InstanceCap: Improving Text-to-Video Generation via Instance-aware Structured Caption

Supplementary Material

In this supplementary material, we present comprehensive details and analyses across the following sections:

- **Section** 1 elucidates our methodology for constructing positive/negative lexical databases, accompanied by their details.
- **Section** 2 provides an extensive compilation of human-designed class hints, demonstrating their diverse applications.
- **Section** 3 delineates the improved Chain-of-Thought prompting strategies employed in Figure 3, with particular emphasis on their methodological improvements.
- **Section** 4 explicates the architectural framework of InstanceEnhancer, supplemented with exemplary prompts utilized in our Large Language Model implementations.
- **Section** 5 elaborates a detailed discussion of the principles behind our metric design for video reconstruction, including mathematical formulations and empirical validations.
- **Section** 6 demonstrates the prompts used by Inseval in both the inference and evaluation stages.
- **Section** 8 presents a evaluation of our methodology across both commercial and open-source models, including experimental results and analytical findings.

1. Positive/Negative Lexicon

To enhance the aesthetic quality of generated videos, we carefully collected prompts from various open-source model galleries, extracting adjectives to build a *Positive Lexicon*. Conversely, we manually constructed a *Negative Lexicon*, which was further enriched using the powerful LLM, GPT-4o. Both lexicons were refined through meticulous manual screening. The detailed contents of the Positive/Negative Lexicons are shown in Figure S1.

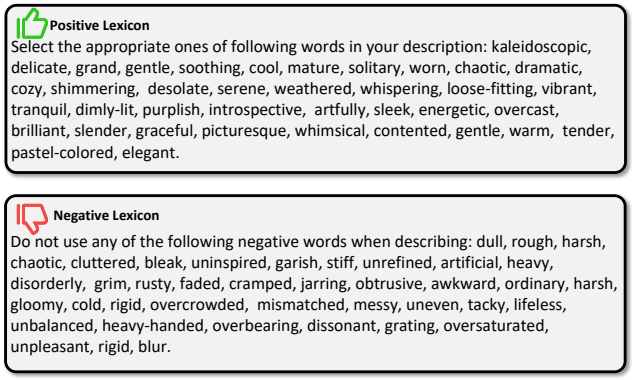

Figure S1. The detail of Positive/Negative Lexicon

2. Human-designed Class Hints

For the Human-designed Class Hints, we carefully crafted additional prompts for over *eighty* categories, each specifically tailored to its specific characteristics. Below, we present twenty of these categories. The full JSON-formatted hints for all classes, ready for direct use, will be provided in the code we plan to release later.

- **Person**: "Please focus primarily on the person's facial expressions, attire, age, gender, and race in the video and give description in detail. Please mention if there are any necklaces, watches, hat or other decoration; otherwise, there's no need to bring them up."
- **Bicycle**: "Please describe the bicycle in terms of color, type, size, condition, and any distinctive marks or decorations. Include details such as the presence of baskets, reflectors, or any branding."
- **Car**: "Please describe the car by its color, make, model, condition, license plate (if visible), and any distinguishing features such as stickers, dents, or modifications."
- **Airplane**: "Please describe the airplane by its type (commercial, private, etc.), airline brand, color scheme, size, and any visible markings such as logos or tail numbers."
- **Bus**: "Please describe the bus by its color, type (public, school, etc.), condition, any branding or advertising on its surface, and the route number or destination if visible."
- **Train**: "Please describe the train by its type (freight, passenger, high-speed, etc.), color, length, condition, and any visible logos or car numbers."
- **Truck**: "Please describe the truck by its type (pickup, semi, etc.), color, make, model, any visible logos or branding, and details such as cargo or modifications."
- **Boat**: "Please describe the boat by its type (sailboat, motorboat, yacht, etc.), size, color, condition, and any identifying features like registration numbers or flags."
- **Traffic Light**: "Please mention the current state of the traffic light (red, yellow, green), its location, and any additional details like the presence of pedestrian signals."
- **Fire Hydrant**: "Please describe the fire hydrant by its color, condition, and any notable features such as signs, markings, or proximity to other objects."
- **Stop Sign**: "Please describe the stop sign's condition, location, and any visible obstructions or markings on it."
- **Parking Meter**: "Please describe the parking meter by its condition, type (modern, traditional), and any visible

information like pricing or operational status."

