

## 1003 A Appendix

|      |  |    |
|------|--|----|
| 1004 | Taxonomy Details                                     | 24 |
| 1005 | Additional Results and Analysis                      | 25 |
| 1006 | Details on Prompt Generation                         | 33 |
| 1007 | Distribution of Taxonomy Categories in SCINE Prompts | 36 |
| 1008 | Annotation Details                                   | 36 |
| 1009 | Statistical Tests                                    | 39 |
| 1010 | Additional VLM Results                               | 39 |
| 1011 | Limitations  | 43 |
| 1012 | Broader Impact                                       | 43 |

### 1013 A.1 Taxonomy Details

1014 We provide additional details on control nodes and their associated values in the taxonomies. Some  
1015 nodes accept open-ended values, for example, a range of stand-alone actions. To simplify evaluation,  
1016 we abstract certain values that can be fine-grained in future work. For instance, we group Aperture  
1017 into wide/medium/narrow, though exact f-stop values can also be studied in the future. Similarly,  
1018 color palette is treated as a discrete value, but can be decomposed into hue, brightness, and  
1019 saturation. Table 3 - 6 details the control nodes and their values of the Camera, Lighting, Setup and  
1020 Events Taxonomies, respectively.

Table 3: Camera Taxonomy Control Nodes and Values

| Name           | Description   | Potential Values  |
|----------------|---|---|
| Lens Size      | Defines the focal length and field of view of the camera lens.                                | Standard, Fisheye, Wide, Medium, Long Lens, Telephoto   |
| Depth of Field | Controls the range of focus in the image, affecting subject isolation.                        | Deep, Shallow, Soft, Rack, Split Diopter, Tilt Shift  |
| Aperture       | The camera lens opening that controls the amount of light propagated through the camera.      | Wide, Medium, Narrow  |
| Shutter Speed  | The duration for which the camera sensor is exposed to light.                                 | Slow, Medium, Fast  |
| ISO            | Sensitivity of the camera sensor to light.  | Low, Medium, High   |
| Angle          | Defines the camera’s viewpoint in relation to the subject.                                    | Low, High, Aerial, Overhead, Dutch, Eye-level, Shoulder, Hip, Knee, Ground, Continuous Values                                     |
| Static         | A fixed camera position without any movement.   | -   |
| 2D             | Camera movements restricted to horizontal or vertical axes.                                   | Pan left, Pan right, Tilt up, Tilt down, Zoom in, Zoom out  |
| 3D             | Camera movements that incorporate spatial depth and multi-axis motion.                        | Push In, Pull Out, Dolly Zoom, Camera Roll, Tracking, Trucking, Arc, Crane  |
| Gear           | Specifies the support systems and stabilization equipment used to facilitate camera movement. | Handheld, Tripod, Pedestal, Cranes, Overhead Rigs, Dolly, Stabilizer, Snorricam, Vehicle Mount, Drones, Motion Control, Steadicam |
| Shot Size      | Determines how much of the subject and surroundings are visible in the frame.                 | Establishing, Master, Wide, Full, Medium-Full, Medium, Medium-Close-up, Close-up, Extreme Close-up                                |
| Framing        | Placements and composition of subjects within the frame.                                      | Single, Two Shot, Crowd, OTS, PoV, Insert   |

Table 4: **Lighting Taxonomy Control Nodes and Values**

| Name                        | Description   | Potential Values  |
|-----------------------------|---|---|
| Natural Light               | Natural sources of light, such as sunlight, moonlight, or firelight.                    | Sunlight, Moonlight, Firelight  |
| Artificial/Practicals Light | Man-made light sources that illuminate the scene.                                       | LED, HMI, Tungsten, Fluorescent, HID                                  |
| Color Temperature           | Defines the hue of the light, typically measured in Kelvin, affecting the scene's mood. | Warm, Cool, Cold  |
| Lighting Conditions         | Describes various lighting scenarios or ambient conditions present in a scene.          | Candlelight, Golden Hour, White Fluorescent, Clear Daylight, Overcast |
| Soft Shadows                | Subtle and diffused shadows resulting from indirect or scattered light.                 | Diffused Light, High Key Lighting, Reflectors                         |
| Hard Shadows                | Sharp, well-defined shadows generated by a direct light source.                         | Direct Light, Low Key Lighting  |
| Reflection                  | The effect of light bouncing off surfaces to create a reflective appearance.            | -   |
| Lighting Position           | Specifies the placement or direction of the light source relative to the subject.       | Back Light, Fill Light, Top Light, Side Light, Key Light              |
| Motion                      | Dynamic changes or movement in the lighting effect.                                     | Flickering, Pulsing   |
| Color Gels                  | Colored filters applied to lights to modify or enhance the color of the illumination.   | -   |

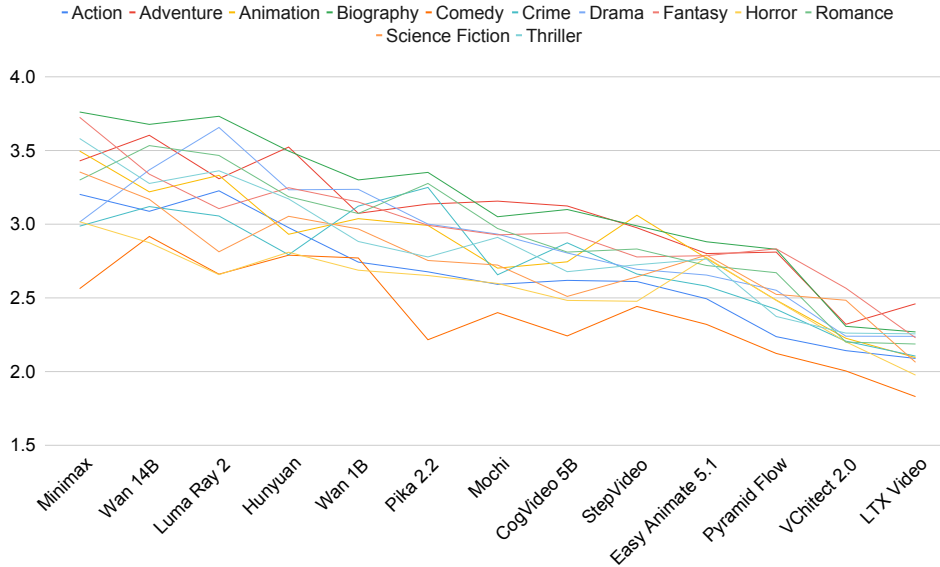


Figure 11: **Model performance on Events across genres.** Across 13 models and 12 genres, Minimax performs best overall; Adventure is the strongest genre, while Comedy is the weakest.

## A.2 Additional Results and Analysis

### A.2.1 Events

Figure 11 shows Events performance across 13 models and 12 genres. Biography and Adventure are strongest whereas Comedy and Horror are the weakest. Minimax leads in 6/12 genres, Luma Ray 2 tops Action and Drama, and WAN-14B is the most consistent, with the lowest standard deviation. Figure 12 shows Events performance across 13 models and 6 Temporal portrayal of Actions. Models handle atomic and concurrent actions well, but struggle with causal, overlapping, and cyclic events.

Figure 13 shows performance of 10 open-source models across 19 emotion classes. Models perform best on remorse and ecstasy, but fare poorly on aggressiveness and rage. As shown in Figure 14, dialogue performance is the weakest in comparison to Emotions and Actions. Models particularly struggle with multi-turn dialogues or when non-verbal reactions follow. Since T2V models do not generate audio, we evaluate whether the correct character delivers the line and/or with appropriate visual expression. Most models fail to localize the speaker, often attributing a single dialogue to multiple characters.

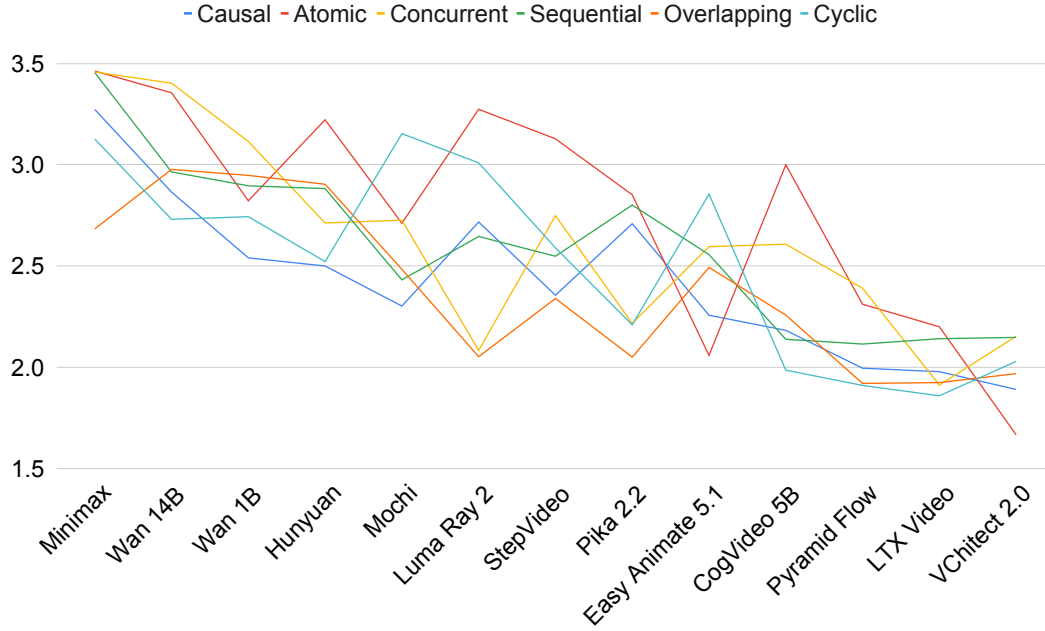


Figure 12: **Model performance on Events across temporal portrayal of Actions.** Atomic actions are handled well, whereas models struggle with causal and overlapping Events.

