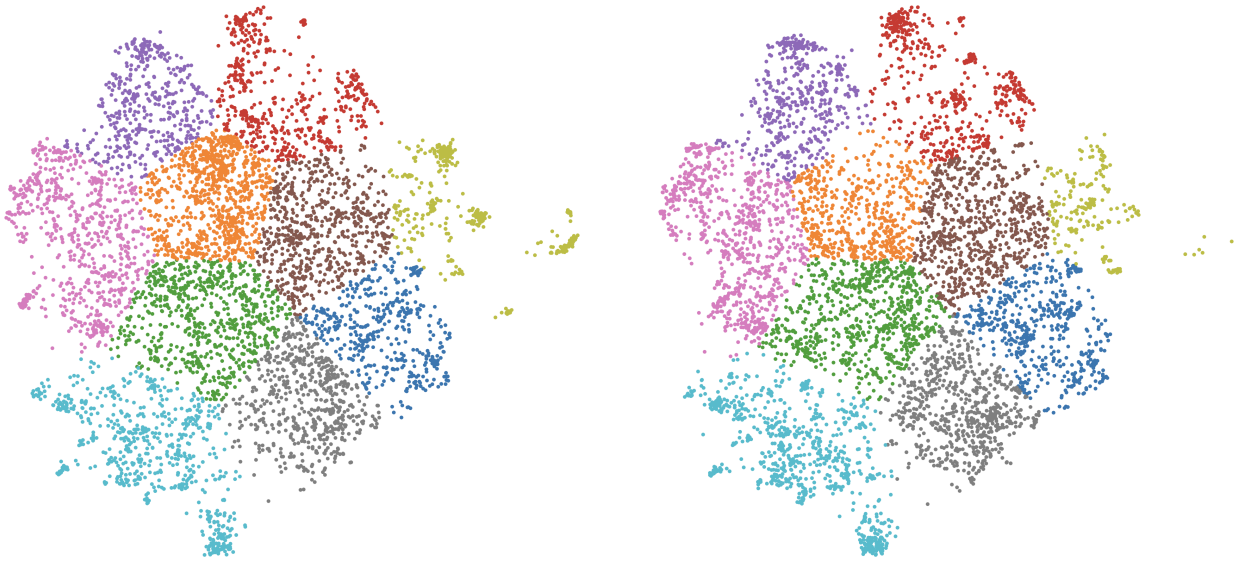
APPENDIX – GeneVA: A Dataset of Human Annotations for Generative Text to Video Artifacts

Jenna Kang Maria Beatriz Silva Patsorn Sangkloy Kenneth Chen Niall L. Williams Qi Sun
New York University
{jennakang, mariasilva, ps5688, kennychen, n.williams, qisun}@nyu.edu

1. Cluster Visualizations



(a) unclean prompts **(b)** clean prompts

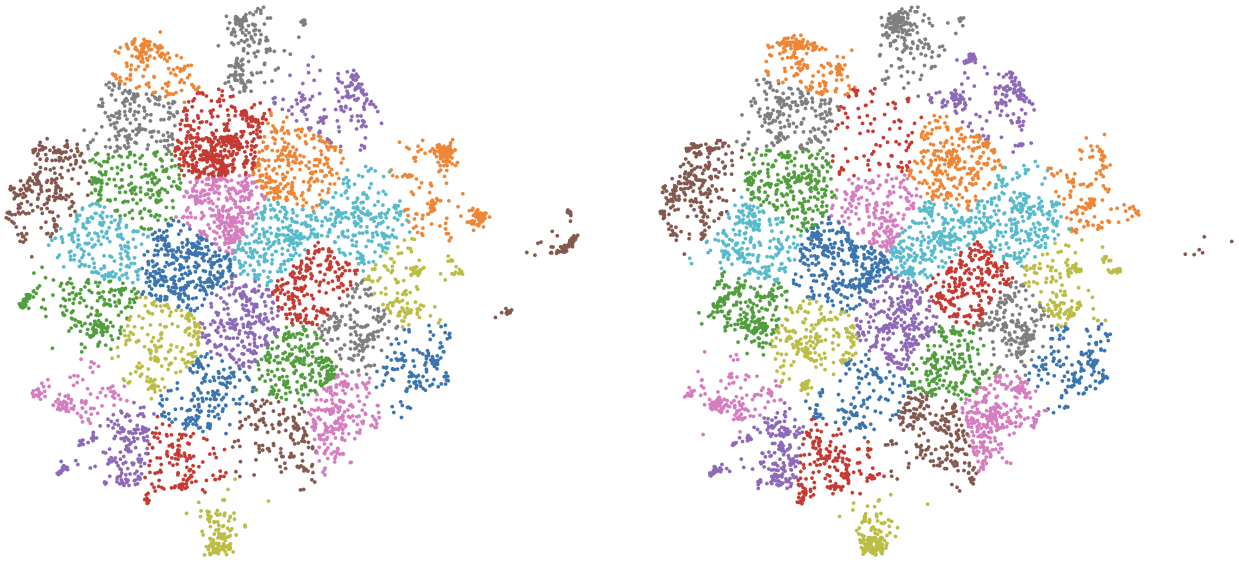
Figure 1. We show the clustered original uncleaned (left) and cleaned (right) prompts for 10 clusters.

