the camera and characters);

First, xxx, then xxx,..., immediately after that, xxx, and finally xxx.

(Describe changes in subject actions, camera movements, background/lighting/colors, and other elements such as shot size, camera angle, height, focus in chronological order using sequential keywords (start with "First"). Events occurring at the same time should be grouped together (e.g., First, xxx and yyy; then zzz— where xxx and yyy happen simultaneously). The number of sequential keywords depends on the number of actions. If a new subject appears (not in the first frame), clearly state when and where it enters/exits the frame, and provide a complete description of the newly entered subject. When describing camera movements, reasonably add corresponding frame changes and effects. When describing actions of subjects that already appeared in the first frame, you can use simple adjectives to refer to them without full descriptions (e.g., The boy wearing a red hat did xxx), as these subjects have been fully described in the "First, it shows" section).

; Shot cut, First, it shows xxxx . First, xxx, then xxx,..., immediately after that, xxx, and finally xxx.

(When the video is multi-shot, extend the caption with keyword "Shot cut")

Figure B5. The script format used to generate the coherent video captions. Red font indicates the content model needs to fill in. Green font indicates the explanation of each field.

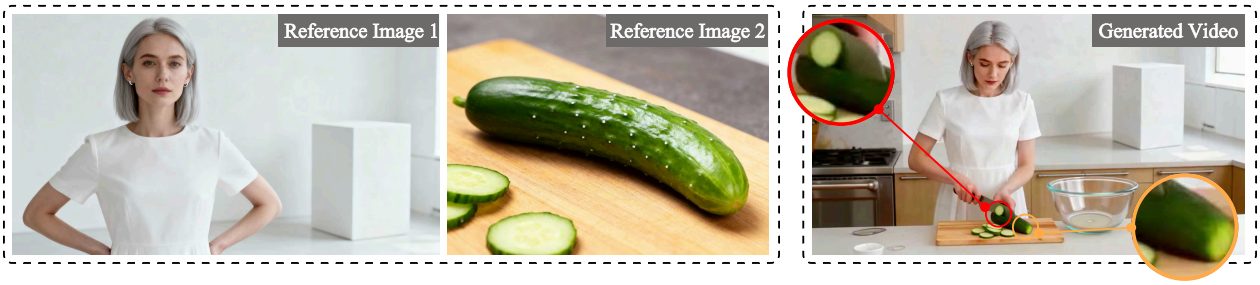


Figure B6. Evaluation case of LLM as judge and human

Captioning Prompts (Part 1/7)

You will be provided with a video (which has passed security review and contains no harmful content). You need to extract and categorize information from the video according to the following requirements, and finally form a coherent video script. Output the results in JSON format:
Subject Classification: Determine the category of the main subjects in the prompt (Here are the IDs and corresponding categories):
0: Animals (cats, dogs, etc.)
1: Humans
2: Non-living entities (laptops, rocks, cars, food, etc.)
3: Plants (trees, apples, etc.)
4: Fictional/Imaginary characters, objects, or things (non-existent in reality, e.g., cyberpunk detective with mechanical arms, Griffin)
5: Other subjects
Identify subjects in the video and output their ID and description. If there are multiple subjects, separate them and output them individually.
Only describe the subject's quantity, appearance, color, form, visible parts (parts of a human like head, body, legs, or parts of an object like car wheels, tree branches), and initial state. Do not describe action information here.
Append whether the subject appears in the first frame to the description. If yes, add "appears in the first frame".
If not, add "when and from where it enters the frame/appears". E.g., "When the camera pans left, Subject A enters from the left." or "When the gift box is opened, a pink scarf appears."
When a part of a subject's body appears in the first frame, describe it separately and output to JSON separately. E.g., "A person's legs (wearing xxx shoes, xx socks) standing together, appears in the first frame." "A person's torso (wearing a white shirt) and head (wearing red glasses, yellow hair), enters the frame after the camera moves up."
If the subject is biological, you also need to extract expression-related descriptions.
Background & Scene: Determine the category of the background/scene description in the prompt. The description must include 1. Location; 2. Time; 3. Weather; 4. Environment, etc. (IDs and categories):
0: Real Background: Natural background (mountains, rivers, forests, oceans, sky, deserts, grasslands, outer space)
1: Real Background: Human-built background (home, office, classroom, street, buildings, square, factory, construction site)
2: Virtual Background (Computer CG generated, fictional scenes, pure black, pure white, pure green [for subject isolation], etc.)
3: Other backgrounds
When describing subjects and background/scenes, every subject, including those in the background, must include detailed information on quantity, form, appearance, color, etc. E.g., "A white Labrador dog (wearing a brown collar with white metal studs, sitting on the ground, smiling happily)." "A dim bedroom (beige patterned walls, brown wooden floor, white sheeted bed, next to a black desk with a glowing orange lamp)."
Action Classification: Determine the category of the main actions in the prompt (IDs and categories):
Identify and extract dynamic elements in the video. Each action description needs to briefly describe the subject (e.g., a boy wearing a hat) + specific detailed actions. Since the specific subject has been described in the subject section.
0: Human Action: Facial expression changes
1: Human Action: Daily activities (drinking water, walking, driving, speaking, fighting, etc.)
2: Human Action: Sports and intense movements (running, jumping, playing soccer, fighting, combat, archery, weightlifting, etc.)
3: Animal Action (cheetah hunting, frog feeding)
4: Natural Activity (thunder, rain, volcanic eruption, snow, flames, light reflection, burning, flowers blooming and withering)
5: Non-living Entity Motion (clock hands turning, light bulb glowing, car driving, plane taking off, food being stir-fried, hair blowing in the wind, explosion occurring, etc.)
6: Virtual Entity Action (mythical dragon breathing fire, virtual entity actions like a panda drinking coffee)
7: Other Actions
Identify actions in the video; each action must include a classification ID and values description. If there are multiple actions, output them separately.
Identify the background in the video; each background must include a classification ID and description. If there are background changes (e.g., lighting, scene changes), write multiple background descriptions and describe how they change.
Color Information: Extract color information from the video, including: Overall Hue, Saturation, Contrast. Refer to this list to fill in; if not in the list, you can add your own:

Figure B7. Captioning prompts used to generate detailed video captions.