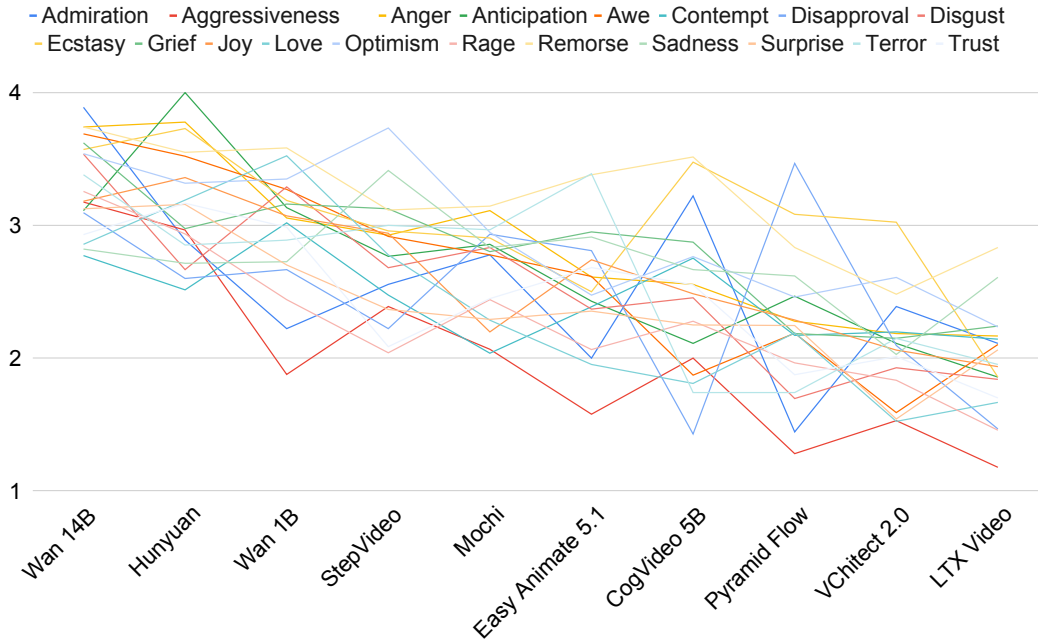


Figure 13: **Model performance on Emotions.** Among 10 models and 19 emotions, Remorse is best portrayed, while Aggressiveness is the weakest.

Table 5: Setup Taxonomy Control Nodes and Values

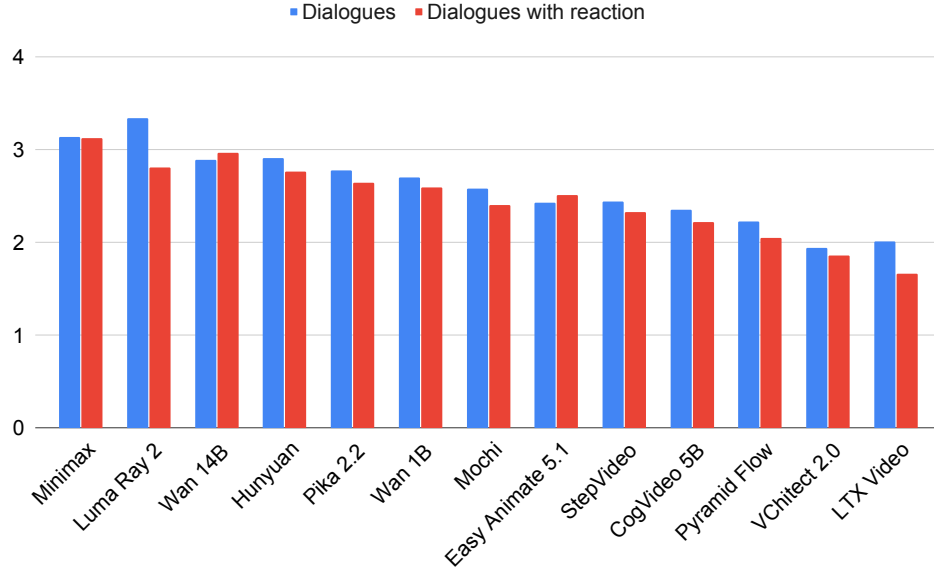
| Name                 | Description   | Potential Values   |
|----------------------|---|--|
| Contrast             | Determines the difference between light and dark areas to enhance visual impact.                            | Low, High  |
| Blur                 | Introduces softness to parts of the image to guide focus or create mood.                                    | Gaussian, Radial, Motion   |
| Noise                | Adds random variations in brightness or color, mimicking film grain or digital sensor noise.                | Gaussian, Salt and Pepper, Poisson   |
| Film Grain           | Emulates the granular texture of traditional film photography for a classic look.                           | -  |
| Color Palette        | Defines the overall range and harmony of colors in the scene, influencing its mood.                         | Open Set   |
| Lines                | Directional elements that guide the viewer’s gaze within the shot.  | Horizontal, Vertical, Diagonal   |
| Regular Shapes       | Structured, geometric forms such as squares, circles, and triangles that add order to the design.           | Square, Circle, Triangle   |
| Natural Shapes       | Unstructured shapes that naturally emerge in the scene, without any geometric constraints.                  | Water-like, Cloud-like   |
| Frame Balance        | Refers to the distribution of visual weight across the composition, ensuring a harmonious layout.           | Rule of Thirds, Symmetry, Right Heavy, Left Heavy  |
| Positional Accuracy  | The absolute position of an object or a subject in a scene.   | Open Set   |
| Relative Positioning | The relative positioning of an object in relationship to other objects in the scene.                        | Open Set   |
| Depth                | Controls the perception of distance between elements, enhancing the three-dimensional feel of the scene.    | Deep, Flat, Limited, Ambiguous   |
| Setting              | Defines if the scene is happening indoors or outdoors.  | INT/EXT  |
| Time of Day          | The time of day the scene is set in.  | DAY, NIGHT, MORNING, EVENING, DAWN, DUSK, LATE NIGHT, MIDDAY, SUNRISE, SUNSET, AFTERNOON |
| Location             | The specific place or setting of the scene.   | Open Set   |
| Negative Space       | Defines if there a lot of empty space.  | -  |
| Positive             | Defines how the space is occupied in the environment.   | Clean, Cluttered   |
| Mood                 | The emotional atmosphere or feeling created by the environment.   | Open Set   |
| Scale                | The relative size or extent of the environment.   | Open Set   |
| Style                | The artistic or visual style of the backdrop.   | Open Set   |
| Background           | The part of the scene that is behind the main subject and does not need to be exactly described.            | Open Set   |
| Elements             | The natural or artificial components of the backdrop.   | Rain, Snow, Fog, Wind, Thunder, Smoke, Dust, Ash, Fire                                   |
| Prop Description     | A general description of the prop.  | Open Set   |
| Prop Class           | The category or type of the prop.   | Open Set   |
| Prop Material        | The substance(s) the prop is made of.   | Wood, Glass, Gold, Paper, Plastic  |
| Prop Pattern         | The design on the prop.   | Grid, Checker, Stripes, Zigzag, Dots, Bricks, Metal, Hexagons                            |
| Prop Utility         | The purpose or function of the prop, whether it just exists in the scene or will it be used by the subject. | Decorative, Functional   |
| Subject Class        | The category or type of the subjects.   | Open Set   |
| Subject Accessories  | Items worn or carried by the subjects that enhance their appearance or functionality.                       | Open Set   |
| Subject Costume      | The clothing worn by the subjects, especially for a performance or to create a specific character.          | Open Set   |
| Subject Hair         | The style and appearance of the subjects’ hair.   | Open Set   |
| Subject Makeup       | Cosmetics applied to the subjects’ face or body to enhance or alter their appearance.                       | Open Set   |
| Subject Pose         | The position or stance of the subjects, especially for a photograph or portrait.                            | Open Set   |
| Subject Silhouette   | The outline or shape of the subjects against a light background.  | Open Set   |
| Subject Proportions  | The relative size and scale of the subjects’ body parts or features.  | Open Set   |
| Text Generation      | The process of creating written content to be displayed on the video.                                       | Open Set   |

### 1035 A.2.2 Camera and Lighting

1036 For Camera and Lighting, we compare model performance across subcategories: Camera Creative  
1037 Intent (Figure 15), Camera Movement (Figure 16), and Lighting Source (Figure 17). Notable  
1038 differences emerge within a parent node: models handle shot size better than framing, indicating  
1039 strength in high-level cues but difficulty with subject placement. Likewise, 3D movements outperform  
1040 2D, and natural lighting is handled better than artificial, reflecting challenges in rendering realistic  
1041 synthetic illumination.