Additional visualizations of clustered prompts are shown here, in a similar style to Figure 3a/b in the main manuscript. Our motivation for showing these cluster plots is to qualitatively validate our text prompt cleaning (described in the next section). The reasoning is that, if clusters are roughly equivalent, then the cleaned prompts capture the intent of the original text prompts. Figure 1 shows 10 clusters and Figure 2 shows 30. Qualitatively, we see that the cleaned prompts' embeddings have similar clustering behavior when projected to a low dimension. Even with a large number of clusters (30 clusters in Figure 2), this observation holds.

2. Text Prompt Cleaning

We note that many of the text prompts from VidProM were not in a human-readable format. As such, we passed all these original unclean text prompts through Gemini to display them in a readable form. The instruction prompt passed to Gemini was the following:

```
system_instruction = f"""
    Your task is to summarize a list of user prompts.
```



(a) unclean prompts **(b)** clean prompts

Figure 2. We show the clustered original uncleaned (left) and cleaned (right) prompts for 30 clusters.

```
    Follow these rules strictly for EACH prompt in the list below:
    1.  Identify the core creative idea.
    2.  Remove all technical parameters, such as '--ar 16:9', '--v 5', '-fps 2', 'T252', etc.
    3.  Remove all style, quality, or rendering keywords like 'photorealistic', '4k', '8k',
        'highly detailed', 'unreal engine', 'cinematic', etc.
    4.  Translate any non-English parts of the prompt into simple, clear English.
    5.  Condense the core idea into one or two simple, human-readable sentences.
    6.  Return the summaries for each prompt in the exact same order they were given.
    7.  IMPORTANT: Separate each summary with the exact delimiter '|||'. Do not add any other
        text, numbering, or commentary.

    Here are the prompts to summarize:
    {formatted_prompts}
"""
```

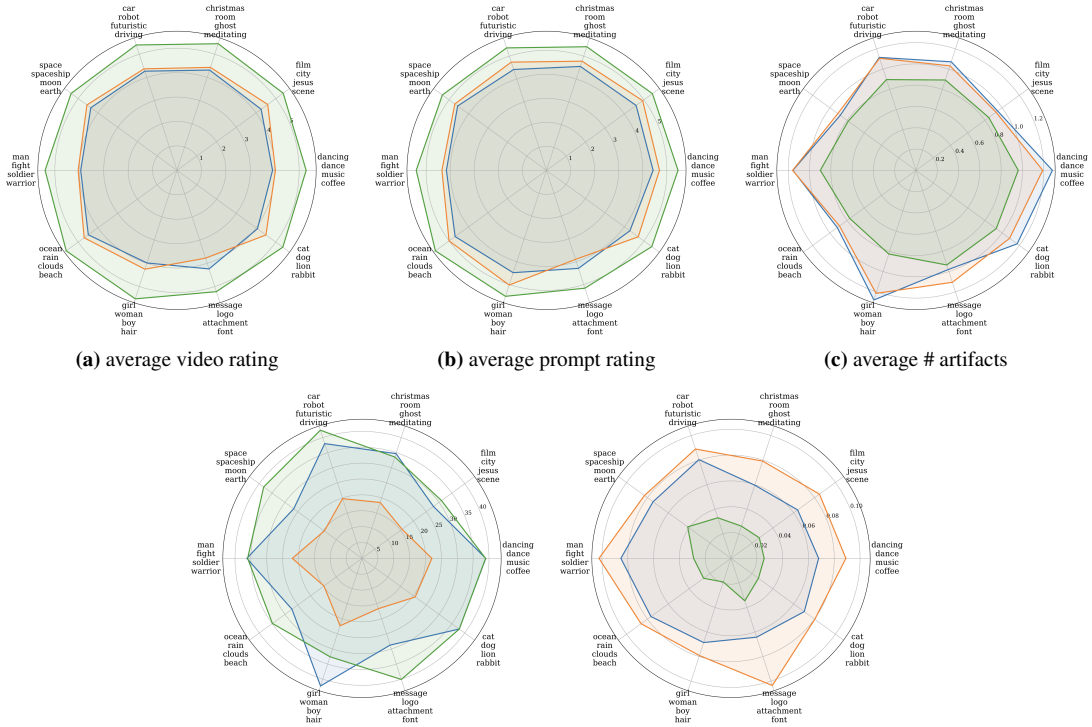
Figure 3a and Figure 3b in the main manuscript show the unclean and clean text prompt embeddings, respectively, projected onto a 2D space; qualitative inspection shows the two align well. In addition to the clustering, we manually confirmed that cleaned text prompts capture the intent of the original prompts.