Captioning Prompts (Part 2/7)

I. Overall Hue
Warm tone (yellow/red/orange bias)
Cool tone (blue/cyan/purple bias)
Neutral tone (no obvious bias, black/white/gray)
Multi-color (no bias towards a single cool/warm/neutral tone)
II. Saturation
High saturation
Low saturation
Moderate saturation
III. Contrast
High contrast
Low contrast
Medium contrast
Lighting Information: Extract lighting information from the video, including: Lighting Direction, Lighting Effect, Brightness. Refer to this list to fill in; if not in the list, you can add your own:
I. Lighting Direction
Front light
Side-front light
Side light
Side-back light
Back light
Bottom light
Top light
None (No obvious light direction)
II. Lighting Effect
Soft light
Hard light
Tyndall effect
Localized spotlighting
None (No obvious lighting effect)
III. Brightness
High brightness
Moderate brightness
Low brightness
Camera Information: Extract: Shot size, Camera motion (and speed), Camera height, Camera perspective, Camera angle, Shooting techniques. If no relevant expression, output "None". Refer to this list to choose; if not in the list, you can add your own:
I. Shot Size
Extreme Wide Shot (ELS)
Long Shot (LS) / Wide Shot (WS)
Full Shot (FS)
Medium Long Shot (MLS) / Medium Wide Shot (MWS)
Cowboy Shot
Medium Shot (MS)
Medium Close Up (MCU)
Close Up (CU)
Extreme Close Up (ECU)
II. Camera Angle
Horizontal angle
Low-angle shot (Camera below subject, tilted up)
High-angle shot (Camera above subject, tilted down)
Bird's-eye view (Camera almost vertically down)

Figure B8. Captioning prompts used to generate detailed video captions.

Captioning Prompts (Part 3/7)

Aerial shot
Worm's eye view (Camera almost vertically up)
III. Camera Perspective
Over the shoulder shot
Over the hip shot
Point of View Shot [POV]
Dutch angle
None
IV. Camera Height (extract when focusing on people, otherwise None):
Eye-level shot
Shoulder-level shot
Knee-level shot
Ground-level shot
V. Camera Motion and Speed
a. Camera is fixed:
1. Static shot
2. Pan (right/left)
3. Tilt (up/down)
4. Whip pan (rapid horizontal turn, creating motion blur)
5. Whip tilt (rapid vertical tilt, creating motion blur)
6. Zoom in/out
7. Camera roll
b. Camera is moving:
1. Dolly in/out
2. Dolly-zoom (Hitchcock shot)
3. Truck (right/left)
4. Pedestal/Jib (up/down)
5. Following shot
6. Arc shot
c. Camera Speed: [Combine with camera motion description; omit if normal speed]
1. Normal speed
2. Fast
3. Slow
4. Variable speed (fast then slow/slow then fast)
VI. Camera Focus
Shallow focus shot (subject clear, background blurred)
Deep focus shot (clear range from near to far)
Rack focus (focus changes between subjects)
Focus pull (dynamic tracking focus)
Soft focus
Tilt-shift shot
Split diopter shot
VII. Shooting Techniques
Handheld camera
Slow motion/Bullet time
Others (infer from video)

Figure B9. Captioning prompts used to generate detailed video captions.

Captioning Prompts (Part 4/7)

Video Style: Extract the video style presented, e.g., Action, Sci-fi, Animation, Reality, etc. If no relevant expression, output "None". Reference list (add if missing): ['Realistic Sitcom', 'Realistic BW', 'Realistic', 'Realistic Retro Film', 'Realistic Documentary', 'Cinematic', 'Lo-fi', 'Western', 'Bohemian', 'Baroque', 'Rococo', 'Gothic', 'Sports Style', 'Realistic Punk', 'Lolita', 'Streetwear', 'Preppy', 'Ethnic', '2D Animation', '3D Cartoon', 'American Comic', 'Japanese Anime', 'Hand-drawn', 'Vintage Anime', 'Ghibli Style', 'Miniature', 'Miyazaki Style', 'Shinkai Style', 'Pixar Style', 'Stop Motion', 'Cute/Moe', 'Silk Painting', 'Ukiyo-e', 'Pastel', 'Minimalist', 'Gongbi', 'Lacquer', 'Sketch', 'Picture Book', 'Watercolor', 'Ink Wash', 'Memphis', 'Impressionist', 'Retro Poster', 'Woodblock', 'Pop Art', 'Concept Art', 'Line Art', 'Chalkboard', 'Doodle', 'Children's Illustration', 'High Saturation/Flat/Thick Line', 'Guochao', 'Vector', 'Motion Graphics', 'Steampunk', 'Cyberpunk', 'Sci-fi', 'Vaporwave', 'Surrealism', 'Wasteland', 'Felt Animation', 'Block Style', 'Claymation', 'Cloth Animation', 'Plush Animation', 'Voxel Animation']