Table 6: **Events Taxonomy** Control Nodes and Values

| Name                      | Description  | Potential Values   |
|---------------------------|--|--|
| Standalone Actions        | If the action is stand-alone   | <i>Open Set</i>  |
| Interactive Actions       | If the action involves subject-subject or object-subject interaction     | <i>Open Set</i>  |
| Temporal (Actions)        | How actions unfold across time.  | Atomic, Concurrent, Sequential, Causal, Overlapping, Cyclic, Reverse |
| Foreground (Actions)      | Describes if the action is taking place in the foreground                | Local, Global, Focal   |
| Background (Actions)      | Describes if the is taking place in the background                       | -  |
| Uncertainty               | The probabilistic nature of the action outcome                           | Probabilistic, Deterministic, Mixed                                  |
| Implicit Emotions         | Emotions that are suggested or implied rather than directly stated.      | <i>Open Set</i>  |
| Explicit Emotions         | Emotions that are clearly and directly shown or stated within the scene. | <i>Open Set</i>  |
| Temporal (Emotions)       | How emotions evolve across time.   | Atomic, Concurrent, Sequential, Overlapping, Causal                  |
| Foreground (Emotions)     | Describes if the emotion is taking place in the foreground               | Local, Global, Focal   |
| Background (Emotions)     | Describes if the emotion is taking place in the background               | -  |
| Type of Dialogue Delivery | How the dialogue is delivered  | Dash, Ellipsis, Monologue  |
| Foreground (Emotions)     | Describes if the dialogue is being spoken in the foreground              | Local, Global, Focal   |
| Background (Emotions)     | Describes if the the dialogue is being spoken the background             | -  |
| Change in Environment     | Change of environment or occurrences within a shot                       | <i>Open Set</i>  |
| Story Structure           | Key narrative elements that shape the scene's progression.               | Turning Point, Climax, Foreshadowing, Conflict                       |
| Pace                      | How fast the events are happening in a shot                              | Slow, Fast   |
| Regularity                | How regularly the events are happening in a shot                         | Regular, Irregular   |

Figure 14: **Model performance on Dialogues.** Compared to Actions and Emotions, models struggle at Dialogues. Within Dialogues, performance drop is seen during multi-turn conversations

1042 We next analyze performance at the value level. In Lighting Source (Figure 18), Sunlight, Strobes,  
1043 and Firelight are handled more reliably, while HMI, Fluorescent, and Tungsten lighting show  
1044 comparatively lower performance. At Camera Angle (Figure 19), Aerial and Knee level angles  
1045 are depicted better, while the Dutch and Shoulder-level angles show lower performance. For Shot  
1046 Size (Figure 20), Medium-Wide, Master, and Establishing shots perform better, while Full and  
1047 Extreme Close-Up shots are less consistent.

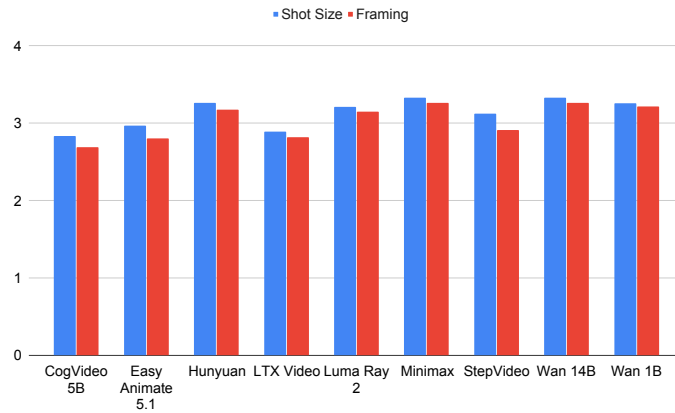


Figure 15: Within the Camera Creative Intent sub-category, models are consistently better at depicting the correct shot size in comparison to camera framing.

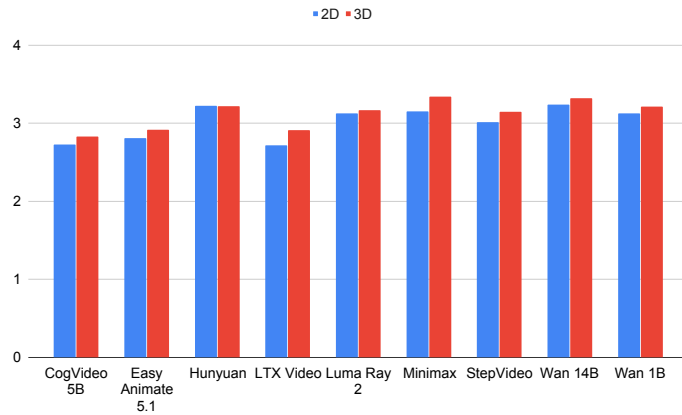


Figure 16: Within the Camera Movement sub-category, models are consistently better at 3D camera movements ("Dolly Zoom") in comparison to 2D camera movements ("Tilt Up").

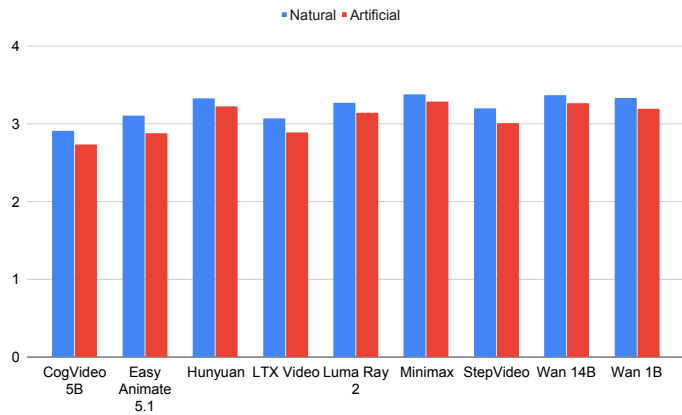


Figure 17: Within the Lighting Position sub-category, models fare better at Natural Sources of Light in comparison to Artificial Sources.

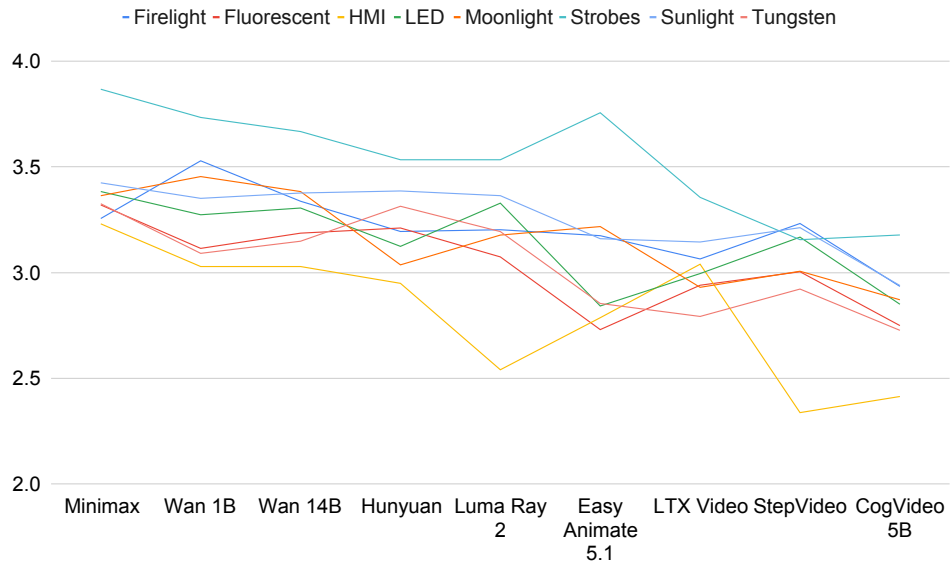


Figure 18: **Model performances across Lighting Source.** Strokes and Sunlight emerge strongest, whereas HMI and Fluorescent are points of weaknesses.

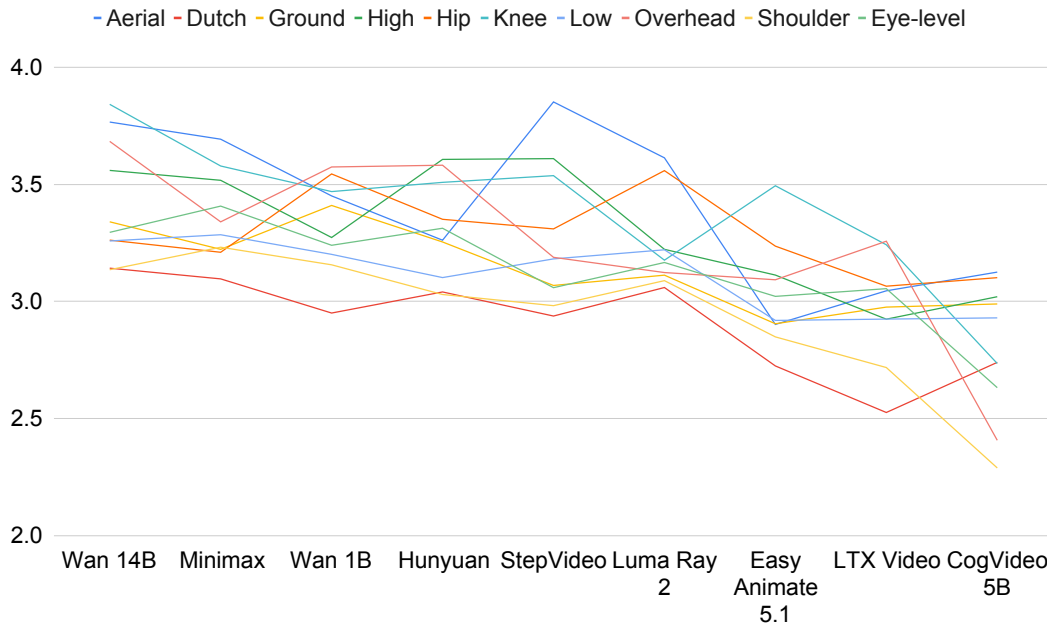


Figure 19: **Model performances across Camera Angles.** The Dutch angle poses a common challenge to all current video generative models

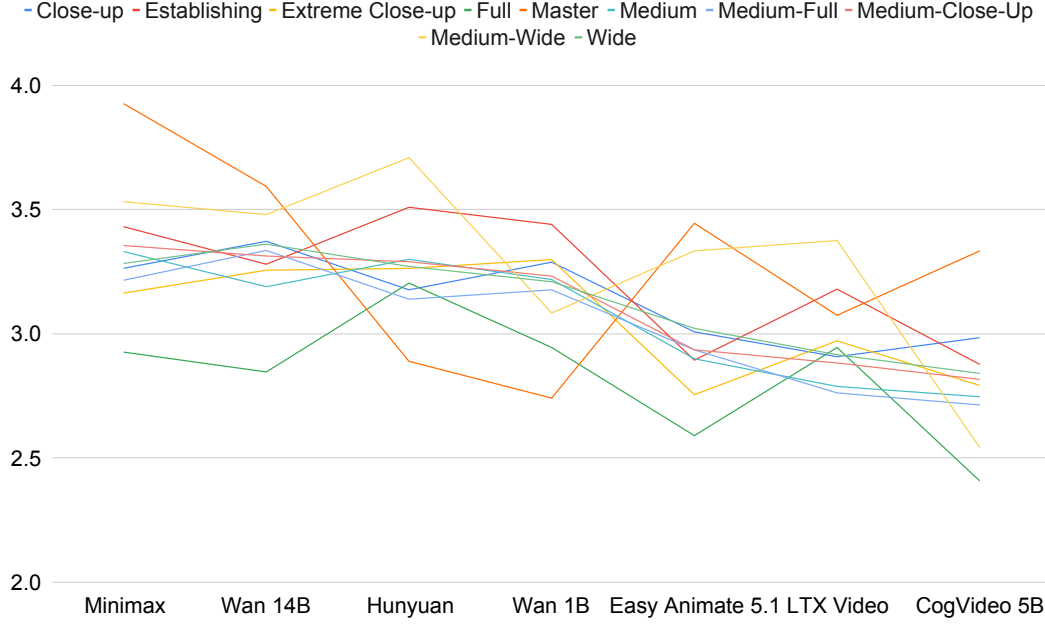


Figure 20: **Model performances across Camera Shot Size.** Models perform well on full shots and struggle at medium-wide and extreme-close-up shots.