3. More Radar Plots

In the main manuscript, we showed radar plots across 20 clusters of embedded text prompts for each of the 3 models in our dataset. Here, we plot these for 10 (Figure 3), 15 (Figure 5), 25 (Figure 6), and 30 (Figure 4) clusters. Of note is that similar trends arise across all cluster counts, in a way described in the main manuscript.

4. Additional Artifact Detection Results

We show additional artifact detection results in Figure 7. Predicted artifact descriptions are at the top of each frame, and detected artifacts are shown as a red bounding box.



(a) average video rating **(b)** average prompt rating **(c)** average # artifacts

(d) average # bboxes **(e)** average total bbox coverage

Figure 3. This figure shows results for 10 clusters.

5. Additional Bounding Box Visualizations

We include additional annotated bounding boxes in the same style as Figure 2. The artifacts are separated by model, where Figure 8 shows labeled artifacts for Pika, Figure 9 for VC, and Figure 10 for Sora. Five annotated videos are shown for each, and for unique artifact categories.

6. Additional Details on our prompt collection

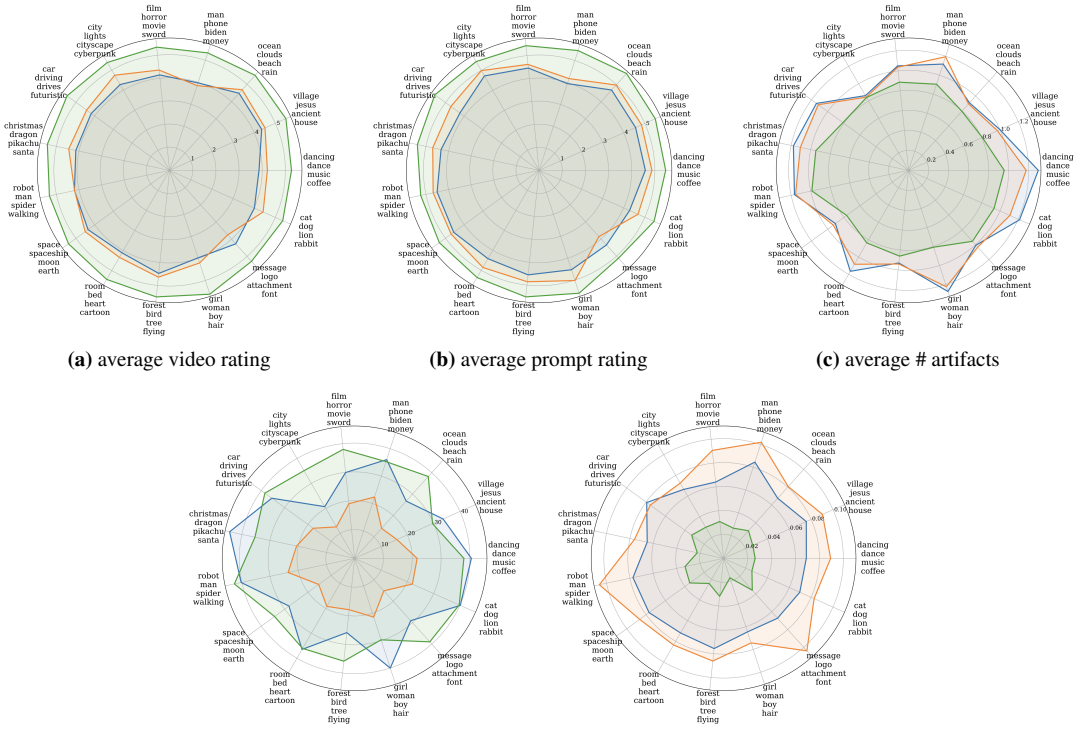
The scale of the VidProM dataset [3] paired with its lack of direct prompt-to-file mappings became a challenge as we sought to extract a specific subset. To address this, we developed a pipeline to retrieve the videos corresponding to the prompts chosen by our kernel herding sampling. This involved downloading all TAR archives to our high-performance computing infrastructure, mapping our herding-selected prompts to their corresponding UUIDs through the dataset's indexing system, and implementing custom extraction scripts to locate and organize target videos. This process was executed separately for Pika [1] and VideoCrafter2 [2] content, with the curated dataset subsequently stored in S3 infrastructure for the human annotation workflow. This approach ensures our dataset captures both the diversity of user intentions through representative prompt sampling and the variety of current generation capabilities through strategic model selection, providing a robust foundation for artifact analysis.

