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, ms14127, ps5688, kennychen, n.williams, qisun}@nyu.edu

Abstract

Recent advances in probabilistic generative models have extended capabilities from static image synthesis to text-driven video generation. However, the inherent randomness of their generation process can lead to unpredictable artifacts, such as impossible physics and temporal inconsistency. Progress in addressing these challenges requires systematic benchmarks, yet existing datasets primarily focus on generative images due to the unique spatio-temporal complexities of videos. To bridge this gap, we introduce GeneVA, a large-scale artifact dataset with rich human annotations that focuses on spatio-temporal artifacts in videos generated from natural text prompts. We hope GeneVA can enable and assist critical applications, such as benchmarking model performance and improving generative video quality.

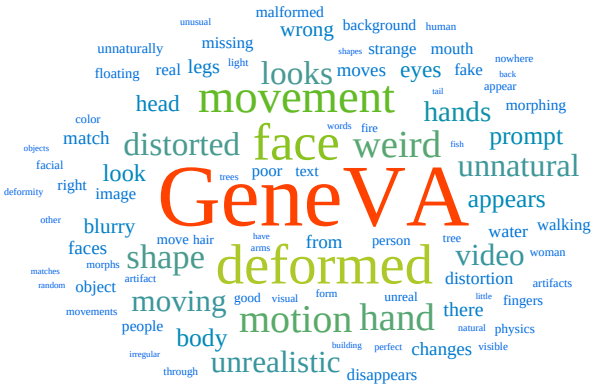


Figure 1. Human annotators were asked to describe, in a free text form, artifacts in AI-generated video. A word cloud of the most frequent artifacts mentioned by the annotators is shown here.

1. Introduction

Recent advancements in diffusion models have significantly advanced text-driven visual generation, extending capabilities from static image generation to dynamic video content [14, 18, 21, 43]. Despite these breakthroughs, state-of-the-art text-video generative models often introduce a number of unique artifacts no present in still images [5]. In AI-generated videos, both spatial (e.g., distorted geometry or inconsistent appearance) and temporal (e.g., physically implausible motion or temporal incoherence) artifacts arise. A fundamental roadblock in characterizing and mitigating these unique artifacts is the lack of data and metrics to assess them. Crucially, while metrics [17, 46] may capture some sense of overall quality, assessment by real human observers is the highest standard for quality evaluation.

A wide body of research has studied the development of computational metrics to predict visual quality with the goal that the metric correlates well with real subjective responses. Popular examples of these automatic metrics include SSIM [46], LPIPS [39]. A number of metrics have also been used to optimize and evaluate text-image generative models. For example, the Fréchet Inception Distance (FID) score [17] has been widely adopted to measure generation realism. However, artifacts that arise from text-video generation may extend beyond what can be captured by these metrics. An extension of FID to video [34, 35] has been found to result in poor correlations with subjective judgements in some cases [12]. To the best of our knowledge, a dataset that comprehensively evaluates and annotates these generative text-video artifacts does not exist, perhaps due to the difficulty in collecting large-scale data of the sort.

In this paper, we present GeneVA, a rich human-annotated dataset for **gene**rative text to **v**ideo **a**rtifacts. GeneVA offers a representative set of 5,452 text prompts sampled from real-world human text inputs [38]. Then, 3 text-video generative models (2 open-source and 1 commercial) were used to generate, with the text prompts as input, 16,356 videos amounting to a sum $52,326$ seconds of content. In total, 16,451 crowd-sourced annotations of spatio-temporal artifacts was collected. Annotated artifacts include both categorical selections and free-response textual descriptions (as shown in Figure 2). Our goal with GeneVA is to provide a robust, challenging dataset to help the research community make systematic progress on understanding complex visual artifacts. The long-term goal is to leverage this understanding to build more reliable and human-centric generative models.