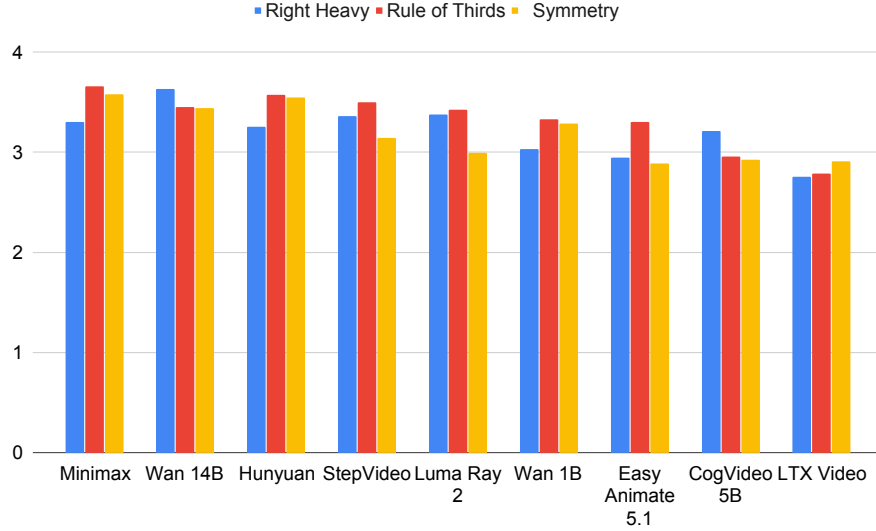


Figure 21: Between different frame compositions, models are better at Rule of Thirds but struggle at maintaining Symmetry.

### 1048 A.2.3 Setup

1049 For the Setup taxonomy, we also analyze performance at the value level. In Balance (Figure 21),  
 1050 models handle rule of thirds framing more effectively but struggle with symmetrical compositions.  
 1051 For Time of Day (Figure 22), among 11 categories, Sunrise and Morning are portrayed well, while  
 1052 Afternoon remains challenging. In Environmental Elements (Figure 23), Snow and Fog are handled  
 1053 better than Rain and Dust.



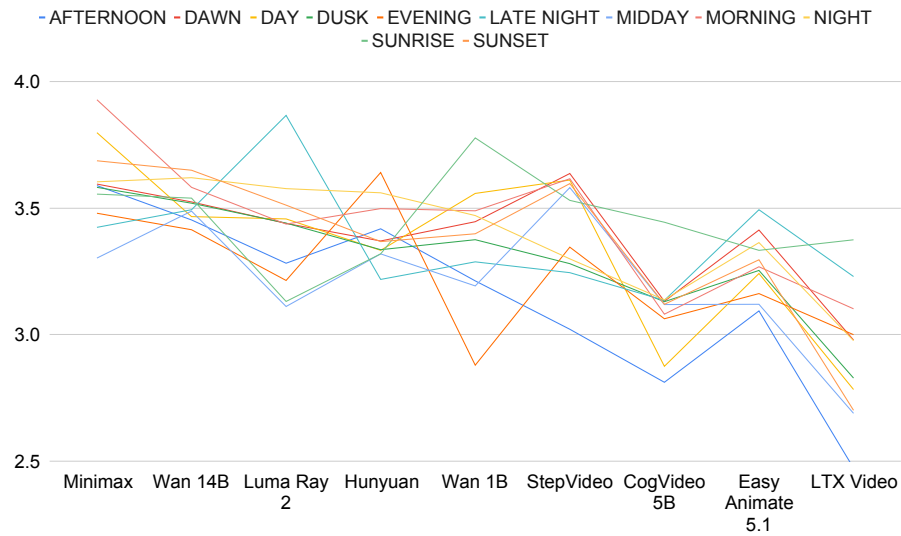


Figure 22: Across Time of Day setups, Sunrise shots are handled better, while Afternoon remains more challenging for models.

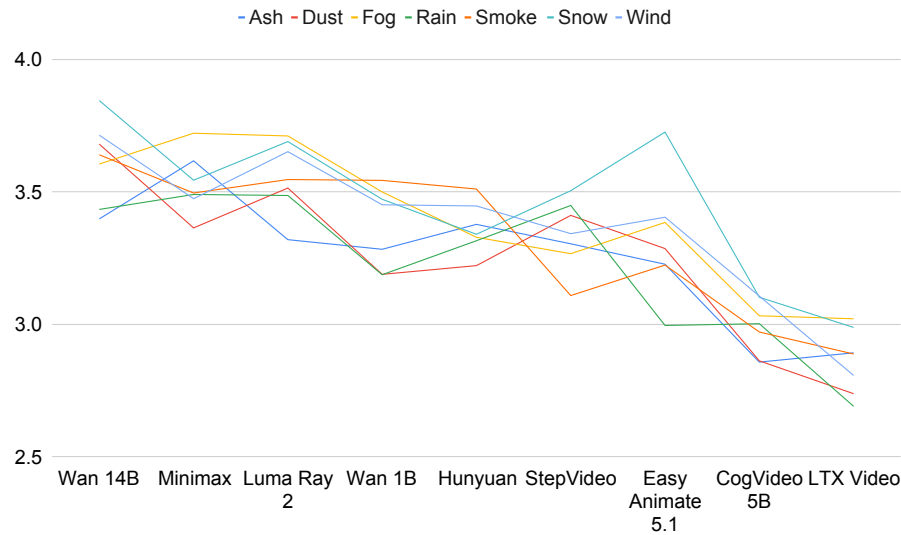


Figure 23: For Environmental Elements, models handle Snow and Fog reasonably well, but struggle with Rain.

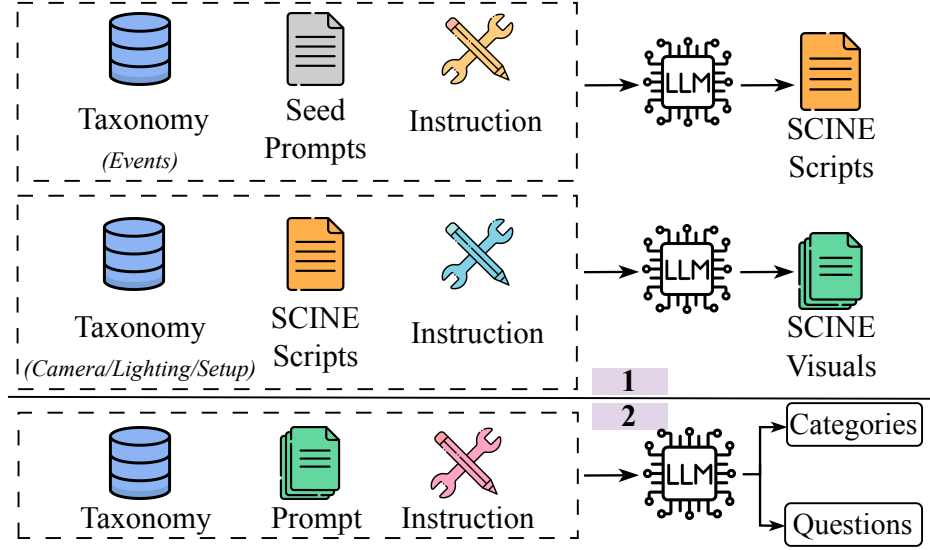


Figure 24: **1. Prompt Generation Pipeline** SCINE-Scripts are created by passing seed prompts and sampled Events taxonomy nodes to an LLM, forming the narrative component of our benchmark. SCINE-Visuals are then generated through structured upsampling, where nodes from the Camera, Lighting, and Setup taxonomies are sampled and injected into each SCINE-Script to create prompts that capture visual exposition. **2. Automatic Categorization and Question Generation** Given a SCINE prompt and taxonomy, we *categorize* each taxonomy element present in the prompt and generate a corresponding *question* to enable isolated evaluation of each control node.