References

[1] Pika art. Accessed: July 10, 2025. 3

[2] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models, 2024. 3

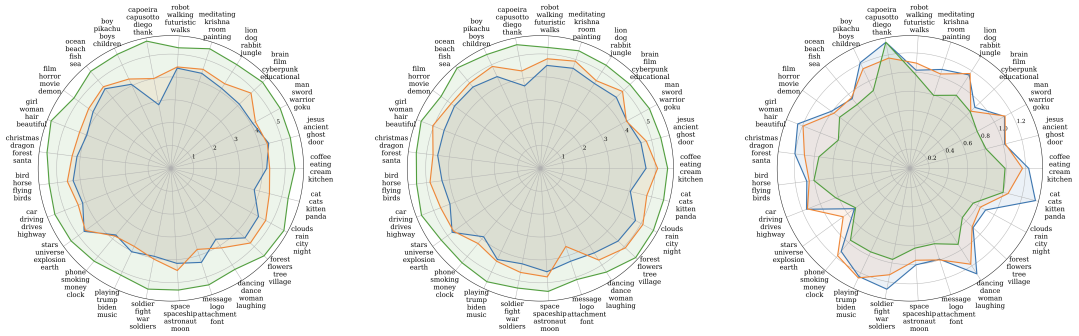
[3] Wenhao Wang and Yi Yang. Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. 2024. 3



(a) average video rating **(b)** average prompt rating **(c)** average # artifacts

(d) average # bboxes **(e)** average total bbox coverage

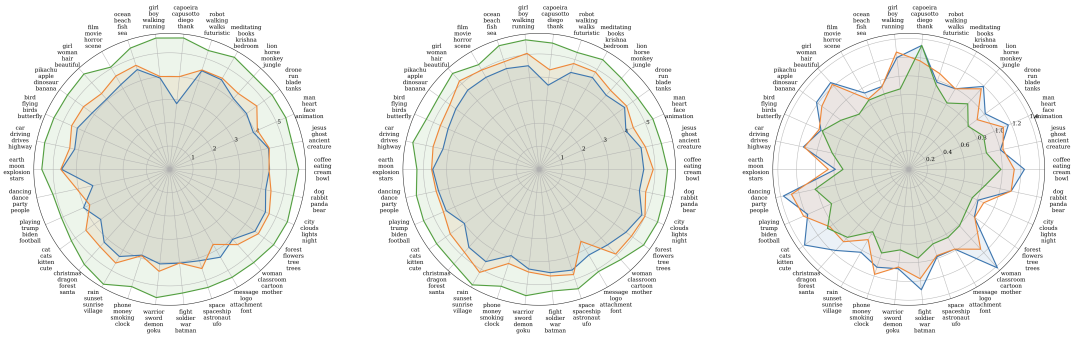
Figure 4. This figure shows results for 15 clusters.