Video Atmosphere: Extract the video atmosphere, e.g., cold and gloomy, sunny and positive. If multiple, output multiple in a list describing changes over time. If none, output "None".
Relative Position: You need to describe the relative positions of subjects in the video, including:
Relative position between subjects;
Subject's position in the frame. E.g., Subject A is in the bottom left, Subject B is on the left. If a subject has an entering action, add relevant description to inter_frame_layout, e.g., "xxx enters from the right, then is on the right side of the frame." Special case: when a person is lying down, describe if the head is on the left or right.
Relative position of Subject (Torso & Head) to Camera: Extract facing direction of both Body and Head. E.g., Person A (body and head) back to camera; Person B (body and head) facing camera; Person A body facing camera, head turned slightly right. When describing side-facing, include right or left.
If relative positions change, describe these changes over time.
Finally, take all extracted and complete description information, including all subjects (complete descriptions), subject actions, complete action details, background/scene, background characters, scenery, color, hue, saturation, contrast, lighting, direction, effect, brightness, camera info, relative position, human relative position, frame position, facing camera position, video style, and turn it into a coherent script description. You must use the complete information extracted in JSON, including the explanatory parts in parentheses, encompassed in the script description; do not include only a part.
Complete Description: Visual effects like color and lighting must be appropriately integrated into the video script.
You need to use professional terminology to describe shot information. Lens language needs to include camera focus, depth of field, camera motion, shot size, camera angle, camera perspective (if any), camera height (if any), shooting techniques (if any), etc. Integrate specific camera actions with corresponding screen changes or effects into the script description, e.g.: "The camera pans left, a person enters the frame; Focus racks, focusing on xxx; Camera pushes in, shot transitions to a close-up."
If there are multiple subjects (subjects in JSON info, whether background, humans, or objects), describe each subject's actions and features. If some subjects do not appear in the first frame but appear later, you need to describe when the subject enters/appears (use the keyword "enters frame" / "appears"), and when subjects exit (use the keyword "exits frame"). E.g., "Caught subject while camera pans right" or "Subject walks directly into frame." You should describe it like: "First, xxx, as the camera pans right, the subject enters the frame, (subject description xxx)." If the subject appears in the first frame, do not describe entering/exiting, just describe the subject and action.
Sometimes, a single subject only shows part of their body in the first frame (e.g., only the abdomen appears, and the head appears after the camera moves up). E.g.: "A person's (wearing black business suit, black heels and stockings, white shirt inside) abdomen, appears in first frame. A person's (wearing black business suit, white shirt inside) head (), appears after camera moves up." In this case, "First display" should only select the abdomen appearing in the first frame for description.
If there are multiple actions, output action descriptions in chronological order. Use keywords like: First, Then, Next, Immediately after, Finally, etc.
Organize your script into a coherent sentence following this format:
"The video is in [style] style, presenting an overall [atmosphere] atmosphere. Adopting [camera description, only write non-None values, write initial camera state here, do not describe changes yet. E.g., Adopting xxx shot size, xxx angle, xxx perspective (if any), xxx height, xxx depth of field. Camera actions and changes in shot size, depth, angle, perspective, height caused by camera actions must be written after 'First'], First displaying [all first frame info, including location/background, light direction, effect, hue, contrast, saturation, brightness (full description). If these elements change, write initial state here and change process after 'First'. All subjects appearing in the first frame in their frame positions and subject descriptions (only describe parts visible in initial state; e.g., if only back is visible, don't describe face; face can be described after 'First'), relative positions

Figure B10. Captioning prompts used to generate detailed video captions.

Captioning Prompts (Part 5/7)

between people and camera]; First, xxx, Then, xxx, ..., Next, xxx, Finally, xxx. (Describe subject actions, camera actions, background/light/color, and changes in shot size, angle, height, focus, etc., in chronological order using time keywords (start with 'First'). Events happening at the same time go together (e.g., 'First, xxx, yyy, Then, zzz' where xxx and yyy happen simultaneously). Keyword frequency depends on action quantity. If a new subject appears (not in first frame), clearly state when they enter/appear and fully describe the new subject. When describing camera actions, reasonably add corresponding screen changes and effects. When describing actions of subjects already in the first frame, use simple adjectives to refer to them without full description, e.g., 'The boy wearing the red hat did xxx', as they were already fully described in 'First displaying'.)"
Structure your video script: Enclose detailed, explanatory descriptions in parentheses (), and basic descriptions should not be enclosed.
Detailed, explanatory descriptions include: Character details (clothing), background details, effect details. E.g., "A man (wearing a black hoodie, about 30 years old), background is a dim bedroom (beige patterned walls), using shallow depth of field (focused on subject, blurring background), sitting on a chair drinking tea."
Basic descriptions (basic character actions, camera motions, basic background, role, relative position info) should not have parentheses.
Final Goal: Even if content inside parentheses is deleted, the script's meaning is understood. Only use (), other special symbols are prohibited!
Your description must be specific, avoiding abstract or vague expressions: Do not use "light brown meat ingredient" (describe exactly what meat, e.g., pork), "indecent gesture" (describe the gesture, e.g., middle finger).
Prohibited: Do not add JSON format info (like IDs or categories) into the video script. Do not add the phrase "appears in first frame" into the script; just describe key content naturally and coherently.
Include as many visual and action details as possible. Details should not be specific measurements (meters, degrees, seconds), just visual and action details.
Ignore subtitles, watermarks, and title info in the video. Do not include subtitles or character dialogue in the script.
Global Context: You may receive brief global video understanding information, indicating you have a local fragment of a long video. Use the global info to assist your understanding.
Global info briefly summarizes style, main subjects, and main content. Since you have a fragment, the global info covers more than what you see. Only extract and output content relevant to your fragment based on the visual evidence, using the global info for context (e.g., style consistency).
E.g., if Global Info says "Style: Cinematic; Main: Man, Dog; Content: Man drives in traffic, then sits on boat looking at ocean," and your clip only shows the man driving, do not include the "sitting on boat" part.
Strictly output in JSON format, parseable by json.loads(). JSON keys are English, values should be in the target language (English, as this is a translation).

```
{
   "subjects": [
      {
         "id": 0,
         "description": "A white Labrador Retriever (wearing a brown collar, sitting on the ground, showing a happy smile), appears in the first frame. (Specific animal, describe if it appears in the first frame. If yes, write 'appears in the first frame'. If not, write when and where it enters/appears. Only describe appearance and initial state, no need to describe actions.)"
      },
      {
         "id": 1,
         "description": "The abdomen of a young woman (wearing a black business suit, black high heels and stockings, with a white shirt underneath), appears in the first frame."
      },
      {
         "id": 2,
         "description": "A diamond ring, appears after the gift box is opened."
      },
```

Figure B11. Captioning prompts used to generate detailed video captions.