Table 7: A working example of a prompt with its corresponding categories and questions. Each question targets a single control node from the taxonomy, enabling human annotators to perform fine-grained, independent evaluations per node.

| Prompt   | Final Category                                   | Question  |
|--|--|---|
| In a stark white laboratory illuminated by cool LEDs casting clinical precision, a scientist carefully drops a single blue chemical into a beaker, the camera framing an intimate close-up as soft depth of field blurs the sterile environment behind. A back light carves a subtle halo around the glassware moments before the liquid erupts into bright green, intensified by a strategic neon-tinted color gel that makes the reaction glow like bottled lightning. | Camera → Creative Intent → Shot Size             | Does the generated video clearly exhibit a well-executed close-up shot that captures the subject with the intended intimacy and detail?         |
|  | Camera → Intrinsic → Depth of Field              | Does the video effectively showcase a soft depth of field that isolates the subject while smoothly blurring the background?                     |
|  | Lighting → Sources → Artificial/Practicals Light | Is the effect of artificial LED source clearly visible and does it emulate the clinical, cool lighting effect as described in the scene?        |
|  | Lighting → Color Temperature                     | Does the video convey a cool color temperature in its lighting setup that reinforces the clinical precision suggested in the prompt?            |
|  | Lighting → Lighting Position                     | Is a back lighting effect evident in the video, such that it effectively carves a halo or outline around the subject as described?              |
|  | Lighting → Advanced Controls → Color Gels        | Does the video incorporate a neon-tinted color gel effect that intensifies the lighting during the chemical reaction as detailed in the prompt? |

### 1054 A.3 Details on Prompt Generation

1055 Figure 24 illustrates the overall prompt generation pipeline, followed by automatic categorization  
 1056 and question generation. Our process mirrors professional workflows—progressing from scripts  
 1057 to visuals—while constraining prompts to taxonomy-defined control nodes. This enables scalable,  
 1058 fine-grained evaluation by human annotators. During the development of SCINE-Visuals, prompts  
 1059 that exceed the text encoder token limits are filtered out from the benchmark. Table 7 presents an  
 1060 example of a prompt and the corresponding categories and questions generated from it.

1061 Below, we show an example of the instruction given to an LLM to upsample a SCINE-Script into  
1062 SCINE-Visuals by incorporating control nodes from the Camera and Lighting taxonomies.

### SCINE Scripts - Cinematographer

#### System Prompt :

You are a world-class cinematographer known for your visionary storytelling, mastery of light, and camera. You have decades of experience working on award-winning films across genres, collaborating with top directors and production teams. Your insights blend technical expertise with artistic sensibility. When describing scenes or advising on visual storytelling, you use cinematic terminology with clarity and inspiration. Think like Roger Deakins, Emmanuel Lubezki, and Greig Fraser—your visual choices always elevate the emotional tone and narrative arc of a project.

#### User Prompt :

#### GOAL

You will be given a prompt and 2 taxonomies that define camera and lighting controls commonly used by cinematic professionals. Your objective is to enrich the given prompt by sampling relevant nodes from both the taxonomies. As a cinematographer, your role is to "shoot" this scene using the best possible cinematic expression, utilizing the camera and lighting control options provided in the taxonomy.

PROMPT: {prompt}

#### MOST IMPORTANT INFORMATION

1. Only Use Nodes from the Provided Taxonomies : - You must never introduce nodes that are not present in the given taxonomies. - While the values within each node can be flexible—allowing for creativity and imagination grounded in your professional experience. For example, the node "Color Gel" is defined, but has no values. It is upto you to define these values. - The structure must strictly adhere to the nodes defined in the taxonomy. Think expansively within the bounds of each node, but never go beyond them.
2. Preserve the Original Prompt Content : - Do NOT remove or add any of the original content from the input prompt. - Your only task is to enrich the prompt by layering in camera and lighting related information. The core semantics and narrative of the prompt must remain entirely intact.
3. Do NOT include the path through which you sample the nodes in the prompt. That is, do NOT add the paths from the taxonomy using '->'.

**GUIDELINES**

1. Input Prompt - The input prompt describes a single continuous event, intended to occur within one uninterrupted shot. Therefore, do not include any cuts or multiple camera setups. Assume this is a one-shot sequence.
2. Each node in the taxonomies contains: - Description: A definition of what the node represents.  
- Example: An example of how the node may appear in a prompt.  
- Values: A non-exhaustive list of possible values for the node. Some notation: a. OPEN SET - Indicates the node supports a wide range of possible values.  
b. [] - Indicates the node may have multiple values, which are not predefined and should be selected based on your reasoning and cinematic knowledge.
3. Enriched Prompt - Your enriched version will serve as input to a text-to-video model. It must be fluent, natural, and interpretable by the model, while incorporating cinematic elements effectively.

**CAMERA TAXONOMY** The Camera Taxonomy defines elements related to the camera's intrinsics, extrinsics, and its cinematic use : {camera\_taxonomy}

**LIGHTING TAXONOMY** The Lighting taxonomy broadly defines all elements of lighting, including source, position of lighting, along with its effects such as shadows and reflections, along with color temperature, lighting motion such as flickering etc : {lighting\_taxonomy}

When incorporating lighting into your enriched prompt, remember that a cinematographer can shape the look and feel of a shot by selectively illuminating different depth planes of the scene. Lighting can be applied to the foreground, mid-ground, background, and

1063

the subject itself—either individually or in combination. Your choices should support the emotional tone, visual focus, and narrative intent of the shot.

1064

1065 Below, we show an example of the instruction given to an LLM to categorize and generate evaluation  
1066 questions for an input prompt using the Camera taxonomy.

### Camera Categorization and Question Generation

#### GOAL

You are an expert prompt evaluator. Your task is to analyze a video generation prompt and categorize it based on a predefined taxonomy.

**PROMPT:** {prompt}

Available Categories (with Examples)

The category presented to you is that of Camera. The Camera taxonomy broadly defines everything related to the camera - the intrinsics, the extrinsics and the cinematic use of camera. {camera\_taxonomy}

#### Notes about the Taxonomy

Each node in the taxonomy contains :

1. Description : Definition of what that node represents.
2. Example : An example of the presence of a node in the form of a prompt.
3. Values : A non-exhaustive list of values of these nodes. Values are a list of values that this node can have. Some nomenclature :
  - a. OPEN SET indicates that this node contains a large number of values.
  - b. [] indicates that this node may have multiple values, but are not defined explicitly and it is upto your reasoning and knowledge.

#### Examples of Categorization

1. Static Medium-Close-Up of David's face showing quiet devastation. Quick Push In as tears well up in his eyes. Shot with a medium ISO to capture the dim apartment lighting.

Static - Camera -> Trajectory -> Camera Movement -> Static

Push In - Camera -> Trajectory -> Camera Movement -> 3D

Medium-Close up - Camera -> Creative Intent -> Shot Size

Medium ISO - Camera -> Intrinsics -> Exposure -> ISO

2. Wide shot of a bustling city street at night. The neon lights of the shops and restaurants cast a colorful glow on the wet pavement. People walk by, their faces illuminated by the bright signs. The camera pans up to reveal the towering skyscrapers that loom overhead, their windows reflecting the city lights.

Wide shot - Camera -> Creative Intent -> Shot Size

Pans Up - Camera -> Trajectory -> Camera Movement -> 2D

#### TASK

Analyze the given prompt and return the following structured output in a valid JSON format:

Words: Extract important keywords or key phrases from the prompt using the following guidance:

- Identify named entities related to a camera in professional use as you would in NER (Named Entity Recognition).
- Extract noun phrases or descriptive terms that relate to a camera.
- Prefer multi-word expressions where meaningful related to a camera.
- Avoid generic or uninformative words like "a", "video", "the", etc.

Categories: For each word or phrase, assign the most appropriate category from the taxonomy. A dictionary of relevant categories from the taxonomy.

1067

Table 8: Lexical Diversity of SCINE Scripts. Compared to existing prompt-based benchmarks, SCINE-Scripts demonstrate higher lexical diversity across multiple metrics.

| Benchmark          | TTR $\uparrow$ | Distinct bi-grams $\uparrow$ | Jaccard Distance $\uparrow$ |
|--------------------|----------------|------------------------------|-----------------------------|
| VBench [63]        | 0.1489         | 0.4605                       | 0.9384                      |
| MovieGenBench [65] | 0.1660         | 0.5311                       | 0.9285                      |
| EvalCrafter [64]   | <b>0.2270</b>  | <u>0.6038</u>                | <u>0.9413</u>               |
| T2V-CompBench [66] | 0.1435         | 0.4781                       | 0.9350                      |
| SCINE Scripts      | <u>0.1760</u>  | <b>0.6177</b>                | <b>0.9445</b>               |

- For each relevant category, assign a score between '0' and '1' representing how strongly the prompt matches the category.
- Provide a reason for each score, referring to the words or phrases extracted and how they relate to the category.
- Generate a question that helps a human evaluator determine whether this category is visually present in the generated video. Use your reasoning to guide the question. The evaluator will use this question to rate the video on a scale from 1 (not at all) to 5 (strongly represented).
- The generated question should evaluate quality, consistency and presence of the node in the video.

**Important Guidelines:**

- The camera information should be explicitly mentioned in the prompt. Do NOT imply, assume or derive anything. Only consider a word or a phrase a match, if it is explicitly mentioned in the prompt.
- Each prompt can have multiple nodes of the Camera taxonomy. You should capture all of the nodes in the prompt and map it back to the taxonomy.
- You must always traverse from the root node, which is Camera in this case. That is, the 'category' should always start as (Camera -> ..)
- You will never create a node that is not in the taxonomy. These nodes can have multiple values, as previously explained and you are expected to be imaginative about the values. But the nodes, should always come from the given taxonomy.
- Since the taxonomy is of Camera, we do not care about objects, subjects, lighting, events, actions or emotions. Your sole focus should be about camera terms that are present in the prompt in accordance with the taxonomy. You will NOT ask any question related to objects, subjects, lighting, events, actions or emotions.

1068

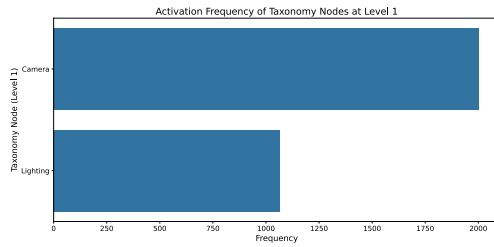
1069 Table 8 compares SCINE-Scripts with existing prompt-based video generation benchmarks. We  
 1070 compute token level metrics: Type-Token Ratio (TTR), Distinct bi-grams, and average pairwise  
 1071 Jaccard Distance, and find that SCINE-Scripts exhibits strong lexical diversity.