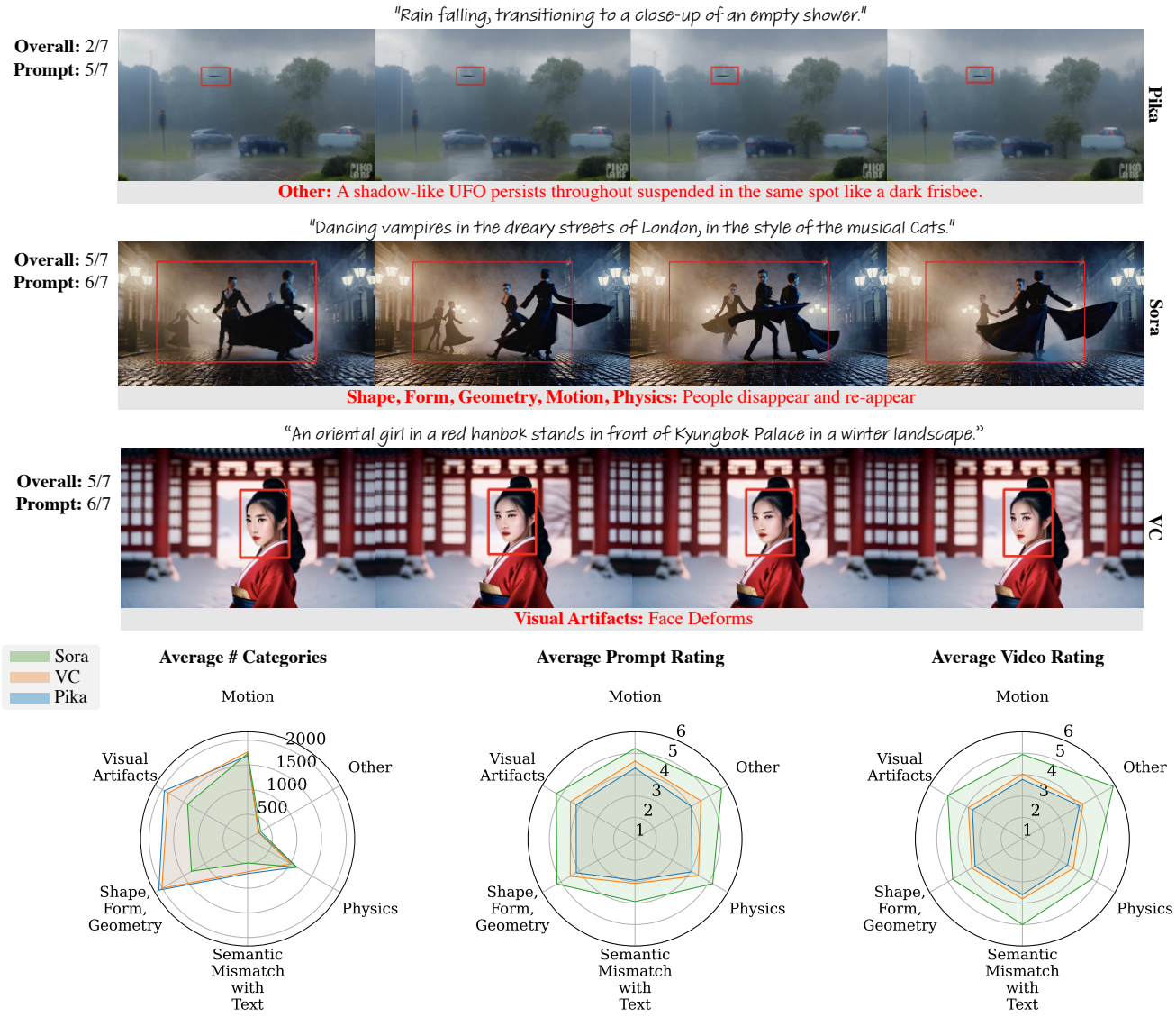


Figure 2. We show example annotated bounding boxes for each model (labeled on the right). The bounding boxes are annotated in red, with their artifact category and user-annotated description below the frames. Video quality ("Overall") and video-prompt alignment ("Prompt") are shown to the left. Summary statistics are shown in the radar plots. Specifically, we show statistics for each artifact category, grouped by category count, average video-prompt alignment, and average video quality rating given the user-selected artifact categories. This is done for the three models in our dataset: Sora [3], VideoCrafter2 [6], and Pika [2]. See additional examples in the Appendix.

As a first step toward this goal, we also leveraged these annotations to develop an interpretable artifact detector pipeline (Section 5.1). In addition to identifying artifact locations, our system also generates human-readable explanations for each artifact, enabling a fine-grained analysis of the generative model's behavior. The artifact detector model achieves 13% Average Precision ($AP_{0.25}$) and demonstrates strong zero-shot generalization when applied to videos from external models not seen during training. This highlights the robust and generalizable nature of our dataset.

We will release the GeneVA dataset and the trained detector to the public upon acceptance. We envision the dataset could serve as a new benchmark for future text-video generative models, and can launch new research in the direction of artifact reduction and detection in AI-generated videos.