- **Bench**: "Please describe the bench by its material, color, condition, and any distinctive features such as inscriptions, decorations, or nearby objects."
- **Bird**: "Please describe the bird by its species (if identifiable), color, size, behavior, and any unique markings or features."
- **Cat**: "Please describe the cat by its color, breed (if identifiable), size, behavior, and any distinguishing features such as collars or patterns."
- **Dog**: "Please describe the dog by its breed (if identifiable), color, size, behavior, and any accessories such as collars or leashes."
- **Horse**: "Please describe the horse by its color, breed (if identifiable), size, behavior, and any accessories such as saddles or reins."
- **Sheep**: "Please describe the sheep by its color, size, behavior, and any distinguishing features such as markings or tags."
- **Cow**: "Please describe the cow by its color, breed (if identifiable), size, behavior, and any distinguishing features such as tags or markings."
- **Elephant**: "Please describe the elephant by its size, tusk length, condition, and any unique features such as markings or behavior."

3. Prompt Design of Figure 3

System prompt. Referring to ShareGPT4Video [3], we divided the System prompt into three parts. Through extensive tests on challenging samples, including multi-instance, complex scenes, and high-intensity motion, we finalized the system prompt shown in Figure S6. Additionally, temporal metadata extracted using the code provided in Figure S7.

Prompts of global description, background detail and camera movement. The global description is derived from a single prompt: *"Please describe this video in one sentence, no more than 20 words."*. To illustrate the acquisition of camera motion and background details, we provide an example of implementing camera hints with movement cues in Figure S8. A similar approach is used for extracting background details included in our code released later.

Prompts of structured caption. In the structured caption section, we we use Actions and Motion as examples, with the CoT prompt shown in Figure S9. The acquisition of Appearance and the injection of Human-designed class hints follow a similar approach.

4. Design of InstanceEnhancer

In InstanceEnhancer, prompt alignment during inference is achieved through a two-stage process (Figure S2). To provide more precise instructions to LLMs, we meticulously designed multiple examples as part of the CoT, which are fed into the LLMs. An example of this is shown in Figure S10.

5. Evaluation Metrics for Video Reconstruction

3DVAE score (3DVAE$_{score}$). The LIPIPS score [35] which is widely used to evaluate image reconstruction quality, measures perceptual distance between ground truth (GT) and reconstructed images. We extent this concept for video data by using 3DVAE [32] to extract latent-space video representations from both GT videos and their caption-reconstructed versions. 3DVAE$_{score}$ computes the distance between latent representations across spatial and temporal dimensions:

$$d(\mathbf{z}_{\text{GT}}, \mathbf{z}_{\text{rec}}) = \sum_{l} \sum_{t} \sum_{h,w} \left\| w_l \odot \left(\mathbf{z}^{l}_{\text{GT},hwt} - \mathbf{z}^{l}_{\text{rec},hwt} \right) \right\|_2^2 \tag{1}$$

where $\mathbf{z}^{l}_{\text{GT},hwt}$ and $\mathbf{z}^{l}_{\text{rec},hwt}$ represent the latent representations at layer l, spatial location (h, w), and temporal frame t, with w_l as the layer-specific weight matrix. We set $(h, w, t) = (224, 224, 8)$ for evaluation.

To ensure consistency, we use the same video generation model across all captioning methods. Following LIPIPS methodology, we validate the 3DVAE score by comparing GT videos against various distorted versions. As shown in Tab. S1, the results demonstrate that our score effectively captures perceptual similarities between GT and reconstructed videos.

Distortion type	3DVAE score↓	Setting
Blurring	7.71	GaussianBlur(kernel=(5, 5), sigma=0)
Compression artifacts	11.19	JPEG compression (quality 5-30)
Corruptions	39.80	Random pixel masking (binary mask)
Random noise	49.70	Gaussian noise (mean=0, stddev=25)
Brightness distortion	63.25	Scaling (factor 0.5-1.5)
Spatial shifts	78.94	Random affine shifts (±10 pixels)
T2V models Avg.	134 ∼ 145	-
Broken video	149.50	-

Table S1. 3DVAE scores for various distortions and video models, showcasing its effectiveness in capturing perceptual similarities and reconstruction accuracy. The setting column provides details of the experimental setup for each distortion type.