#### 1072 A.4 Distribution of Taxonomy Categories in SCINE Prompts

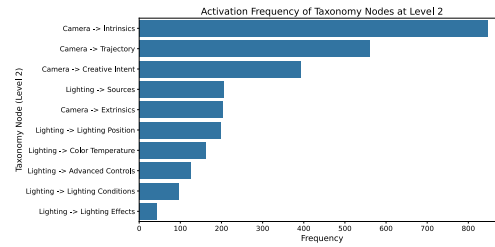
1073 Figures 25, 26, and 27 show the distribution of activated nodes in SCINE Visuals, aggregated at the  
 1074 node level, across the roles of Cinematographer, Production Designer, and Director, respectively. As  
 1075 shown, our prompts cover a broad distribution across all taxonomies.

#### 1076 A.5 Annotation Details

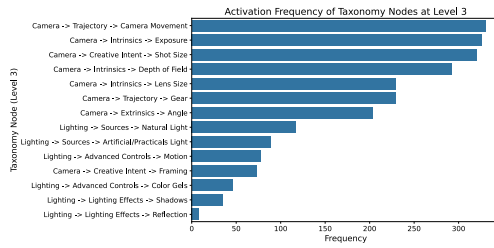
1077 Figure 28 shows the annotation interface used by human annotators during evaluation. We also  
 1078 present the distribution of annotators' years of experience in film production in Figure 29. While  
 1079 annotations for cinematic controls can be subjective, especially given the large number of control  
 1080 nodes, we try our best to mitigate this by providing clear rating guidelines to annotators for each  
 1081 control node. Table 9 presents a minimal example of the rating guidelines shared with the annotators.



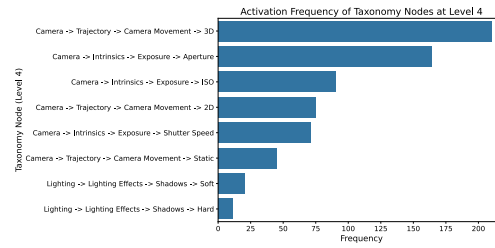
(a) Level 1 Activations



(b) Level 2 Activations

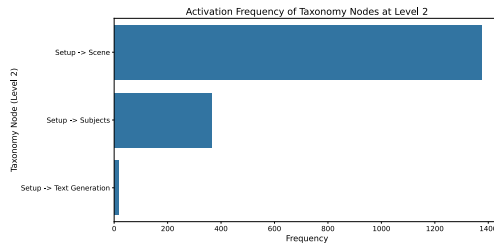


(c) Level 3 Activations

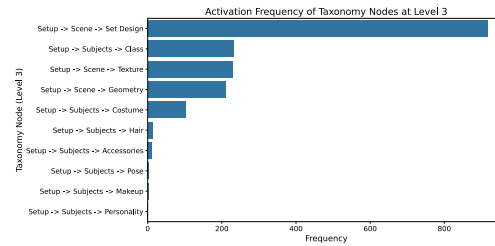


(d) Level 4 Activations

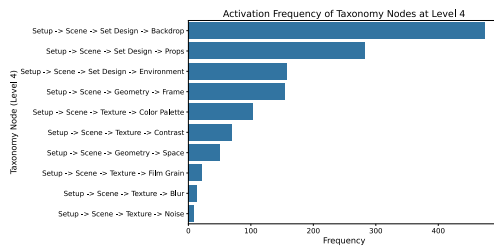
Figure 25: Node activations in Camera and Lighting taxonomies for the Cinematographer role in SCINE Visuals.



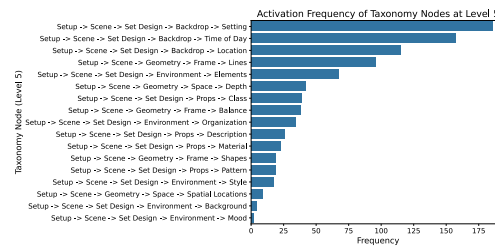
(a) Level 2 Activations



(b) Level 3 Activations



(c) Level 4 Activations



(d) Level 5 Activations

Figure 26: Node activations in Setup taxonomy for the Production Designer role in SCINE Visuals.

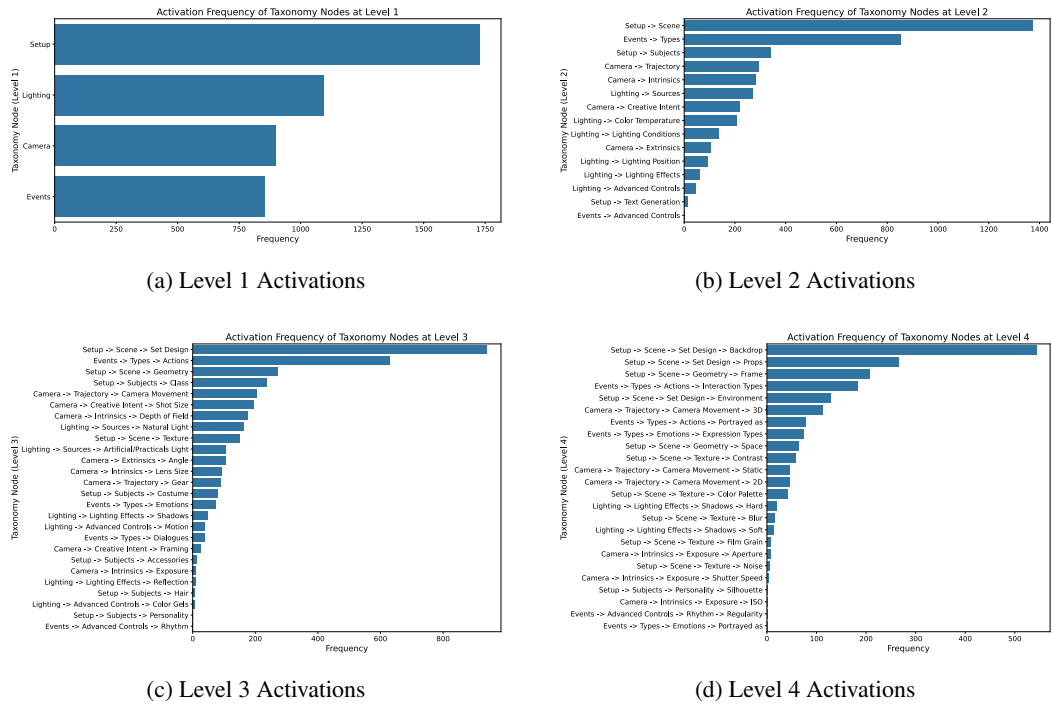


Figure 27: Node activations in All taxonomies for the Director role in SCINE Visuals.

**Video A**

**Video B**

**Prompt**

On a city rooftop at night, music pulses as people dance under string lights. Other guests stand in small groups near the railing, talking and admiring the glittering city view, while a bartender mixes drinks behind a portable bar.

**Events -> Types -> Change in Environment**

Does the video clearly depict pulsating music as a dynamic environmental element that contributes to the overall atmosphere?

1. Score A (select 1) \*

1  2  3  4  5

1. Score B (select 1) \*

1  2  3  4  5

**Events -> Types -> Actions**

Does the video effectively show people dancing with lively, coordinated movements that capture the intended energy of the scene?

1. Score A (select 1) \*

1  2  3  4  5

1. Score B (select 1) \*

1  2  3  4  5

Figure 28: User Interface used by annotators to perform evaluations.

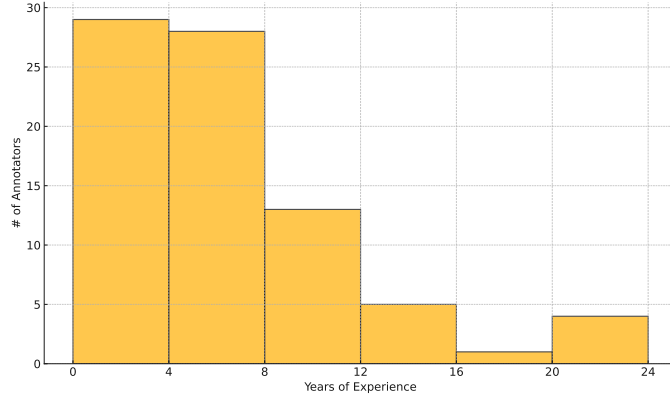


Figure 29: Distribution of years of film production experience among human annotators in our evaluation.

Table 9: Examples of rating guidelines provided to human annotators for different control nodes, across all taxonomies

| Dimension                 | Score | What to Look For  |
|---------------------------|-------|---|
| Low Angle                 | 1     | Camera is at or above eye level, not low angle at all.  |
|                           | 2     | Slight upward tilt, but still feels neutral.  |
|                           | 3     | Below subject, mild upward view, light impact.  |
|                           | 4     | Clear low angle, subject looks larger or imposing.  |
|                           | 5     | Strong low angle, subject dominates, towering presence.   |
| Overlapping Actions       | 1     | No action or one action is present.   |
|                           | 2     | Actions are isolated or unrelated.  |
|                           | 3     | Timing is off, they start or end awkwardly.   |
|                           | 4     | Some overlap, but hard to follow.   |
|                           | 5     | Fluid overlap, actions feel natural and dynamic together.   |
| Back Light Position       | 1     | Light clearly comes from front or side, no rim light or background separation.  |
|                           | 2     | Some edge lighting, but not consistent or strong, subject may still blend into background.                                    |
|                           | 3     | Back light is partially visible, outline is hinted but not clear on full subject.   |
|                           | 4     | Back light is clearly present, rim light separates subject from background.   |
|                           | 5     | Strong back light effect, glowing edges around hair or shoulders. Subject clearly pops against the background. Perfect match. |
| Symmetrical Frame Balance | 1     | Composition is clearly asymmetrical.  |
|                           | 2     | Some repeating elements, but no visual mirror.  |
|                           | 3     | Partial symmetry or mirrored clutter that’s not clean.  |
|                           | 4     | Almost perfect symmetry, small inconsistencies exist.   |
|                           | 5     | Clear and precise symmetry, mirrored subjects, reflections, or centered framing. Strong and intentional.                      |