2. Related Work

In this section, we summarize prior datasets on text-video generation. We make an effort to focus on datasets that collect human quality assessments of these videos. A summary of the most relevant datasets is tabulated in Table 1.

Table 1. Here, we shows existing datasets that are the most relevant to this work. We list the total number of models each work used to generate their video dataset, the total number of videos in the dataset, the total number of human ratings, whether participants had to describe artifacts, and whether participants localized artifacts by annotating bounding boxes. Notably, our dataset is the only one to include bounding boxes that specify the artifact locations and text annotations that describe their nature.

Dataset	Total Model Count	Total Video Count	Total Human Rating Count	Description?	Bounding Box?
T2V [10]	5	1,005	48,240	✗	✗
Vbench [20]	4	∼1,700	–	✗	✗
FETV [28]	4	2,613	28,116	✗	✗
EvalCrafter [27]	5	2,500	8,647	✗	✗
VideoScore [16]	11	37,600	–	✗	✗
GeneVA	3	16,356	16,356	✓	✓

2.1. Datasets for Annotating Generative Videos

The proliferation of generative AI has created a need for the quality assessment of AI-generated images and videos [26]. Numerous datasets have been made available that include human assessments of AI-generated images [8, 23, 42, 44]. Fewer datasets of this kind have been proposed for text-video content. Chen et al. [9] collected quality data for 9,180 videos, focusing specifically on video-action pairs. VideoReward [25] and IPO [45] collected pairwise comparison data for content generated by a number of different video generation models. Chivileva et al. [10] and Huang et al. [20] collected human quality ratings and measured their alignment with quality metrics. The VideoScore [16], FETV [28], and EvalCrafter [27] datasets collected human rating across various aspects (temporal consistency, text alignment, etc.) on a large set of generated videos. None of these works, however, include bounding boxes on artifacts or free-form text descriptions of them.

2.2. Artifact Detection

The detection and localization of artifacts in text-image and text-video content can facilitate identification of AI-generated content. This could be useful for detecting fraud or mitigating the spread of misinformation, for example [30, 31]. Several datasets are available for annotated bounding boxes in text-image content [4, 11, 37, 40]. In this work, we focus on the localization of artifacts in AI-generated videos. Very few works have collected a dataset of this sort for text-video generation; one such example is for the task of deepfake detection by Hondru et al. [19]. No prior works that we are aware of provide annotated bounding boxes for general text-video content.

3. The GeneVA Dataset

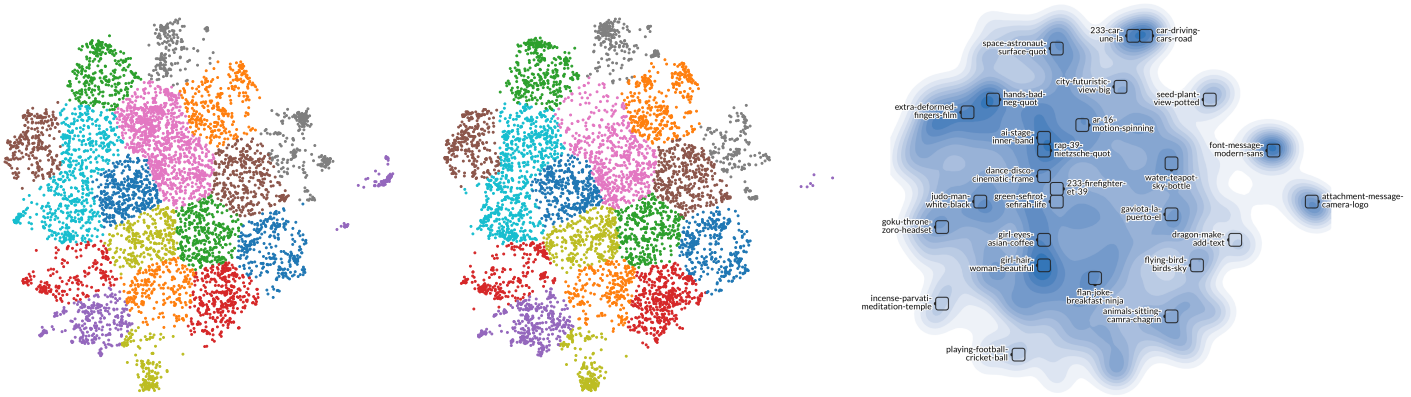
In this section, we describe how the text-video dataset is collected and how human annotators label each video.