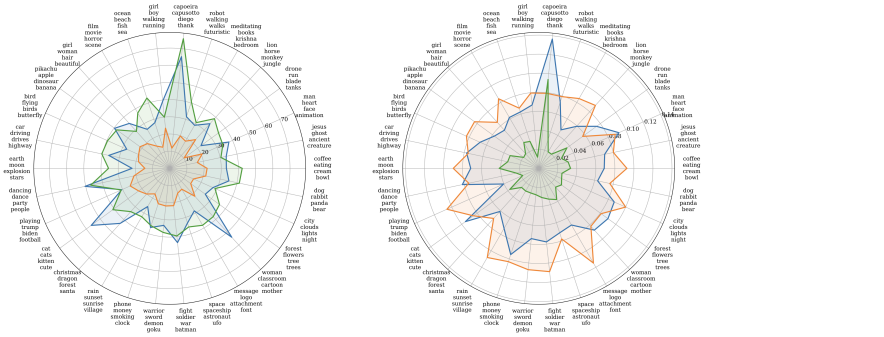
(a) average video rating **(b)** average prompt rating **(c)** average # artifacts

(d) average # bboxes **(e)** average total bbox coverage

Figure 5. This figure shows results for 25 clusters.



(a) average video rating **(b)** average prompt rating **(c)** average # artifacts



(d) average # bboxes **(e)** average total bbox coverage

Figure 6. This figure shows results for 30 clusters.

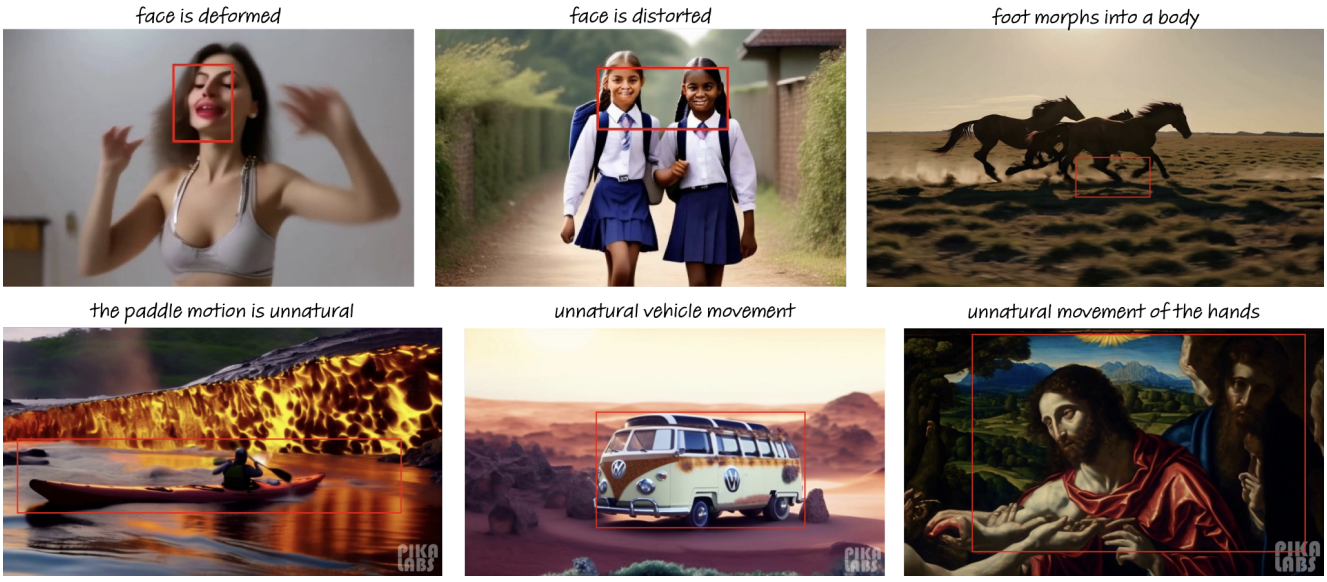


Figure 7. Additional results for the artifact detection model.

"A barbarian warrior with an axe and two heads in his hands."

Overall: 3/7
Prompt: 2/7

Shape, Form, Geometry: arm/hand deformed and weird

"A female TED Talk speaker receives a standing ovation."

Overall: 2/7
Prompt: 3/7

Semantic Mismatch with Text: the female speaker does not receive standing ovation

"Nefertiti looking at conspirators while holding a poisoned chalice."