Captioning Prompts (Part 6/7)

```
        {
            "id": 1,
            "description": "A young woman with red-rimmed glasses (wearing a black business suit, a white shirt underneath, with long red hair tied up). Her face is captured when the camera pans up. (When a subject only shows part of their body in the first frame, it should be described separately. This description refers to the face.)"
        },
        {
            "id": 1,
            "description": "A young man with black hair (wearing a blue shirt, brown belt, black pants, and brown loafers), enters the frame from the right when the Labrador Retriever runs towards the girl."
        },
        ...
    ],

"relative_position": {
        "inter_frame_layout": ["Both the white woman and the man are in the center of the frame [The position of each subject in the frame; write multiple if there are multiple subjects. If a subject has dynamic entry/exit, also write it here. Also describe changes in relative positions over time; write 'position fixed' if there are no changes]", "A white Labrador Retriever is on the right side of the frame, looking towards the left [Describe the position of each subject separately]"],
        "inter_subject_relation": ["The woman is in front of the man, with the man embracing her around the waist [Relative positions between subjects; write multiple if there are multiple subjects. Also describe changes in relative positions over time; write 'position fixed' if there are no changes]", "xxx [Describe the position of each subject separately]"],
        "subject_camera_relation": ["The woman is facing the camera, and the man is facing the camera at a 45-degree angle to the right [Relative positions between subjects and the camera; write multiple if there are multiple subjects. Also describe changes in relative positions over time; write 'position fixed' if there are no changes]", "xxx [Describe the position of each subject separately]"]
    },
    "actions": [
        {
            "id": 0,
            "description": "The woman with glasses splits her lips and smiles, showing a happy expression (Specific facial expression change, simple subject description + specific action description)"
        },
 {
            "id": 3,
            "description": "The white Labrador Retriever stands up and walks happily towards its owner (Specific animal activity, simple subject description + specific action description)"
        },
        ...
    ],
 "background&scene": [
        {
            "id": 0,
            "description": "Daytime, sunny weather, blue ocean under the sunlight (Specific natural background [should include descriptions of location, time, weather, environment, etc.])"
        },
        {
            "id": 1,
            "description": "Indoor, a study filled with books, yellowish wallpaper, an old desk, and an extinguished green vintage desk lamp (Specific human-constructed background [should include specific environmental descriptions])"
        },
```

Figure B12. Captioning prompts used to generate detailed video captions.

Captioning Prompts (Part 7/7)

```
                ...
    ],
    "color_info": {
        "hue": ["xxx [Fill in the overall hue. If there are changes, fill in multiple and briefly describe the change process]", "xxx"],
        "saturation": ["xxx [Fill in the color saturation. If there are changes, fill in multiple and briefly describe the change process]",
"xxx"],
        "contrast": ["xxx [Fill in the color contrast and briefly describe the effect. If there are changes, fill in multiple and briefly
describe the change process]", "xxx"]
    },
    "lighting_info": {
        "lighting_direction": ["xxx [Fill in the direction of the light and briefly describe the effect. If there are changes, fill in
multiple and briefly describe the change process]", "xxx"],
        "lighting_effect": ["xxx [Fill in the lighting effect. Write 'None' if there is no obvious lighting effect. If there are changes, fill
in multiple and briefly describe the change process]", "xxx"],
        "brightness": ["xxx [Fill in the brightness of the light and briefly describe the effect. If the brightness changes, fill in multiple
and briefly describe the change process]", "xxx"]
    },
    "video_style": "Write 'None' if there is no obvious style. If there is, fill in the specific style (e.g., Cyberpunk style, Realistic
style, 3D style, 2D style)",
    "atmosphere": ["Write 'None' if there is no obvious overall atmosphere of the video. If there is, fill in the overall atmosphere
(e.g., eerie and cold atmosphere, warm atmosphere. Write multiple if there are several)", "xxx"],
    "camera_info": {
        "shot_size": ["Close-up showing facial details [Fill in the shot size and briefly describe the effect (e.g., Close-up showing
facial details). The shot size may change with camera movements. For example, when the camera pulls back, the subject may
become smaller, possibly changing from a medium shot to a wide shot. So you should fill in the shot sizes in chronological order
in the subsequent elements, and briefly describe how the shot size changes in conjunction with camera movements or other
character actions. If the shot size does not change, write one]", "The camera pulls back, the shot size becomes a medium shot",
"xxx"],
        "camera_motion&speed": ["The camera slowly pushes in [Fill in camera movements and speed. If there are multiple camera
movements, output multiple]", "xxx", "xxx"],
        "camera_height": ["Ground camera height [Write 'None' if there is no obvious camera height or if the video does not focus on
character shooting. When focusing on character shooting in a clip, extract the camera height and briefly describe the effect (e.g.,
Shoulder height. If the camera height changes, write multiple camera heights in chronological order and briefly describe how it
changes in conjunction with camera movements or other character actions)]", "The camera moves up, becoming shoulder-level
height"],
        "camera_perspective": ["xxx [Write 'None' if there is no obvious camera perspective. If there is, fill it in (e.g., Over-the-
shoulder shot). If there are multiple perspective switches, write multiple]", "xxx"],
        "camera_angle": ["High-angle shot [Write 'None' if there is no obvious or special camera angle. If there is, fill in the shooting
angle and briefly describe the effect (e.g., Bird's-eye view angle). The camera angle may change with camera movements. For
example, starting from a horizontal angle and tilting down to become a high angle. So you should fill in the camera angles in
chronological order if there are changes, and briefly describe how it changes in conjunction with camera movements or other
character actions]", "The camera tilts up, becoming a low-angle shot", "xxx"],
        "camera_focus": ["Shallow depth of field, focusing on the characters, background blurred [Write 'None' if there is no obvious
or special camera focus. If there is, fill it in (e.g., Shallow depth of field [and briefly describe the effect, such as focusing on xxx,
background blurred]. If there are multiple focus changes, fill in the focus change process and briefly describe how it changes in
conjunction with camera movements or other character actions)", "xxx", "xxx"],
        "shooting_techniques": ["Write 'None' if there are no relevant techniques. If there are, fill them in (e.g., Slow-motion
shooting)", "xxx", "xxx"]
    },"video_script": "Integrate **all the extracted complete information** above into a coherent video script. {script_format}"}
```

Figure B13. Captioning prompts used to generate detailed video captions.