3.1. Synthetic Video Collection from Text-to-Video Models

To effectively analyze artifacts in text-to-video generation, we first require a robust and representative dataset of synthetic videos. Our goal is to capture realistic scenarios and diverse content typically encountered in real-world text-to-video generation use cases.

We constructed our dataset by leveraging the large-scale VidProM dataset [38], which contains 1.67 million unique prompts derived from real-world user interactions (scraped from Pika Discord servers) and corresponding videos from multiple text-video generation models. To extract a representative subset that maintains the distributional properties of this extensive collection, we employed kernel herding [7] on 3,072-dimensional embeddings of the prompts, which were generated using OpenAI's text-embedding-3-large model [1]. We sampled 6,000 prompts using this approach, with the goal of ensuring fair coverage across the semantic space of prompts, avoiding sampling bias while preserving the diversity of the original dataset. Figure 3c visualizes the distribution of our selected prompts. More details are available in the Appendix.

To maximize diversity in terms of generated videos and potential artifacts, our dataset incorporates videos from three recent generative text-video models: Sora [3], Pika [2], and Videocrafter2 (VC) [6]. Videos generated by Pika and VC are directly extracted from the VidProM dataset [38]. We supplemented these with newly generated Sora videos to capture the current state-of-the-art performance and ensure our analysis reflects the most recent advances in text-to-video generation. For the Sora model, we generated 5-second videos for each sampled prompt. Some prompts from our original prompt sampling were excluded due to moderation restrictions of the Sora model as well as our own manual review to ensure appropriate video content. This careful curation resulted in a final dataset of 5,452 unique prompts and 16,356 synthetic videos across the three models, yielding a total of $52,326$ seconds of diverse video content.



(a) original embedding space **(b)** cleaned embedding space **(c)** language prompts

Figure 3. (a)/(b) visualize the embedded feature space before and after our cleaning procedure. We note that the two plots look very similar. Plots with different cluster counts are displayed in the Appendix. (c) We visualize the embedding space of prompts used in our dataset via Wizmap [41]. Darker colors represent denser regions, which are labeled with their descriptors.

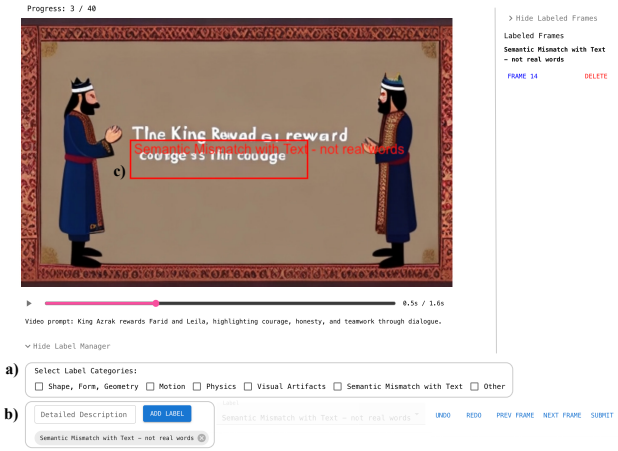


Figure 4. A screenshot of our video annotation interface. Users annotate artifacts by (a) selecting a label category, (b) describing the artifact in additional detail (free response), and (c) drawing a bounding box around the artifact. Bounding boxes for the same labeled artifact are interpolated across multiple frames as long as the user draws an initial and final keyframe bounding box for the artifact. Annotated bounding boxes are listed in the right side column (enumerated per-frame) to allow for easy understanding of which frames have been annotated and simple editing if users want to change their boxes before submitting.

3.2. Human Annotations

Our goal is to gather human annotations for our curated text-and-video pairs. Specifically, we aim to collect participants' quality ratings of each video, as well as annotations of artifacts, including their category, description, and location. To gather this data, we run a large-scale crowdsourced study described below.

Participants We recruited study participants through the online crowdsourcing platform *Prolific*. All study protocols were approved by an institutional review board (IRB), and subjects were compensated at a rate of $10/h.