Human evaluation. Automated machine-based scoring systems, while offering enhanced objectivity and efficiency, often fail to align with human preferences or fully grasp the nuances of context and meaning in a given task. To ensure a comprehensive and balanced evaluation, we adopted a human-based assessment framework. This evaluation is

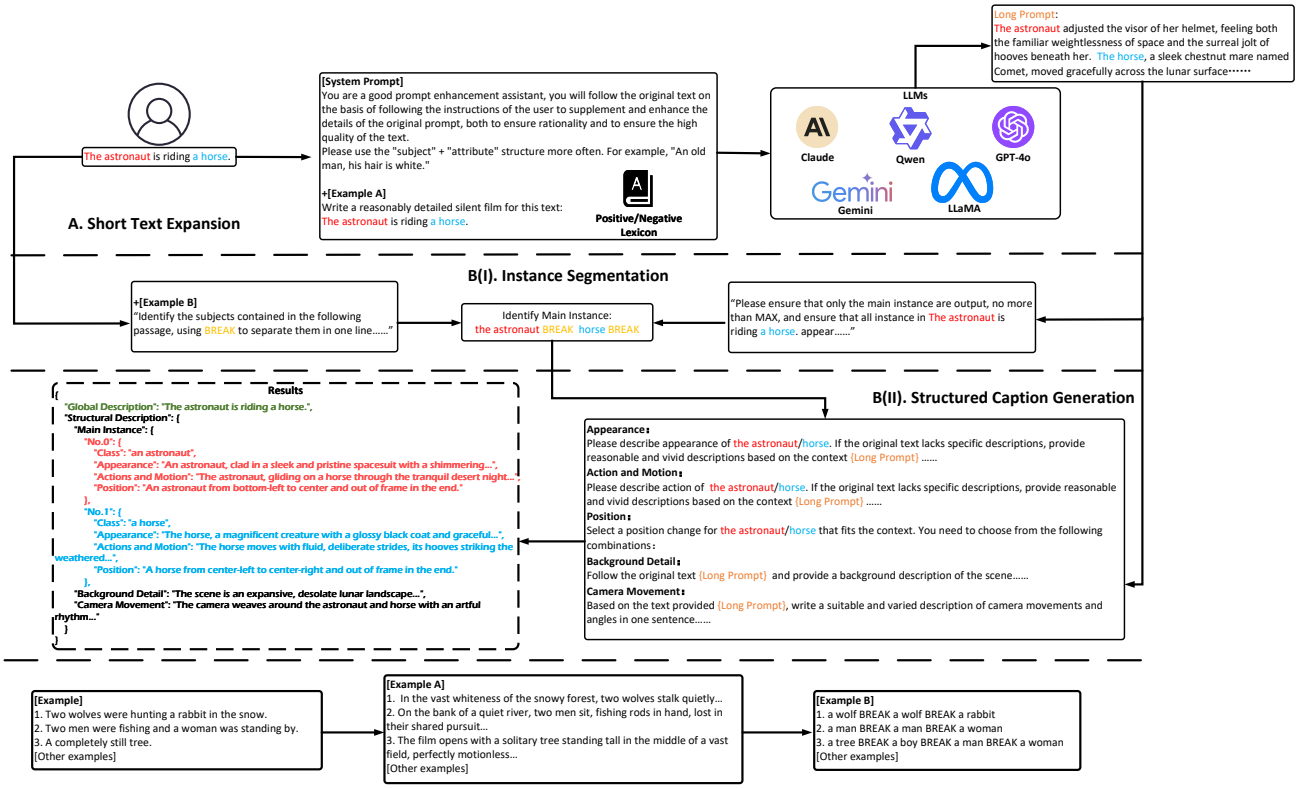

Figure S2. Detailed overview of the InstanceEnhancer pipeline. Example No.1 as shown in Figure S10.

Single

Action

{"sentence": "A tall man stands at the left.", "instance": {"class": "man", "action": "standing"}}

Color

{"sentence": "A neon pink pen lies on the desk.", "instance": {"class": "pen", "color": "neon pink"}}

Shape

{"sentence": "A round chair is pushed against the far left corner of the desk.", "instance": {"class": "chair", "shape": "round"}}

Texture

{"sentence": "A rough towel hangs on the rack beside the bathroom door.", "instance": {"class": "towel", "texture": "rough"}}

Detail

{"sentence": "A fluffy gray cat with bright green eyes sits on the windowsill, gazing outside at the fluttering leaves in the breeze. Its tiny pink nose twitches as it watches a bird land on a nearby branch, and a distinct purple stripe runs down its belly, adding to its unique charm. A small silver bell dangles from its ear, jingling softly with each movement.", "instance": {"detail": "a distinct purple stripe runs down its belly", "class": "cat"}}