Overall: 5/7
Prompt: 6/7

Visual Artifacts: Person with six fingers

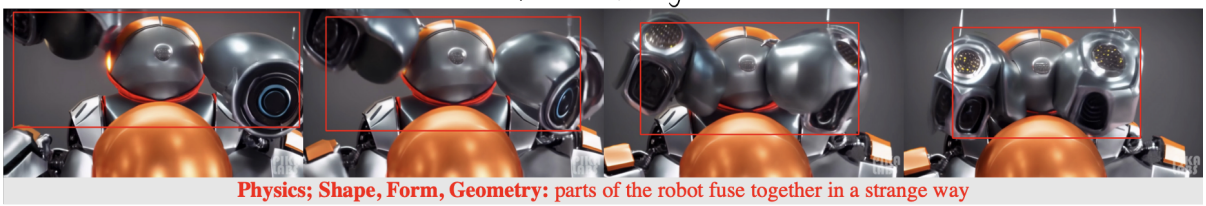
"A cute dog playing with a girl."

Overall: 2/7
Prompt: 6/7

Shape, Form, Geometry; Motion: Dogs face expands greatly.

"A robot is dancing."

Overall: 2/7
Prompt: 1/7

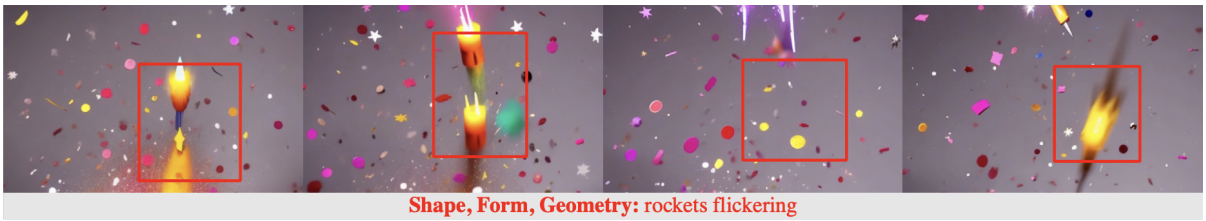


Physics; Shape, Form, Geometry: parts of the robot fuse together in a strange way

Figure 8. Annotations for Pika model.

"Confetti rockets."

Overall: 5/7
Prompt: 7/7



Shape, Form, Geometry: rockets flickering

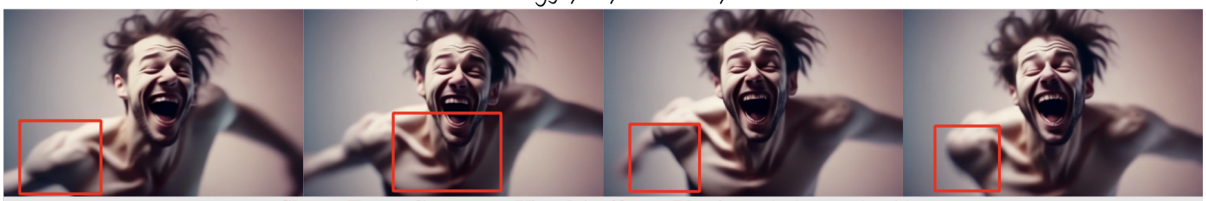
"A time-traveling rabbit wizard stands on a time platform in space, next to a rotating clock."

Overall: 6/7
Prompt: 6/7

Semantic Mismatch with Text: Clock does not rotate

"A man dancing joyfully in a dreamy scene."

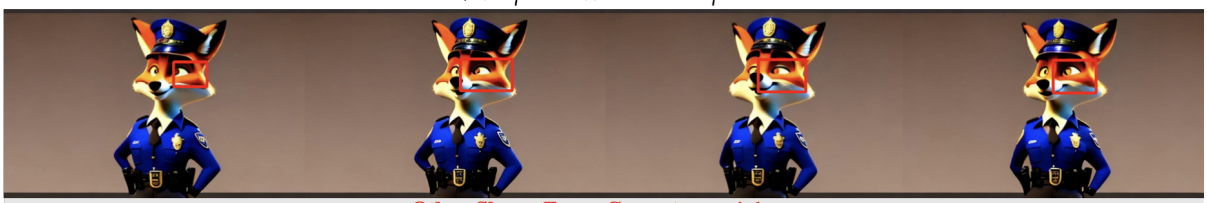
Overall: 3/7
Prompt: 4/7



Shape, Form, Geometry; Visual Artifacts; Physics: misshapen shoulder

"A fox police officer in Zootopia."