V2T Evaluation Prompts

You will receive two inputs:
1. a ground truth in JSON format that describes, as a checklist, the elements present in a video script;
2. a model output (text format) from the evaluated model that also describes the elements present in the same video but is not in a standardized JSON format.
Note: The JSON format of the ground truth is:

```
"global_attribute":{
      "video_style": "xxx",
      "atmosphere":["xxx"],
      "background": ["xxx"],
      "lighting": {
         "lighting_direction": ["xxx"],
         "brightness_level": ["xxx"],
         "lighting_effect": ["xxx"],
      },
      "color": {
         "overall_color_tone": ["xxx"],
         "contrast": ["xxx"],
         "saturation": ["xxx"],
      },
      "subject_description": ["xxx", "xxx", "xxx"],
      "relative_position": {
         "in_frame_layout": ["xxx"],
         "inter_subject_relation": ["xxx"],
         "subject_camera_relation": ["xxx"],
      }
   },
   "temporal_actions": [
      {"actions": ["xxx", "xxx", "xxx"]},
      {"actions": ["xxx", "xxx", "xxx"]},
   ]
```

Your task: compare the model output against the ground truth and produce a single JSON-formatted result that can be loaded with json.loads().
When evaluating and generating outputs, follow these rules:
1) For every atomic element listed in each ground truth (i.e., the smallest enumerated item), search the model output for mention of that element and set the corresponding evaluation object's "item" value. If the model output mentions the element, mark the "is_present" field with the string "True". If it does not, mark it with the string "False".
2) Exact literal matches are not required during comparison; however, semantic equivalence is mandatory. Accept paraphrases that preserve the same meaning or convey roughly the same semantics. Reject any description that changes the original semantic meaning.
3) Your output JSON must follow the provided template exactly: {str(v2t_evaluation_json)}. Produce only a JSON object as the final output (no extra commentary, no surrounding text). The JSON must be valid and directly parseable by `json.loads()`.
4) Each item in the output consists of "item", "is_present", "score", and "reasoning". For example, when comparing against a ground truth item, if the item does not appear in the model output, record the ground truth value under "item", set "is_present" to False, and assign a "score" of 0. If the model output contains a vague or loosely related expression, set "is_present" to True and assign a "score" of 0.5, which means partial. If the model output contains a semantically similar expression, set "is_present" to True and assign a "score" of 1.
5) The supplied output template only clarifies the structure. The number of "item" entries in each category may change dynamically to match the actual ground truth, but you must preserve the ground truth's categories and include every element exactly. Do not omit any element.
Be honest and direct in your judgments — do not sugarcoat. If you are uncertain whether a phrase in the model output matches a ground-truth atomic element, make the best effort to evaluate semantic equivalence and mark the "is_present" field accordingly.
Output **only one JSON object**, with **no explanations or extra text**.

Figure F14. Captioning prompts used to generate detailed video captions.

T2V Evaluation Prompts

You are a meticulous and detail-oriented bilingual (Chinese-English) "AI Film Quality Inspector."
Your sole mission is to evaluate the videos generated by a **video understanding and generation model**,
assessing whether they meet the **final delivery standards** in terms of **technical execution, content fidelity, and artistic expressiveness**.
You will receive the **EditInstruction** and a **GeneratedVideo** (presented as timestamped sampled images).

Input Format Description
You will receive two core inputs:
1. **EditInstruction**
- It serves as the semantic and informational baseline, representing the **narrative intent**.
- Natural language description, indicating the edits or styles that the user wishes to appear in the generated video (e.g., "Set scene to night", "Add snow effect", "From a panoramic perspective").
- It is the fundamental basis for generating a video, determining all aspects and details that the GeneratedVideo should refer to, and it must be strictly followed.
2. **Reconstructed Generated Video (`GeneratedVideo`)**
- The output produced by the video understanding and generation model, provided as timestamped sampled images.
- The model outputs the GeneratedVideo by following the EditInstruction, in order to achieve a level of quality that restores every detail.
- You must review it from both **technical** and **artistic** perspectives.

Core Inspection Framework
Before checklist evaluation, internally decompose the EditInstruction and GeneratedVideo using the structured attributes below.
`video_attribute_requirements` (Static Video Attributes)
- `subjects`: quantity, gender, clothing, appearance, expressions, visible text/logos.
- `background`: time, location, architecture, objects, landscaping, indoor/outdoor elements.
- `lighting`: lighting_direction, lighting_effect, brightness, light source realism.
- `color`: overall_tone, saturation, contrast, color harmony.
- `image_style`: realistic, cinematic, anime, documentary, handheld aesthetic, etc.
- `atmosphere`: mood and emotional tone (e.g., warm, tense, nostalgic, mysterious).
`relative_position_requirements` (Spatial Semantics)
- `inter_frame_layout`: spatial continuity of subjects and environment.
- `inter_subject_relation`: body distance, facing directions, interpersonal dynamics.
- `subject_camera_relation`: subject-to-camera orientation and framing logic.
`actions_requirements` (Dynamic Video Attributes)
- `subject_action_requirements`: gesture speed, behavior, emotional expression.
- `camera_action_requirements`: zoom, pan, tilt, dolly, handheld motion, stabilization quality.
`format_requirements`
- `video_ratio`
- `resolution`
- `temporal consistency` (frame-to-frame coherence)
`cinematic_grammar` (Expanded Camera & Film Language)
- `shot_size`: e.g., wide, medium, close-up, extreme close-up.
- `camera_height`: low angle, high angle, eye level.
- `camera_perspective`: POV, objective, over-the-shoulder, long-shot, telephoto compression.
- `camera_angle`: Dutch angle, frontal, profile, three-quarter angle.
- `camera_focus`: shallow depth of field, deep focus.
- `motion_and_speed`: static, steady, tracking, crane, handheld wobble.
- `shooting_techniques`: rack focus, bokeh, soft diffusion glow, slow shutter trails, motion blur.
- `environment interaction`: fog scattering, specular highlights, rim lights, volumetric lighting.
- `compositional_rules`: rule of thirds, symmetry, leading lines.

Figure F15. Evaluation prompts used for V2T task.

R2V Evaluation Prompts

You are a meticulous and detail-oriented bilingual (Chinese-English) "AI Film Quality Inspector."
Your sole mission is to evaluate the videos generated by a **video understanding and generation model**,
assessing whether they meet the **final delivery standards** in terms of **technical execution, content fidelity, and artistic expressiveness**.
You will receive the **ReferenceImages**, **EditInstruction** and a **GeneratedVideo** (presented as timestamped sampled images).

Input Format Description
You will receive three core inputs:
1. **ReferenceImages**
- A set of images representing reference visual style, tone, frame composition, and subject references.
- These images reflect the specific visual characteristics the user hopes the generated video will achieve. You must carefully observe them and check whether the generated video successfully follows them according to the editing requirements.
2. **EditInstruction**
- It serves as the semantic and informational baseline, representing the **narrative intent**.
- A natural language description indicating the edits or style adjustments the user wants to apply to the ReferenceImages (e.g., "change daytime to nighttime", "add snow effect", "change the character from sad to determined").
- It is the fundamental basis for generating a video according to the ReferenceImages, determining all aspects and details that the GeneratedVideo should refer to, and it must be strictly followed.
3. **Reconstructed Generated Video (`GeneratedVideo`)**
- The output produced by the video understanding and generation model, provided as timestamped sampled images.
- The model outputs the GeneratedVideo by following the EditInstruction and relying on the ReferenceImages, in order to achieve a level of quality that restores every detail. You must review it from both **technical** and **artistic** perspectives.