Multiple

Action

{"sentence": "A horse running as a butterfly flutters by.", "instance": {"0":{"class": "horse", "action": "running"}, "1":{"class": "butterfly", "action": "fluttering"}}}

Color

{"sentence": "A sapphire blue keychain hangs from a light brown bag.", "instance": {"0":{"class": "keychain", "color": "sapphire blue"}, "1":{"class": "bag", "color": "light brown"}}}

Shape

{"sentence": "A square speaker lies on a round shelf.", "instance": {"0":{"class": "speaker", "shape": "square"}, "1":{"class": "shelf", "shape": "round"}}}

Texture

{"sentence": "A satin pillow sits on a faux leather sofa.", "instance": {"0":{"class": "pillow", "texture": "satin"}, "1":{"class": "sofa", "texture": "faux leather"}}}

Detail

{"sentence": "A fluffy gray cat with a purple stripe down its belly sits on the windowsill. Beside the fireplace, a golden retriever with a heart-shaped birthmark on its front paw rests comfortably.", "instance": {"0": {"detail": "a distinct purple stripe on its belly", "class": "cat"}, "1": {"detail": "a heart-shaped birthmark on its front paw", "class": "retriever"}}}

Figure S3. Inference examples of Inseval.

carried out across several key dimensions, including: 1) Instance Detail (**ID**): Evaluate whether the text provides accurate descriptions of the details of the examples in the video. 2) Intrinsic Hallucination: Evaluate whether the text hallucinates descriptions of things present in the video. 3) Extrinsic Hallucination: Evaluate whether the text introduces content that is not present in the video. For convenience, the latter two have been combined into a single metric called the Hallucination Scores (**HS**) [9]. The specific guidelines and scoring criteria for each metric refers to Table S2.