Study Protocol A screenshot of our web-based annotation tool is shown in Figure 4. The study was conducted via a web browser, and run on the participant's computer screen. We did not enforce any requirements on users' display specifications to better reflect real-world video viewing conditions. Subjects were shown a 3-minute tutorial video on example artifact labeling, controls, and procedures before proceeding with the annotation task. Each video was played in its entirety before participants were allowed to begin the annotation phase. During the annotation phase, participants are shown the video and its corresponding text prompt. For clarity, we provide the cleaned versions of the original VidProM prompts that preserves the original's semantic intent (see Figures 3a and 3b and Appendix for more details). Participants annotated artifacts by selecting a label category from a pre-defined set (Figure 4 (a)) and describing the artifact in additional detail (Figure 4 (b)). Artifact categories are selected to cover the most representative artifact types in AI-generated video, and includes 1) Shape, Form, Geometry, 2) Motion, 3) Physics, 4) Visual Artifacts, and 5) Other. Participants were free to scrub through the video timeline to jump to specific frames of the video. For each label, participants had the choice to draw a bounding box around the artifact (Figure 4 (c)). Bounding boxes for the same labeled artifact are interpolated across multiple frames.

To encourage users to prioritize reporting salient artifacts rather than over-scrutinizing the videos, we set a limit of five artifact annotations per video. This ensures that our data primarily captures artifacts most likely to be noticed by a

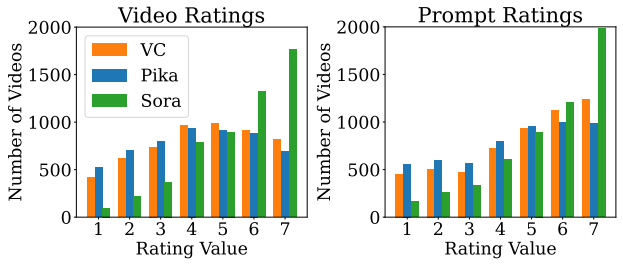


Figure 5. Distributions of user-submitted scores for each video's (left) visual quality and (right) alignment between text prompt and generated content. Both the video and text quality scores are notably higher for Sora videos compared to the videos generated by VC and Pika, which share very similar score distribution shapes.

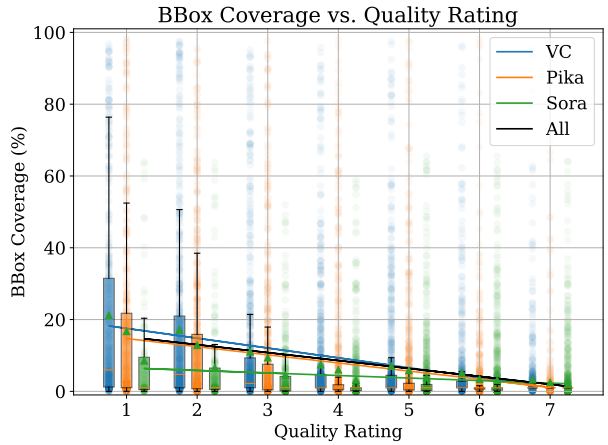


Figure 6. Box plots showing the distribution of spatio-temporal bounding box coverage across all videos, grouped by generative model and video quality score. Whiskers represent 95% confidence intervals. We also show the regression fits that denote the correlation between video quality and bounding box coverage. The black line shows the regression for the pooled data of all models.

typical user. Once the user submits their labeled artifacts, they rate the video's visual quality and its alignment with the text prompt. Users submitted ratings on two separate 7-point Likert scales (where 1 being the lowest quality and 7 the highest).

4. Dataset Characteristics

To better understand what kinds of video content our subjects annotated, in this section, we analyze and summarize various properties of the AI-generated videos. A base summary of the free-form artifact descriptions users submitted is visualized in Figure 1, which shows the most common words used to describe artifacts. The distribution of video quality and text prompt-alignment ratings is shown in Figure 5, separated by model. Interestingly, text prompt alignment scores were right-skewed, but seem normally-distributed for video quality (except for Sora).

Table 2. We show per-model characteristics, comparing videos having a semantic match with the text prompt and those with a mismatch. These results highlight the importance of ensuring that generated videos are aligned with the semantics of the text prompts used to generate them.