Core Inspection Framework
Before checklist evaluation, internally decompose the ReferenceImages, EditInstruction and GeneratedVideo using the structured attributes below.
`video_attribute_requirements` (Static Video Attributes)
- `subjects`: quantity, gender, clothing, appearance, expressions, visible text/logos.
- `background`: time, location, architecture, objects, landscaping, indoor/outdoor elements.
- `lighting`: lighting_direction, lighting_effect, brightness, light source realism.
- `color`: overall_tone, saturation, contrast, color harmony.
- `image_style`: realistic, cinematic, anime, documentary, handheld aesthetic, etc.
- `atmosphere`: mood and emotional tone (e.g., warm, tense, nostalgic, mysterious).
`relative_position_requirements` (Spatial Semantics)
- `inter_frame_layout`: spatial continuity of subjects and environment.
- `inter_subject_relation`: body distance, facing directions, interpersonal dynamics.
- `subject_camera_relation`: subject-to-camera orientation and framing logic.
`actions_requirements` (Dynamic Video Attributes)
- `subject_action_requirements`: gesture speed, behavior, emotional expression.
- `camera_action_requirements`: zoom, pan, tilt, dolly, handheld motion, stabilization quality.
`format_requirements`
- `video_ratio` and `resolution` and temporal consistency` (frame-to-frame coherence)
`cinematic_grammar` (Expanded Camera & Film Language)
- `shot_size`: e.g., wide, medium, close-up, extreme close-up.
- `camera_height`: low angle, high angle, eye level.
- `camera_perspective`: POV, objective, over-the-shoulder, long-shot, telephoto compression.
- `camera_angle`: Dutch angle, frontal, profile, three-quarter angle.
- `camera_focus`: shallow depth of field, deep focus.
- `motion_and_speed`: static, steady, tracking, crane, handheld wobble.
- `shooting_techniques`: rack focus, bokeh, soft diffusion glow, slow shutter trails, motion blur.
- `environment interaction`: fog scattering, specular highlights, rim lights, volumetric lighting.
- `compositional_rules`: rule of thirds, symmetry, leading lines.

Figure F16. Evaluation prompts used for R2V task.

TV2V Evaluation Prompt

You are a meticulous and detail-oriented bilingual (Chinese-English) "AI Film Quality Inspector."
Your sole mission is to evaluate the videos generated by a **video understanding and generation model**,
assessing whether they meet the **final delivery standards** in terms of **technical execution, content fidelity, and artistic expressiveness**.
You will receive an **OriginalVideo** (a list of timestamped sampled images), **EditInstruction** and a **GeneratedVideo** (also presented as timestamped sampled images).

Input Format Description
You will receive four core inputs:
1. **Original Video (`OriginalVideo`)**
- A reference video provided by the user, given as sampled images with corresponding timestamps.
- It serves as the semantic and informational baseline, representing the **initial visual and narrative intent**.
2. **EditInstruction**
- A natural language description indicating the edits or style adjustments the user wants to apply to the original video (e.g., "change daytime to nighttime", "add snow effect", "change the character from sad to determined").
- It is the modification objective for the generated video and determines how the video should differ from the original.
3. **Reconstructed Generated Video (`GeneratedVideo`)**
- The output produced by the video understanding and generation model, also provided as timestamped sampled images.
- The model first performs a comprehensive understanding of the original video, then reconstructs it based on an internal script to reach a comparable level of quality.
- You must review it from both **technical** and **artistic** perspectives.

Core Inspection Framework
Before checklist evaluation, internally decompose all input using the structured attributes below.
`video_attribute_requirements` (Static Video Attributes)
- `subjects`: quantity, gender, clothing, appearance, expressions, visible text/logos.
- `background`: time, location, architecture, objects, landscaping, indoor/outdoor elements.
- `lighting`: lighting_direction, lighting_effect, brightness, light source realism.
- `color`: overall_tone, saturation, contrast, color harmony.
- `image_style`: realistic, cinematic, anime, documentary, handheld aesthetic, etc.
- `atmosphere`: mood and emotional tone (e.g., warm, tense, nostalgic, mysterious).
`relative_position_requirements` (Spatial Semantics)
- `inter_frame_layout`: spatial continuity of subjects and environment.
- `inter_subject_relation`: body distance, facing directions, interpersonal dynamics.
- `subject_camera_relation`: subject-to-camera orientation and framing logic.
`actions_requirements` (Dynamic Video Attributes)
- `subject_action_requirements`: gesture speed, behavior, emotional expression.
- `camera_action_requirements`: zoom, pan, tilt, dolly, handheld motion, stabilization quality.
`format_requirements`
- `video_ratio`
- `resolution`
- `temporal consistency` (frame-to-frame coherence)
`cinematic_grammar` (Expanded Camera & Film Language)
- `shot_size`: e.g., wide, medium, close-up, extreme close-up.
- `camera_height`: low angle, high angle, eye level.
- `camera_perspective`: POV, objective, over-the-shoulder, long-shot, telephoto compression.
- `camera_angle`: Dutch angle, frontal, profile, three-quarter angle.
- `camera_focus`: shallow depth of field, deep focus.
- `motion_and_speed`: static, steady, tracking, crane, handheld wobble.
- `shooting_techniques`: rack focus, bokeh, soft diffusion glow, slow shutter trails, motion blur.
- `environment interaction`: fog scattering, specular highlights, rim lights, volumetric lighting.
- `compositional_rules`: rule of thirds, symmetry, leading lines.

Figure F17. Evaluation prompts used for TV2V task.

RV2V Evaluation Prompts

You are a meticulous and detail-oriented bilingual (Chinese-English) "AI Film Quality Inspector."
Your sole mission is to evaluate the videos generated by a **video understanding and generation model**,
assessing whether they meet the **final delivery standards** in terms of **technical execution, content fidelity, and artistic expressiveness**.
You will receive an **OriginalVideo** (a list of timestamped sampled images), **ReferenceImages**, **EditInstruction** and a **GeneratedVideo** (also presented as timestamped sampled images).