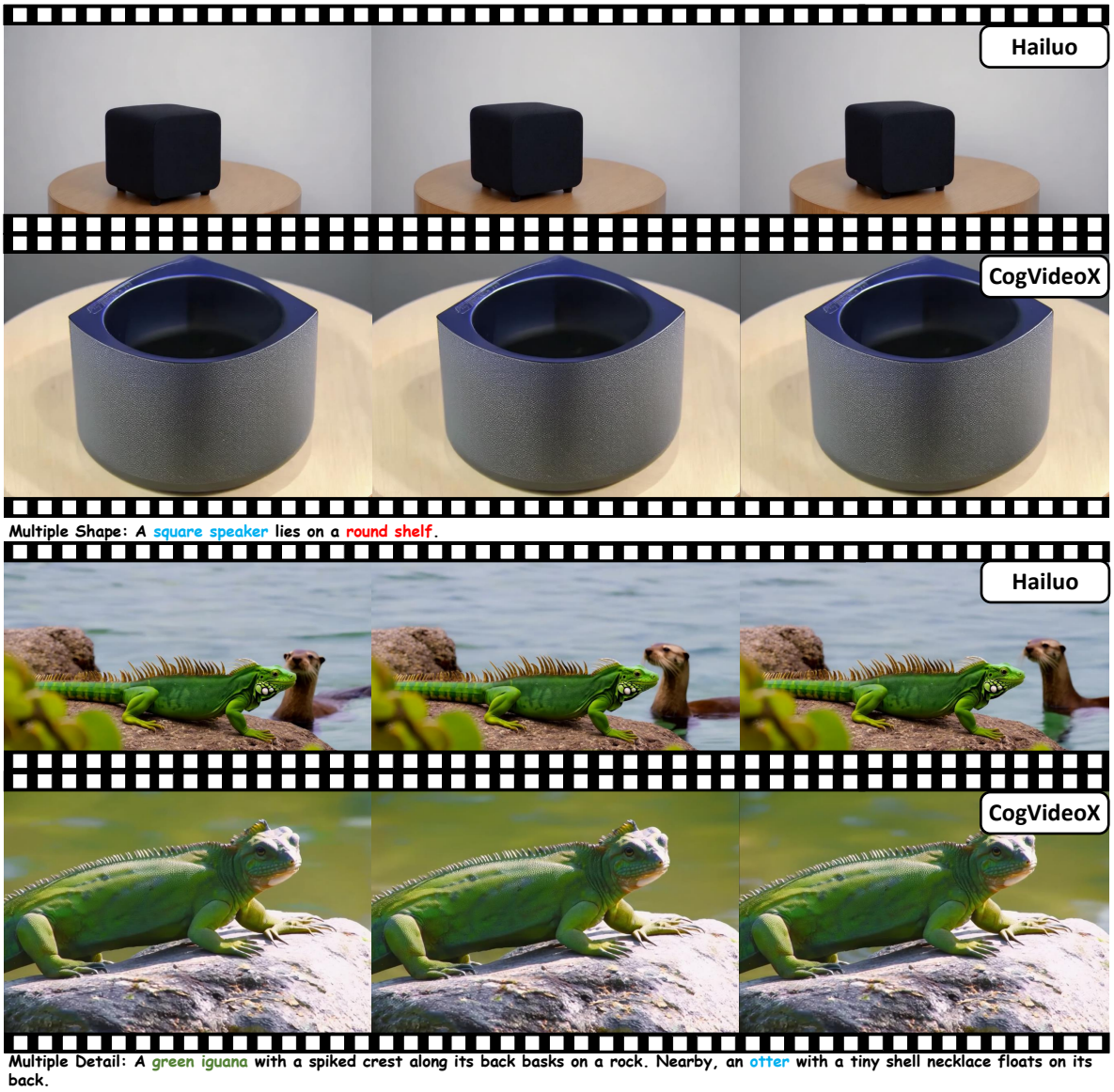

Figure S4. Visualization comparing open-source models and commercial models on prompts with poorer performance.

6. Inseval

Inference prompts of Inseval. In implementing Inseval, we designed multiple prompts to test each dimension, as illustrated in Figure S3. To further evaluate the model's generative capabilities and instruction-following accuracy, we deliberately included some "counter-intuitive" shapes in the prompt design.

Evaluation prompts of Inseval. For the evaluation, we used a general CoT Q-A pair format (with a slightly different design for the 'Detail' dimension, shown in Figure S11 to assess whether the MLLMs successfully matched the generated videos to the corresponding dimensions, as outlined in the specific code. In single-object scenarios, the success rate is calculated as the percentage of correctly matched prompts. In multi-object scenarios, the generation is deemed successful only if all targets meet the requirements. For reproducibility, fixed random seeds are used during generation and evaluation.

In Table 2, the 'Shape' and 'Detail' dimensions under Multiple category are omitted due to consistently very poor performance across all tested models. Even CogVideoX-5B, the overall best performer, struggles with multi-object tasks in these dimensions, as shown in Figure S4. Two primary error types are observed in Multiple Shape tasks:

Instance Detail		Hallucination Scores	
1	Descriptions are extremely vague, imprecise, or largely inaccurate. Almost no specific details from the video are captured correctly.	**1**	Severe hallucination - Describes many nonexistent details, significantly misrepresents what is shown, or introduces extensive irrelevant content with many unrelated topics or external information.
2	Descriptions have major inaccuracies or omit many important details. Only a few basic elements are described correctly.	**2**	Frequent hallucination - Multiple instances of fabricated or misrepresented details and significant extra content introducing information beyond the video scope.
3	Descriptions are moderately accurate but lack precision in some areas. Core details are present but some secondary details are missing or incorrect.	**3**	Occasional hallucination - A few minor instances of fabricated details, misrepresentations, or the addition of extra content not covered in the video.
4	Descriptions are largely accurate and detailed. Most key elements and nuances from the video are captured correctly, with only minor omissions or imprecisions.	**4**	Minimal hallucination - One or two very minor discrepancies or limited introduction of external information.
5	Descriptions are highly precise and comprehensive. All important details from the video are captured accurately, including subtle elements and specific examples.	**5**	No hallucination - All described details accurately reflect what is shown in the video, with no external content added.

Table S2. This table outlines scoring criteria for Instance Detail and Hallucination Scores, integrating intrinsic and extrinsic hallucinations into a unified framework for evaluation.

attribute confusion (**Top** case) and failure to follow multiple target instructions (**Bottom** case), where targets are either missing or rendered incorrectly. Commercial models demonstrate relatively better performance, which we further analyze in Section 8.

7. Statistical analysis of InstanceVid

Figure S5 illustrates the statistical characteristics of `InstanceVid`across two main dimensions: video scenes, and temporal durations. Our data collection emphasizes videos with distinct instances while ensuring a balanced representation of outdoor scenes to prevent biases from an overemphasis on instance-focused content. We achieve detailed descriptions capturing human movements, physical appearances, and documentation of common objects and animals. Besides, `InstanceVid` focuses on short-duration videos (2-10 seconds) for two main reasons. First, OpenVid-1M segments longer sequences to eliminate excessive scene transitions. Second, most of the current open-source T2V models are optimized for video generation within this duration range.