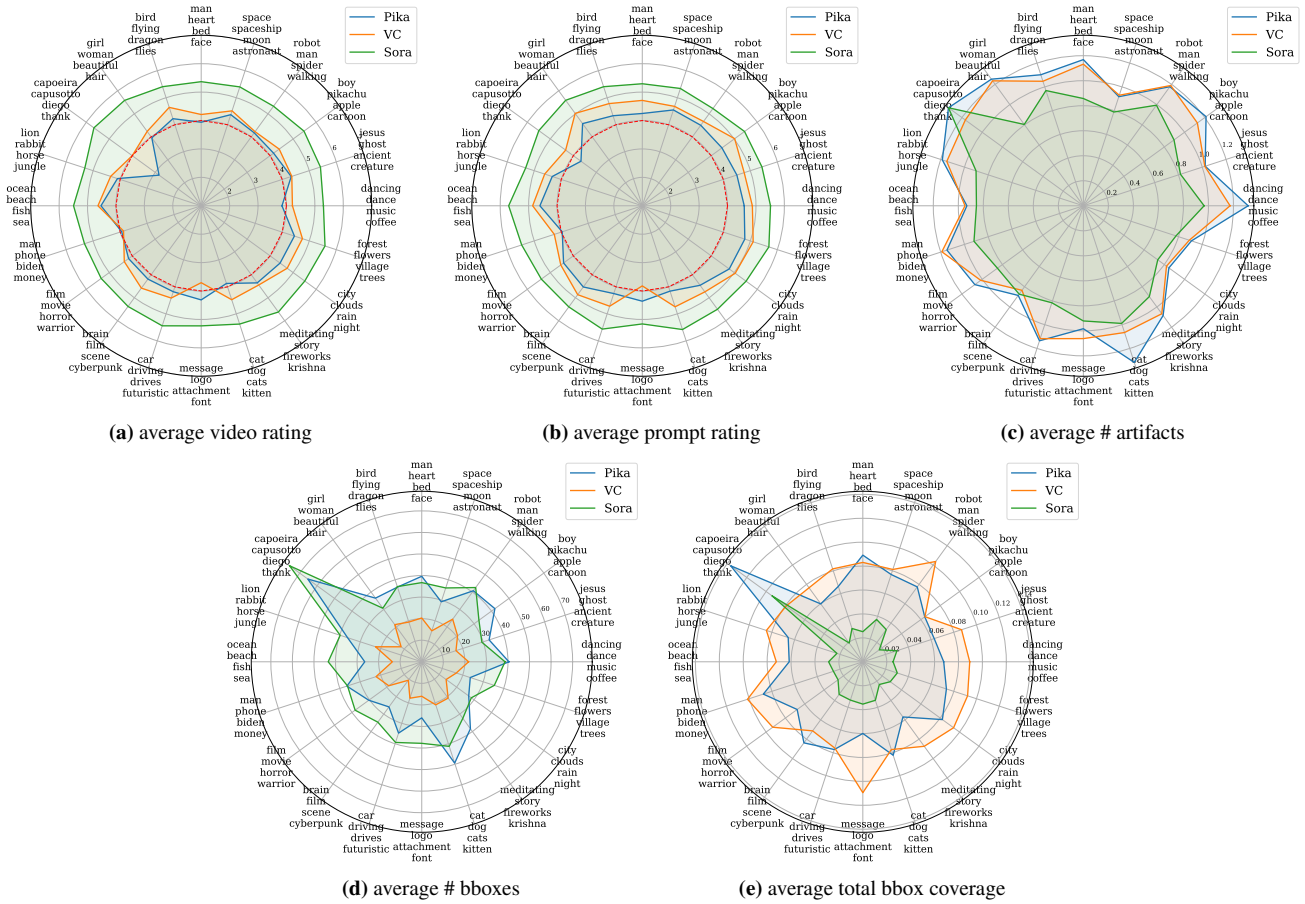
Model	Category	Quality	Prompt	# Artifacts	# BBoxes	# Samples
Pika	Semantic Mismatch	3.604	2.938	1.59	48.12	704
Pika	Semantic Match	4.268	4.677	1.02	32.25	4743
VC	Semantic Mismatch	3.799	3.084	1.54	27.29	667
VC	Semantic Match	4.453	4.971	0.99	17.75	4785
Sora	Semantic Mismatch	4.998	3.933	1.48	76.00	489
Sora	Semantic Match	5.460	5.600	0.81	32.25	4963

Video summary statistics are shown in Figure 7, grouped into 20 different semantic clusters (results for additional cluster counts are displayed in the Appendix). To interpret the discovered text prompt clusters, we analyzed their content using TF-IDF and embedding-based proximity. First, we filtered for prompts from clean sources and computed their TF-IDF representations using a vocabulary of the top 10,000 words after removing English stopwords. For each cluster, we calculated the mean TF-IDF scores and compared them to global means across all prompts to identify cluster-specific keywords. A log-ratio-based scoring function was applied to highlight words that are both frequent within a cluster and relatively rare globally. The top 4 scoring words were selected as representative keywords for each cluster.