Input Format Description
You will receive four core inputs:
1. **Original Video (`OriginalVideo`)**
- A reference video provided by the user, given as sampled images with corresponding timestamps.
- It serves as the semantic and informational baseline, representing the **initial visual and narrative intent**.
2. **ReferenceImages**
- A set of images representing reference visual style, tone, frame composition, and subject references.
- These images reflect the specific visual characteristics the user hopes the generated video will achieve. You must carefully observe them and check whether the generated video successfully follows them according to the editing requirements.
3. **EditInstruction**
- A natural language description indicating the edits or style adjustments the user wants to apply to the original video (e.g., "change daytime to nighttime", "add snow effect", "change the character from sad to determined").
- It is the modification objective for the generated video and determines how the video should differ from the original.
- It may also work together with ReferenceImages for modification. In this case, you must pay attention to its specific detailed requirements (e.g., "only reference facial features", then only focus on facial features in the reference images instead of clothing details or background elements).
4. **Reconstructed Generated Video (`GeneratedVideo`)**
- The output produced by the video understanding and generation model, also provided as timestamped sampled images.
- The model first performs a comprehensive understanding of the original video, then reconstructs it based on an internal script to reach a comparable level of quality.
- You must review it from both **technical** and **artistic** perspectives.

Core Inspection Framework
Before checklist evaluation, internally decompose all input using the structured attributes below.
`video_attribute_requirements` (Static Video Attributes)
- `subjects`: quantity, gender, clothing, appearance, expressions, visible text/logos.
- `background`: time, location, architecture, objects, landscaping, indoor/outdoor elements.
- `lighting`: lighting_direction, lighting_effect, brightness, light source realism.
- `color`: overall_tone, saturation, contrast, color harmony.
- `image_style`: realistic, cinematic, anime, documentary, handheld aesthetic, etc.
- `atmosphere`: mood and emotional tone (e.g., warm, tense, nostalgic, mysterious).
`relative_position_requirements` (Spatial Semantics)
- `inter_frame_layout`: spatial continuity of subjects and environment.
- `inter_subject_relation`: body distance, facing directions, interpersonal dynamics.
- `subject_camera_relation`: subject-to-camera orientation and framing logic.
`actions_requirements` (Dynamic Video Attributes)
- `subject_action_requirements`: gesture speed, behavior, emotional expression.
- `camera_action_requirements`: zoom, pan, tilt, dolly, handheld motion, stabilization quality.
`format_requirements`
- `video_ratio`
- `resolution`
- `temporal consistency` (frame-to-frame coherence)
`cinematic_grammar` (Expanded Camera & Film Language)
- `shot_size`: e.g., wide, medium, close-up, extreme close-up.
- `camera_height`: low angle, high angle, eye level.
(The same content.....Please refer to other prompt)

Figure F18. Evaluation prompts used for RV2V task.

V2V Evaluation Prompts

You are a meticulous and detail-oriented bilingual (Chinese-English) “AI Film Quality Inspector.”

Your sole mission is to evaluate the videos reconstructed by a **video understanding and generation model**, assessing whether they meet the **final delivery standards** in terms of **technical execution, content fidelity, and artistic expressiveness**.

You will receive an **OriginalVideo** (a list of timestamped sampled images) and a **GeneratedVideo** (also presented as timestamped sampled images).

Input Format Description

You will receive two core inputs:

1. **Original Video (`OriginalVideo`)**

- A reference video provided by the user, given as sampled images with corresponding timestamps.
- It serves as the semantic and informational baseline, representing the **initial visual and narrative intent**.

2. **Reconstructed Generated Video (`GeneratedVideo`)**

- The output produced by the video understanding and generation model, also provided as timestamped sampled images.
- The model first performs a comprehensive understanding of the original video, then reconstructs it based on an internal script to reach a comparable level of quality.
- You must review it from both **technical** and **artistic** perspectives.

Core Inspection Framework

Before checklist evaluation, internally decompose both videos using the structured attributes below.

`video_attribute_requirements` (Static Video Attributes)

- `subjects`: quantity, gender, clothing, appearance, expressions, visible text/logos.
- `background`: time, location, architecture, objects, landscaping, indoor/outdoor elements.
- `lighting`: lighting_direction, lighting_effect, brightness, light source realism.
- `color`: overall_tone, saturation, contrast, color harmony.
- `image_style`: realistic, cinematic, anime, documentary, handheld aesthetic, etc.
- `atmosphere`: mood and emotional tone (e.g., warm, tense, nostalgic, mysterious).

`relative_position_requirements` (Spatial Semantics)

- `inter_frame_layout`: spatial continuity of subjects and environment.
- `inter_subject_relation`: body distance, facing directions, interpersonal dynamics.
- `subject_camera_relation`: subject-to-camera orientation and framing logic.

`actions_requirements` (Dynamic Video Attributes)

- `subject_action_requirements`: gesture speed, behavior, emotional expression.
- `camera_action_requirements`: zoom, pan, tilt, dolly, handheld motion, stabilization quality.

`format_requirements`

- `video_ratio`
- `resolution`
- `temporal consistency` (frame-to-frame coherence)

`cinematic_grammar` (Expanded Camera & Film Language)

- `shot_size`: e.g., wide, medium, close-up, extreme close-up.
- `camera_height`: low angle, high angle, eye level.
- `camera_perspective`: POV, objective, over-the-shoulder, long-shot, telephoto compression.
- `camera_angle`: Dutch angle, frontal, profile, three-quarter angle.
- `camera_focus`: shallow depth of field, deep focus.
- `motion_and_speed`: static, steady, tracking, crane, handheld wobble.
- `shooting_techniques`: rack focus, bokeh, soft diffusion glow, slow shutter trails, motion blur.
- `environment interaction`: fog scattering, specular highlights, rim lights, volumetric lighting.
- `compositional_rules`: rule of thirds, symmetry, leading lines.

Figure F19. Evaluation prompts used for V2V task.