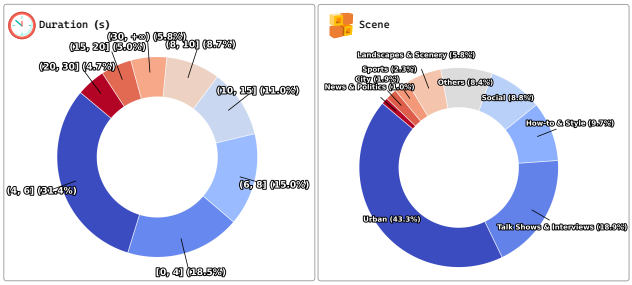

Figure S5. `InstanceVid` provides structured captions for videos in open-domain scenarios, featuring diverse instance, expansive scenes, precise and instance-aware captions, and video-generation-friendly durations.

8. Analysis on Commercial Products vs. Open-source Models

Prompt processing analysis. Commercial T2V products excel at processing complex input prompts, effectively handling long-form text in structured formats while preserving semantic coherence. They can seamlessly interpret detailed scene descriptions, character interactions, and sequential events within a single prompt, producing coherent visual narratives, have shown surprising results in many situations.

Open-source T2V models, however, are *unable to directly process long-text structured prompts*, requiring an ad-

ditional alignment step (Figure S12). This preprocessing can lead to potential information loss and inconsistencies in the final output, restricting the ability to capture nuanced details from the original prompt.

Information retention capabilities. Different models exhibit notable differences in information retention (Figure S4). Commercial products (*e.g.*, Hailuo AI) excel in maintaining fidelity between text and visual content, effectively preserving detailed instructions and translating multiple attributes into video sequences. This strength is particularly apparent when our caption contains *complex scenes* that demand temporal consistency and fine-grained details.

Open-source models face challenges in consistently representing instance information (Figure S4), exhibiting variability in detail preservation and limited capability with complex attribute combinations. These shortcomings are particularly evident when processing prompts with multiple interrelated instances or maintaining consistent visual characteristics across temporal sequences.

System Prompt

 # Character

You are an excellent video frame analyst. Utilizing your incredible attention to detail, you provide clear, sequential descriptions for video frames. You are good at identifying and describing the properties of each target in the video frame, the actions and movement.

 # Skills

Skill 1: Describing Objects Appearances

- Describe the appearances of instance.
- Determine which parts are colored parts, as the goal of the main description.
- Focus mainly on the color part, the black and white part only as an auxiliary role.
- Highly sensitive to person, describe they in detail, such as the style and color of hat, the style and color of clothes, age, gender, body type, expression, etc.

Skill 2: Describing Objects' Actions and Behaviors

- Elaborate the action of instance.
- Notice and describe changes in the actions or behaviors.
- Determine which Objects are main instance and give more detailed description.

Skill 3: Use Fine Words to Describe.

- Select the appropriate ones of following words in your description: kaleidoscopic, delicate, grand, gentle, soothing, cool, mature, solitary, worn, chaotic, dramatic, cozy, shimmering, desolate, serene, weathered, whispering, loose-fitting, vibrant, tranquil, dimly-lit, purplish, introspective, artfully, sleek, energetic, overcast, brilliant, slender, graceful, picturesque, whimsical, contented, gentle, warm, tender, pastel-colored, elegant.
- State facts objectively without using any rhetorical devices such as metaphors or personification.
- Do not use any of the following negative words when describing: dull, rough, harsh, chaotic, cluttered, bleak, uninspired, garish, stiff, unrefined, artificial, heavy, disorderly, grim, rusty, faded, cramped, jarring, obtrusive, awkward, ordinary, harsh, gloomy, cold, rigid, overcrowded, mismatched, messy, uneven, tacky, lifeless, unbalanced, heavy-handed, overbearing, dissonant, grating, oversaturated, unpleasant, rigid, blur.

 # Constraints

- State facts objectively without using any rhetorical devices such as metaphors or personification.
- Exclude sounds-related aspects, given the unavailability of audio signals.
- Descriptions should be fluent and precise, avoiding analyzing and waxing lyrical.
- Descriptions need to be concise, describing only the information that can be determined, without analysis or speculation.
- Do not mention the frame number and timestamp of the current frame.
- The main object will occupy most of the content of the picture, and there may be more than one main object, and there may be no main object in the landscape type of video.
- Only the main object needs to be described in detail, and the other objects only need to be described briefly.
- Strictly follow the format of the structured output, containing all of its elements.