Based on the trends shown in Figure 7, we observe that Sora generally outperforms Pika and VC, scoring higher on average prompt and video quality ratings, as well as lower on the average number of artifacts and total bounding box coverage. Furthermore, VC and Pika video ratings hover around the midpoint of the 7-point rating rating scale (red dashed line in Figures 7a and 7b, while Sora is much higher than the midpoint rating. Interestingly, VC scores notably lower on the average number of bounding boxes per cluster.

Since we are studying the quality of text-to-video generative models, we also need to consider the extent to which the semantic alignment between generated videos and their input text prompts influences users' subjective rating of video quality. Table 2 shows the average quality score for videos, grouped by model and whether or not users reported a semantic match between the video content and the text prompt. Unsurprisingly, the average quality and prompt alignment scores are lower for semantic mismatches. Furthermore, videos that had a semantic mismatch received more reports of artifacts ('# Artifacts') and had more frames on average with at least one bounding box annotation ('# BBoxes').

Finally, we inspected the correlation between video quality scores and the average spatio-temporal coverage of bounding boxes across all videos. Spatio-temporal coverage was defined as the percentage of frames and pixels that were covered by at least one bounding box for a given



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

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

Figure 7. Dataset statistics grouped by the top 20 semantic clusters and grouped by model. The example words for each cluster are the top 4 words in that cluster (zoom for better view). The dashed red line in subplots (a) and (b) indicate the midpoint of the 7-point rating scale.

video. Figure 6 shows the box plot distributions of bounding box coverages for each video, as well as the regression lines for the correlation between video quality scores and bounding box coverage. Unsurprisingly, we see a weak but statistically significant inverse correlation between quality scores and bounding box coverage for all models. As the quality ratings of videos increase, the amount of bounding box coverage decreases, indicating that higher-rated videos have fewer perceived artifacts. The strongest correlation (Pearson) between scores and coverage was found in the VC model ($r = -0.27, p < 0.0001$) , with Pika having the next strongest correlation ($r = -0.25, p < 0.0001$), and Sora with the weakest correlation ($r = -0.13, p < 0.0001$). For the data pooled across all models, this trend continues ($r = -0.26, p < 0.0001$). Note that although the correlation is statistically significant, the strength is relatively weak, which suggests that the prevalence of bounding boxes in videos cannot be simply predicted by video quality score alone and that other features (e.g. temporal consistency, shape deformation, etc.) may also contribute to the prevalence of bounding box annotations.

5. How Can We Use GeneVA?

To demonstrate how GeneVA can be utilized, we introduce a system (Section 5.1) that automatically detects and describes artifacts in generated videos. These outputs can then, in theory, be used to improve video quality via downstream tasks such as video inpainting.

5.1. Automated Artifact Detection

Given a video input, our goal is two-fold: 1) identify the location of any immersion-breaking artifact, and 2) describe the type of artifact detected. To accomplish these goals, we propose a dual-stage pipeline that first identifies an artifact's coordinates with the **Artifact Detector**, then generates a corresponding textual description with the **Artifact Caption Generator**. To leverage the strength of pre-trained image-based models, we also introduce a **Temporal Fusion Module** that learns to attend to temporally relevant information across consecutive frames. Figure 8 shows the overview of our pipeline.

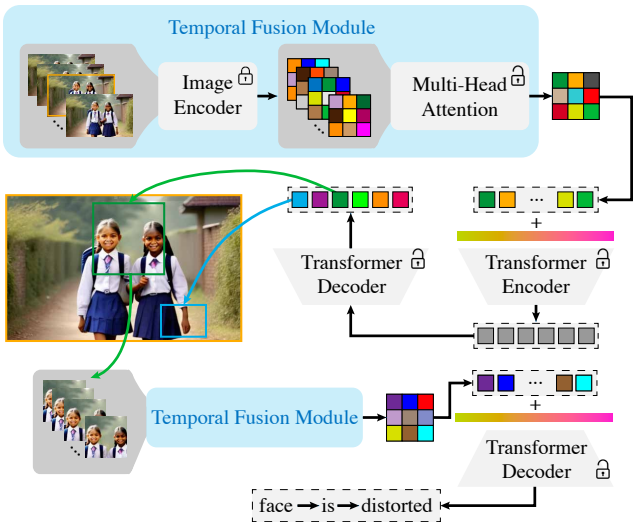


Figure 8. Overview of our proposed pipeline for interpretable artifact detection. First, the model processes a sequence of frames to identify the location of an artifact in the central anchor frame (highlighted in yellow). Following this localization, the system then generates a textual description to explain the nature of the detected artifact.