Overall: 6/7
Prompt: 6/7



Other, Shape, Form, Geometry: weird eye

"An owl hooting in the distance."

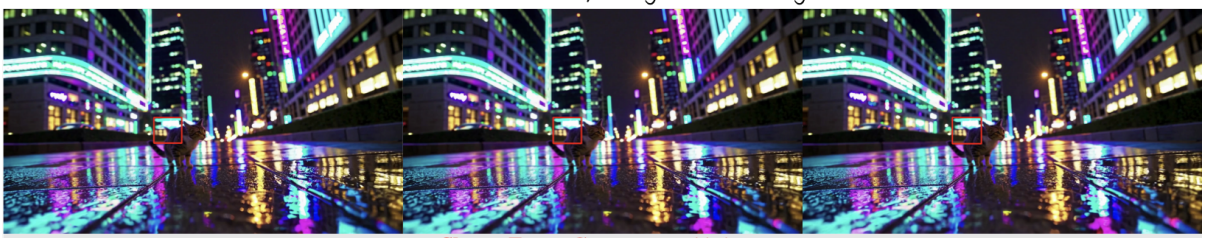
Overall: 4/7
Prompt: 6/7

Motion; Semantic Mismatch with Text: Owl is not hooting

Figure 9. Annotations for VC model.

"A confused cat is in a city during a rain of LED lights."

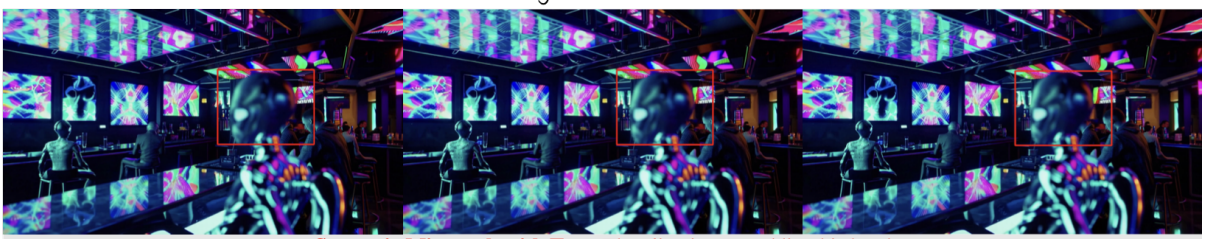
Overall: 7/7
Prompt: 7/7



Shape, Form, Geometry: odd shape

"An alien nodding his head in a futuristic bar."

Overall: 5/7
Prompt: 5/7



Semantic Mismatch with Text: The alien is not nodding his head

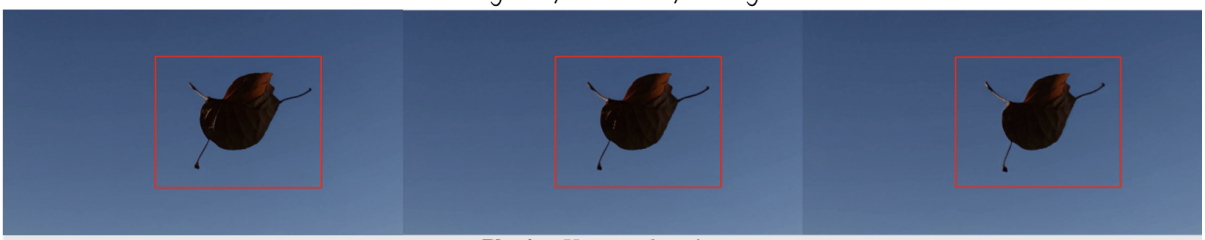
"A technician operating an AOI machine to inspect fabrics."

Overall: 3/7
Prompt: 5/7

Motion: unrealistic motion of the material

"A leaf falling slowly from the sky to the ground."

Overall: 2/7
Prompt: 4/7



Physics: Un natural motion

"A Roman general."

Overall: 4/7
Prompt: 5/7

Shape, Form, Geometry; Visual Artifacts: Unusual two objects

Figure 10. Annotations for Sora model.