Figure S6. System prompt of `InstanceCap`.

```python
def load_video(video_path, max_frames_num,fps=1,force_sample=False):
    if max_frames_num == 0:
        return np.zeros((1, 336, 336, 3))
    vr = VideoReader(video_path, ctx=cpu(0),num_threads=1)
    total_frame_num = len(vr)
    video_time = total_frame_num / vr.get_avg_fps()
    fps = round(vr.get_avg_fps()/fps)
    frame_idx = [i for i in range(0, len(vr), fps)]
    frame_time = [i/fps for i in frame_idx]
    if len(frame_idx) > max_frames_num or force_sample:
        sample_fps = max_frames_num
        uniform_sampled_frames = np.linspace(0, total_frame_num - 1, sample_fps, dtype=int)
        frame_idx = uniform_sampled_frames.tolist()
        frame_time = [i/vr.get_avg_fps() for i in frame_idx]
    frame_time = ",".join([f"{i:.2f}s" for i in frame_time])
    spare_frames = vr.get_batch(frame_idx).asnumpy()
    return spare_frames,frame_time,video_time

def build_video_prompt(image_processor, video, frame_time, video_time):
    video = image_processor.preprocess(video, return_tensors="pt")["pixel_values"].cuda().bfloat16()
    video = [video]
    time_instruciton =
    f"The video lasts for {video_time:.2f} seconds, and {len(video[0])} frames are uniformly sampled from it. "\
    f"These frames are located at {frame_time}.Please answer the following questions related to this video."
    pre_instruction = DEFAULT_IMAGE_TOKEN + f"{time_instruciton}\n"

    return video, pre_instruction
```

Figure S7. Code of getting video temporal metadata.

Camera Movement Prompt

 # User

Let's think step by step... Try to separate the camera movement from the video.

if camera_motion == "Undetermined":
The motion of the video camera is very complex, can you infer the possible motion of the camera and the shooting Angle (long distance/medium distance/overhead Angle/POV, etc.) from the changes in the video?
elif camera_motion == "static":
Is the camera static or moving? Can you deduce the possible motion of the camera and the shooting Angle (long distance/medium distance/overhead Angle/POV, etc.) in the video?"
else:
The motion of the camera in this video is {camera_motion}. According to my tips, can you deduce the possible motion of the camera and the shooting Angle (long distance/medium distance/overhead Angle/POV, etc.) in the video?

Summarize the camera movement and shooting angle, use degree adverbs appropriately(Sharply, rapidly, slowly, etc), try to give description in detail.

Figure S8. Prompt of camera movement.

Actions and Motion

STEP 1

 # USER

Let's think step by step...
Can you tell what color the background of this video is? Unimportant parts have been intentionally obscured with a mosaic; please focus only on the clearly visible portions.
If I remind you that there's a {target_object}, can you see it in the video?
This is supplementary information to the full video to aid your description:
{global_description}
Can you find the {target_object} corresponding to this paragraph in the video?

 # ASSISTANT

Answer of STEP 1

STEP 2

 # USER

Continue ignore mosaic background, if {target_object} has any movement, answer what you think {target_object} is doing in video.
Extract the information related to {target_object} that you are currently describing.
Do not answer any information unrelated to {target_object}'s state of motion or you're not sure. Observe carefully, use appropriate adjectives, and give description in detail.
Do not answer in bullet points. Do not mention any objects that are not related to {target_object}.
Do not speculate about the environment in which the {target_object} is located.
Do not mention anything about a \"blurred background.\" Use a continuous paragraph.

 # ASSISTANT

Answer of STEP 2(Final answer)

Figure S9. Prompt of actions and motion.