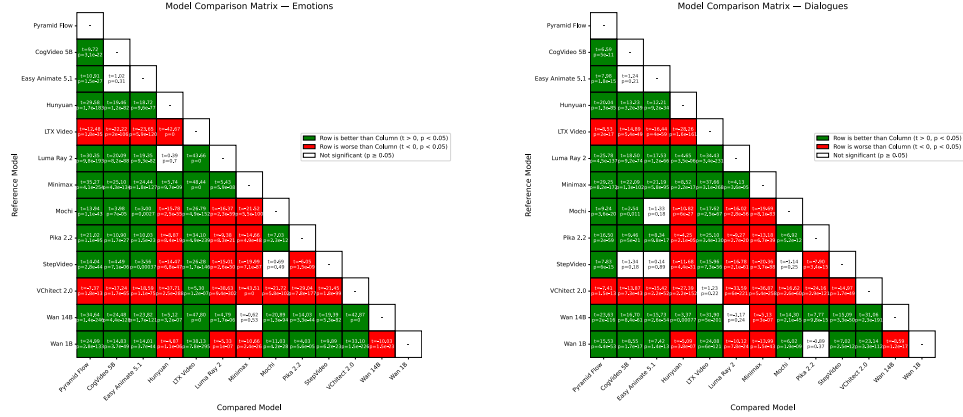
## 1082 A.6 Statistical tests

1083 Pairwise t-tests show that the vast majority of model comparisons in our human evaluation, across all  
 1084 taxonomies are statistically significant at the 5% level ( $p < 0.05$ ). Figure (30 - 32) presents the t-test  
 1085 results of Events, Lighting and Camera, and Setup, respectively.

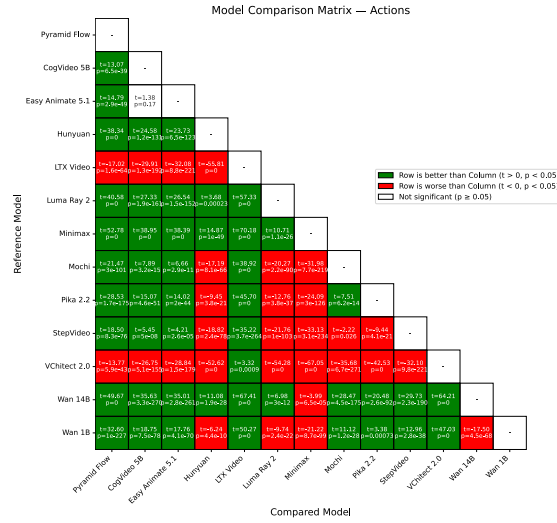
## 1086 A.7 Additional VLM results

1087 **Comparison with Closed Source Models** We extend our validation to closed-source, flagship SOTA  
 1088 models. Specifically, we evaluate two recent models from the Gemini family with distinct purposes:





(a) Pair-wise t-test matrix of model comparison on emotions (b) Pair-wise t-test matrix of model comparison on dialogues

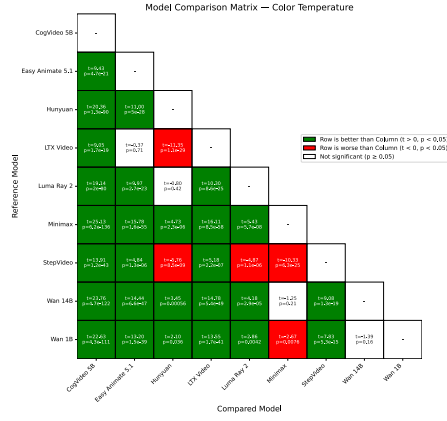


(c) Pair-wise t-test matrix of model comparison on actions

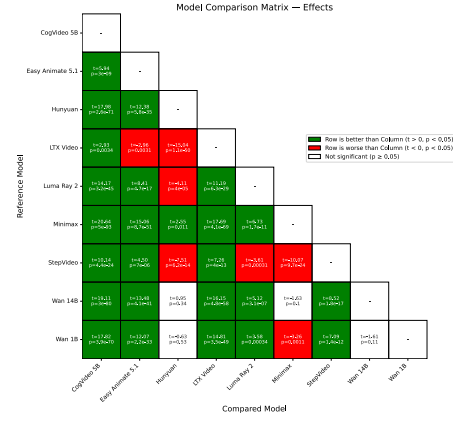
Figure 30: Statistical comparison matrices for Events: Emotions, Dialogue, and Actions.

1089 Gemini-2.0-Flash, optimized for fast inference, and Gemini-2.5-Pro-Preview-05-06, optimized for  
 1090 complex reasoning. We use the same human-aligned preference accuracy metric as with open-  
 1091 source models. Due to the lack of public details on model sizes, we cannot draw conclusions about  
 1092 scaling effects. However, Gemini-2.5-Pro consistently outperforms open-source models, including  
 1093 QwenVL-2.5-72B, across all categories. Notably, as shown in Figure [33](#), our 7B model outperforms  
 1094 Gemini-Flash across all categories and performs competitively with Gemini-2.5-Pro. This highlights  
 1095 the strength and scalability of our approach for professional video evaluation.

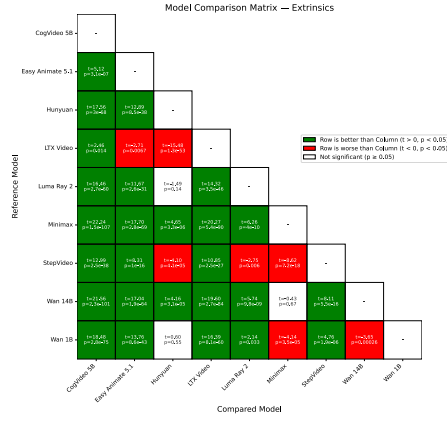
1096 **Reliability in VLMs** A reliable VLM-as-Judge should produce consistent scores when given the  
 1097 same video, prompt, and focus aspect. In this analysis, we evaluate the raw scores generated by  
 1098 VLMs rather than preference rankings, and measure their stability under Best-of-5 sampling. Since  
 1099 VLMs are probabilistic, we evaluate reliability via the standard deviation of scores across runs. We  
 1100 use temperature=0 to sample to make ensure that the highest probability is selected at each sampling  
 1101 step. We exclude our model from this analysis, as its architecture includes a dedicated value head,  
 1102 unlike zero-shot VLMs that produce rewards as text. Our results show that Qwen-2.5VL-3B exhibits  
 1103 a high variance, making it unreliable under repeated sampling. In contrast, the flagship models and  
 1104 the strongest open-source model, QwenVL-2.5-72B, demonstrate high reliability, with consistently  
 1105 low variance (Table [10](#)).



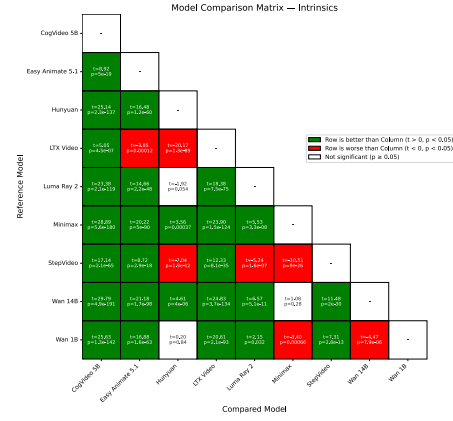
(a) Pair-wise t-test matrix of model comparison on Color Temperature



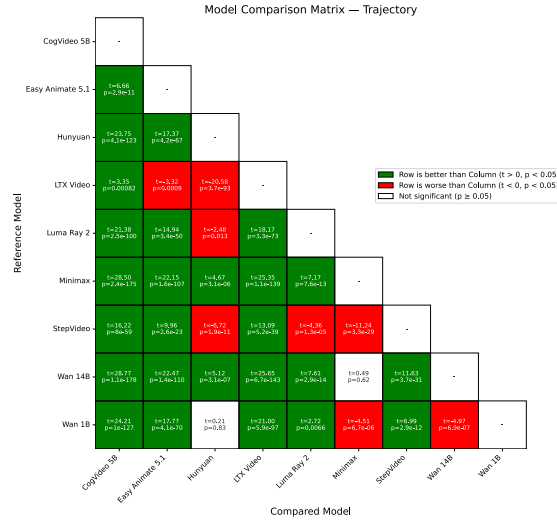
(b) Pair-wise t-test matrix of model comparison on Lighting Effects



(c) Pair-wise t-test matrix of model comparison on Camera Extrinsic

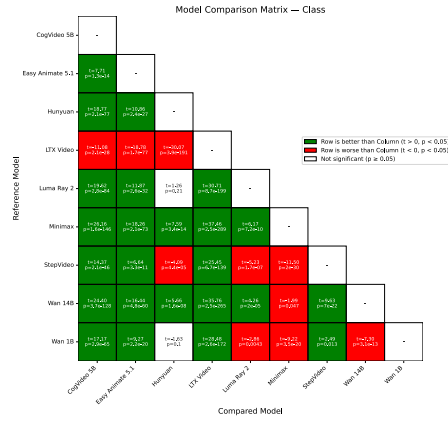


(d) Pair-wise t-test matrix of model comparison on Camera Intrinsic

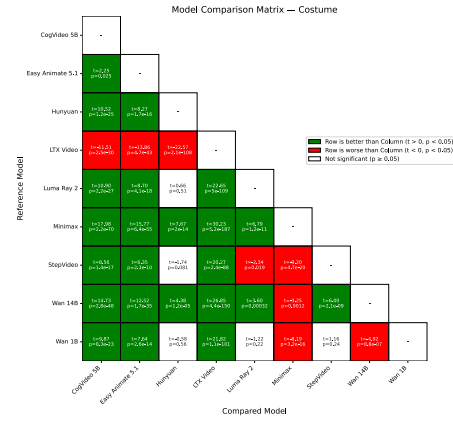


(e) Pair-wise t-test matrix of model comparison on Camera Trajectory

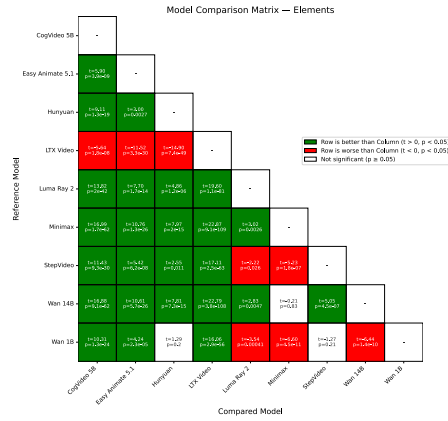
Figure 31: Statistical comparison matrices for Camera and Lighting