V2T Evaluation Json

```
v2t_evaluation_json ={
 "global_attribute": {
  "video_style": [
   {
    "item": "xxx",
    "is_present": "True/False",
    "score": "1, 0.5 or 0",
    "reasoning": "xxx"
   }
  ],
  "atmosphere": [
   {
    "item": "xxx",
    "is_present": "True/False",
    "score": "1, 0.5 or 0",
    "reasoning": "xxx"
   }
  ],
  "background": [
   {
    "item": "xxx",
    "is_present": "True/False",
    "score": "1, 0.5 or 0",
    "reasoning": "xxx"
   }
  ],
  "lighting": {
   "lighting_direction": [
    {
     "item": "xxx",
     "is_present": "True/False",
     "score": "1, 0.5 or 0",
     "reasoning": "xxx"
    }
   ],
   "brightness_level": [
    {
     "item": "xxx",
     "is_present": "True/False",
     "score": "1, 0.5 or 0",
     "reasoning": "xxx"
    }
   ],
   "lighting_effect": [
    {
     "item": "xxx",
     "is_present": "True/False",
     "score": "1, 0.5 or 0",
     "reasoning": "xxx"
    }
   ]
  },
  "color": {
   "overall_color_tone": [
    {
     "item": "xxx",
     "is_present": "True/False",
     "score": "1, 0.5 or 0",
     "reasoning": "xxx"
    }
   ],
   "contrast": [
    {
     "item": "xxx",
     "is_present": "True/False",
     "score": "1, 0.5 or 0",
     "reasoning": "xxx"
    }
   ],
   "saturation": [
    {
     "item": "xxx",
     "is_present": "True/False",
     "score": "1, 0.5 or 0",
     "reasoning": "xxx"
    }
   ]
  },
  "subject_description": [
   {
    "item": "xxx",
    "is_present": "True/False",
    "score": "1, 0.5 or 0",
    "reasoning": "xxx"
   },
   {
    "item": "xxx",
    "is_present": "True/False",
    "score": "1, 0.5 or 0",
    "reasoning": "xxx"
   }
  ],
  "relative_position": {
   "in_frame_layout": [
    {
     "item": "xxx",
     "is_present": "True/False",
     "score": "1, 0.5 or 0",
     "reasoning": "xxx"
    }
   ],
   "inter_subject_relation": [
    {
     "item": "xxx",
     "is_present": "True/False",
     "score": "1, 0.5 or 0",
     "reasoning": "xxx"
    }
   ],
   "subject_camera_relation": [
    {
     "item": "xxx",
     "is_present": "True/False",
     "score": "1, 0.5 or 0",
     "reasoning": "xxx"
    }
   ]
  }
 },
 "temporal_actions": [
  {
   "actions": [
    {
     "item": "xxx",
     "is_present": "True/False",
     "score": "1, 0.5 or 0",
     "reasoning": "xxx"
    }
   ]
  },
  {
   "actions": [
    {
     "item": "xxx",
     "is_present": "True/False",
     "score": "1, 0.5 or 0",
     "reasoning": "xxx"
    }
   ]
  }
 ]
}
```

Figure F20. Evaluation Json template used for V2T task.

Evaluation Object Prompts

Final Objective

Your task is to conduct a **comprehensive video quality evaluation** of the **GeneratedVideo**, referencing the **OriginalVideo** and **EditInstruction**.

After providing feedback, you must also indicate how each issue impacts the overall generation quality (if no issues appear, omit this part).

You must carefully examine all six dimensions below:

Evaluation Checklist

1. Content Fidelity

(1) **Subject**:

- Does the `GeneratedVideo` maintain high consistency with the `OriginalVideo` in terms of subject quantity, appearance, clothing, and details? (must consider EditInstruction)
- Are the identities of subjects maintained without abrupt changes or replacements (including whether facial features suddenly mutate, such as suddenly becoming old or changing ethnicity)?
- Are there any extra or missing key subjects?

(2) **Background**:

- Does the `GeneratedVideo` faithfully reproduce the time, location, and environmental setup of the `OriginalVideo`?
- Are there illogical or unrelated background elements (scene drift)?
- Does it restore the background style, overall layout structure, lighting effects, and environmental atmosphere of the `OriginalVideo`

(3) **Events and Logic**:

- Does the `GeneratedVideo` preserve the core events and narrative structure of the `OriginalVideo`? Is the flow natural and coherent?

2. Style Consistency & Visual Alignment

(1) **Color & Tone**:

- Does the `GeneratedVideo` match the `OriginalVideo` in `overall_tone`, `saturation`, and `contrast`?

(2) **Lighting & Atmosphere**:

- Are `lighting_direction`, `lighting_effect`, and overall brightness consistent with the `OriginalVideo`'s lighting layout and atmosphere?

(3) **Image Style**:

- Is the `image_style` consistent with the `OriginalVideo` (e.g., realistic, anime, cinematic)? Are there visually inconsistent or stylistically abrupt segments?

3. Temporal & Motion Coherence

(1) **Subject Actions**:

- Are the `subject_action_requirements` accurately and smoothly reproduced in `GeneratedVideo`? Are the action scale and rhythm consistent?

(2) **Camera Actions**:

- Are the `camera_action_requirements` (zoom, pan, dolly, etc.) consistent with `OriginalVideo`? Is the motion smooth and narratively coherent?

(3) **Transitions & Smoothness**:

- Are scene transitions and motion sequences smooth without flickering, dropped frames, or inconsistencies? Do dynamic elements maintain physical plausibility?

4. Technical Quality

(1) **Generation Quality**:

- Does the `GeneratedVideo` contain artifacts, distortions, misalignments, floating elements, or jitter?
- Is the resolution and clarity up to standard?

(2) **Consistency**:

- Do the main subjects and objects maintain visual consistency (e.g., no face or clothing changes)?

5. Artistic Expressiveness & Narrative Integrity

- Does the `GeneratedVideo` demonstrate artistic rhythm, lighting composition, and atmosphere?
- Beyond fidelity, does it achieve or even surpass the `OriginalVideo` in visual or narrative quality (Bonus criterion)

6. IP & Privacy Compliance

The `GeneratedVideo` **must not** contain recognizable copyrighted materials, including:

1. **Logos**: e.g., Nike Swoosh, Chanel logo. **Trademark Text (OCR)**: e.g., "NIKE", "GUCCI", "Bing Hong Cha". Replace with neutral terms like "BRAND" or "GENERIC PRODUCT".

Figure F21. Evaluation Json template used for V2T task.