Example Prompt of Stage A&B(I)

Stage A. Short Text Expansion

USER

Write a reasonably silent film for this text: Two wolves were hunting a rabbit in the snow. **(Short prompt)**

ASSISTANT (human-provided)

In the vast whiteness of the snowy forest, two wolves stalk quietly. One is large and heavyset, its fur a mix of dark grays and black, blending into the shadows of the trees. The other is leaner and quicker, with a pale, almost silver coat that glimmers in the winter light. Their eyes are sharp and focused, and each step is measured—one wolf pacing powerfully, the other darting with agile precision. Ahead of them, a small white rabbit crouches in the snow, nearly invisible but for the black tips of its ears and nose. Suddenly, the rabbit bolts, kicking up a spray of powder. The dark wolf charges forward with brute force, while the silver one sprints with smooth, loping strides. The rabbit's movements are frantic and unpredictable, weaving between tree roots and leaping over snowdrifts. The wolves flank it from both sides, their breaths clouding the air, closing in like shadows. At the last moment, the rabbit dives into a narrow burrow, disappearing into safety. The dark wolf sniffs the ground, frustrated, while the silver wolf paces briefly before the two melt back into the forest, leaving only their crisscrossing prints behind in the snow. **(Long prompt)**

Stage B(I). Instance Segmentation

USER

Identify the subjects contained in the following passage, using BREAK to separate them in one line:
Long prompt
Please distinguish between instances and scenes, and return only instances, such as: "a man BREAK two cat BREAK a cup..."Please ensure that only the main instance are output, no more than X, and ensure that all objects in **Short prompt** appear.
Note that the instance you extract must be an entity that can be touched.
When there are multiple targets, you need to give them separately.

ASSISTANT (human-provided)

a wolf BREAK a wolf BREAK a rabbit

Figure S10. Designed example for LLMs.

Evaluation Prompt(Single)

STEP 1

USER

Describe this video in one sentence, care about the {dimension_type}.

ASSISTANT

Answer of STEP 1

STEP 2

USER

Tell me if '{instance_class}' is in the video? Note the synonyms of '{instance_class}'. Your answer can only be YES or NO.Do not output any answer that is not YES or NO.

ASSISTANT

Answer of STEP 2(if answer is"No", return False)

STEP 3

USER

Based on your previous answer, tell me what is the '{dimension_type}' of '{instance_class}' in the video? Be careful to ignore camera movement.

ASSISTANT

Answer of STEP 3

STEP 4 of Others

USER

Do you think the '{dimension_type}' of '{instance_class}' in the video Approximatly close to to '{instance_specific}'?
Your answer can only be YES or NO. Do not output any answer that is not YES or NO.

ASSISTANT

Answer of STEP 4(Final answer)

STEP 4 of Detail

USER

Is the '{instance_specific}' of '{instance_class}' partly reflected in the video? Your answer can only be YES or NO. Do not output any answer that is not YES or NO.

ASSISTANT

Answer of STEP 4(Final answer)

Figure S11. Evaluation prompts of Inseval.

Aligning Prompt

Step 1

User

Let's think step by step...
Read the following JSON and summarize it to continuous text paragraph, ensuring that all main ideas and crucial details are preserved:
{InstanceCap}
What you need to pay attention to is the "Global Description" and "Structural Description" sections. The "Global Description" provides an overall summary of the video, while the "Structural Description" contains detailed information about various aspects of the video, such as main characters, background details, and camera movements. Please focus on the details in the "Structural Description" and combine them with the "Global Description" to create a summary.
Select the appropriate ones of following words in your description: kaleidoscopic, delicate, grand, gentle, soothing, cool, mature, solitary, worn, chaotic, dramatic, cozy, shimmering, desolate, serene, weathered, whispering, loose-fitting, vibrant, tranquil, dimly-lit, purplish, introspective, artfully, sleek, energetic, overcast, brilliant, slender, graceful, picturesque, whimsical, contented, gentle, warm, tender, pastel-colored, elegant.
Do not use any of the following negative words when describing: dull, rough, harsh, chaotic, cluttered, bleak, uninspired, garish, stiff, unrefined, artificial, heavy, disorderly, grim, rusty, faded, cramped, jarring, obtrusive, awkward, ordinary, harsh, gloomy, cold, rigid, overcrowded, mismatched, messy, uneven, tacky, lifeless, unbalanced, heavy-handed, overbearing, dissonant, grating, oversaturated, unpleasant, rigid, blur.

Step 2

User

1. Please use the "subject" + "attribute" + "position" structure more often. For example, "An old man, his hair is white."
2. Please tell me the content of the video directly, don't use "The video shows..." Or other similar forms, you should begin with a direct description of the content, for example:"An old man..."
3. If there are multiple objects, such as people, introduce them with phrases like "A man... and a woman..., and a man...","A car..., and a car...".
4. When you need to describe The background in detail, use "The scene is..." As an opening sentence.
5. Summarize it to approximately 180 words

Figure S12. Aligning prompt used during alignment with the open source model.