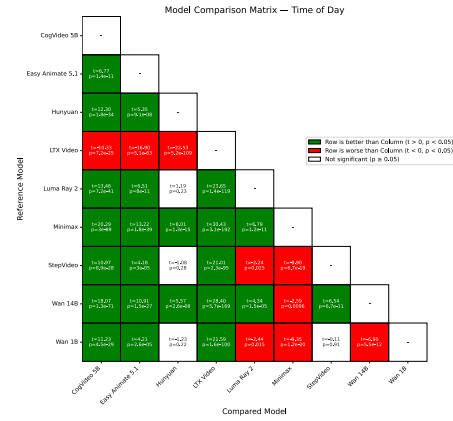
(a) Pair-wise t-test matrix of model comparison on Subject Class



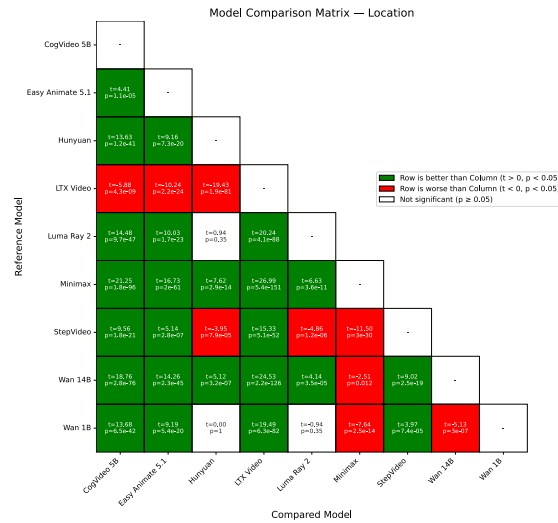
(b) Pair-wise t-test matrix of model comparison on Subject Costume



(c) Pair-wise t-test matrix of model comparison on Elements



(d) Pair-wise t-test matrix of model comparison on Time of Day



(e) Pair-wise t-test matrix of model comparison on Location

Figure 32: Statistical comparison matrices for Setup

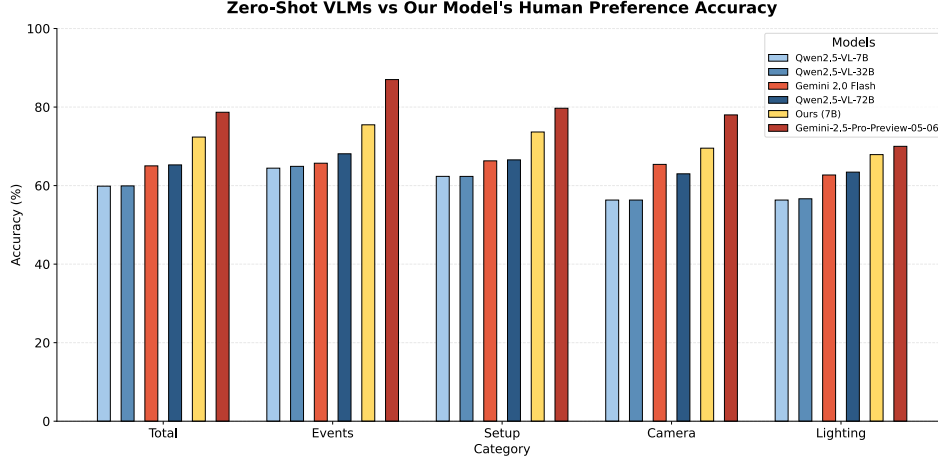


Figure 33: Preference Accuracy of open and closed-sourced VLMs in rating videos generated for Professional Use

Table 10: Measuring VLM Reliability across best-of-5 sampling

| Model                | Standard-Deviation ↓ | Krippendorff-alpha ↑ |
|----------------------|----------------------|----------------------|
| Qwen2.5-VL-3B        | 2.34                 | 0.36                 |
| Qwen2.5-VL-7B        | 0.37                 | 0.84                 |
| Qwen2.5-VL-32B       | 0.47                 | 0.65                 |
| Qwen2.5-VL-72B       | 0.23                 | <b>0.95</b>          |
| Gemini2.5-Flash      | 0.198                | 0.9                  |
| <b>Gemini2.5-Pro</b> | <b>0.136</b>         | <b>0.95</b>          |

1106 **Additional VLM Training Details** Input Videos are preprocessed at 2 FPS and their native resolution.  
 1107 The validation set contains 12,763 samples with unique prompts to test generalization.

## 1108 A.8 Limitations

1109 Although our taxonomy was developed in consultation with domain experts, it is limited by the scope  
 1110 of our collaborator network. Filmmaking terminology and interpretive nuance vary across regions and  
 1111 cultures, greater expert diversity would enable broader incorporation of global cinematic controls into  
 1112 the taxonomy. Some taxonomy nodes (e.g., Color Temperature, ISO) were abstracted for evaluation,  
 1113 as we found it difficult for annotators to consistently perceive fine-grained values (such as 2000K or  
 1114 ISO 800). Prompt generation is based on LLMs, whose proprietary nature and potential biases can  
 1115 influence the language and structure of the prompts. Our zero-shot VLM evaluations were bounded  
 1116 by compute and data resources, limiting the scale and scope of the experiments.

## 1117 A.9 Broader Impact

1118 We hope SCINE encourages the generative AI and computer vision communities to engage more  
 1119 deeply with the elements required to produce a professional cinematic shot. While our taxonomy is  
 1120 currently used for evaluation, it also offers a structured foundation for broader tasks such as captioning,  
 1121 creating control aware video datasets, and guiding model training toward explicit cinematic intent.  
 1122 We also envision that our fine-grained control nodes can drive new directions in open computer  
 1123 vision problems, such as estimating camera intrinsics from video, inferring lighting position, and  
 1124 understanding frame compositionality. In conclusion, we see our taxonomy as a first step toward  
 1125 systematically understanding cinematic controls in generative models, and we are hopeful it will be  
 1126 meaningfully adopted and extended across both generative AI and filmmaking communities.

## 1127 Additional References

- 1128 [63] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang,  
 1129 Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang,  
 1130 Dahua Lin, Yu Qiao, and Ziwei Liu. Vbench: Comprehensive benchmark suite for video  
 1131 generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and  
 1132 Pattern Recognition (CVPR)*, pages 21807–21818, June 2024.
- 1133 [64] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu,  
 1134 Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large  
 1135 video generation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and  
 1136 Pattern Recognition (CVPR)*, pages 22139–22149, June 2024.
- 1137 [65] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv  
 1138 Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, David Yan, Dhruv Choudhary, Dingkan  
 1139 Wang, Geet Sethi, Guan Pang, Haoyu Ma, Ishan Misra, Ji Hou, Jialiang Wang, Kiran Ja-  
 1140 gadeesh, Kunpeng Li, Luxin Zhang, Mannat Singh, Mary Williamson, Matt Le, Matthew Yu,  
 1141 Mitesh Kumar Singh, Peizhao Zhang, Peter Vajda, Quentin Duval, Rohit Girdhar, Roshan  
 1142 Sumbaly, Sai Saketh Rambhatla, Sam Tsai, Samaneh Azadi, Samyak Datta, Sanyuan Chen,  
 1143 Sean Bell, Sharadh Ramaswamy, Shelly Sheynin, Siddharth Bhattacharya, Simran Motwani,  
 1144 Tao Xu, Tianhe Li, Tingbo Hou, Wei-Ning Hsu, Xi Yin, Xiaoliang Dai, Yaniv Taigman, Yaqiao  
 1145 Luo, Yen-Cheng Liu, Yi-Chiao Wu, Yue Zhao, Yuval Kirstain, Zecheng He, Zijian He, Albert  
 1146 Pumarola, Ali Thabet, Artsiom Sanakoyeu, Arun Mallya, Baishan Guo, Boris Araya, Breena  
 1147 Kerr, Carleigh Wood, Ce Liu, Cen Peng, Dmitry Vengertsev, Edgar Schonfeld, Elliot Blan-  
 1148 chard, Felix Juefei-Xu, Fraylie Nord, Jeff Liang, John Hoffman, Jonas Kohler, Kaolin Fire,  
 1149 Karthik Sivakumar, Lawrence Chen, Licheng Yu, Luya Gao, Markos Georgopoulos, Rashel  
 1150 Moritz, Sara K. Sampson, Shikai Li, Simone Parmeggiani, Steve Fine, Tara Fowler, Vladan  
 1151 Petrovic, and Yuming Du. Movie gen: A cast of media foundation models, 2025. URL  
 1152 <https://arxiv.org/abs/2410.13720>.
- 1153 [66] Kaiyue Sun, Kaiyi Huang, Xian Liu, Yue Wu, Zihan Xu, Zhenguo Li, and Xihui Liu. T2v-  
 1154 compbench: A comprehensive benchmark for compositional text-to-video generation, 2025.  
 1155 URL <https://arxiv.org/abs/2407.14505>.