5.1.1. Implementation Details

Temporal Fusion Module Given a video of T consecutive frames $\mathcal{V} = \{I_0, \ldots, I_{T-1}\}$, a frozen backbone (e.g. CNN) is used to extract a sequence of per-frame feature vectors, $\phi(\mathcal{V}) \in \mathbb{R}^{T \times d}$. This is passed through a temporal self-attention layer which projects the features at each spatial location into queries (Q), keys (K), and values (V) and calculates the attention-weighted representation $\mathrm{Softmax}\big(QK^\top/\sqrt{d_h}\big)\,V$ that fuses information across frames. Attention parameters are learned jointly during the training of each stage. In all experiments, we set $T = 5$ frames, sampled at 10 FPS, to give the model temporal context for predicting artifacts of the central anchor frame.

Artifact Detector Existing object-detection pipelines often rely on objecness scores (e.g., Faster R-CNN [33], YOLOs [13, 22, 32]) which prioritize regions that resemble objects. Video artifacts, however, can emerge anywhere in the frame and may not resemble any conventional objects. To avoid this bias, we adopt the end-to-end transformer-based architecture proposed in RT-DETR [29, 47], which employs an efficient hybrid encoder to achieve state-of-the-art real-time performance. The model is initialized with pre-trained weights (ResNet18 [15]) and then jointly trained with the temporal fusion module for video artifact localization.

Artifact Caption Generator The goal of this stage is to generate a textual description for each artifact region. For training, the model uses regions defined by ground truth annotations, while during inference, it processes the regions provided by our detector. We adopt the GIT architecture [36], which utilizes a transformer decoder to autoregressively generate text from visual embeddings. We use the pre-trained weights (Vision Transformer from [36]) as our initialization and jointly train with our temporal fusion module to generate human-like textual descriptions of the localized artifacts.

Table 3. Quantitative results for our artifact detection system on the held-out test set. We report Average Precision (AP) at specific IoU thresholds (AP_{25}, AP_{50}, AP_{75}). AP denotes the standard COCO metric [24], average over IoU thresholds from 0.5 to 0.95 in the step of 0.05.

Method	$AP_{.25}$	$AP_{.50}$	$AP_{.75}$	AP
Ours	**0.132**	**0.091**	**0.004**	**0.032**
w/o temporal fusion	0.103	0.057	0	0.020

5.1.2. Results

For this proof-of-concept study, we randomly sampled a subset of 1,000 videos from our full dataset. This subset is then split at the video level into an 80/10/10 for training, validation, and testing, respectively. We used the validation set to determine the optimal training epoch via early stopping and report final performance on the held-out test set.

As shown in Table 3, for artifact detection, our system achieves an Average Precision ($AP_{.25}$) of 13% on the held-out test set. This reflects the model's ability to identify the general location of an artifact, where a detection is considered correct if it overlaps the true region by at least 25%. This demonstrates that our model is learning a meaningful signal from the data. For context, a random baseline achieves an $AP_{.25}$ of $\approx 1\%$ on the same task. We also conducted an ablation study using single-frame inputs to isolate the effect of temporal fusion. The lower performance (-3% in $AP_{.25}$) when removing temporal context indicates that temporal context provides important clues in identifying video artifacts, many of which are inherently temporal in nature.

While these results are promising, they also highlight the difficulty of the task and the substantial gap that remains to fully match the subtle way humans identify artifacts in generated content. Understanding this aspect is a critical step toward creating a new generation of generative models that align with human perception.

5.2. Cross-model generalization

Our dataset includes labeled artifacts from three different generative models. The goal is to capture a model-agnostic representation, thereby avoiding overfitting to the unique artifacts of any single generative model. To validate this, we apply our pre-trained artifact detector system (Section 5.1)

Figure 9. Results of our artifact detection pipeline on an unseen generative model. Our system identifies and describes artifacts by making predictions on the highlighted central frame (in orange) of a video sequence. The detected artifacts are visualized as red bounding boxes, and the predicted descriptions are included above the sequence. The videos are generated by Google's Veo model using text prompts from our dataset. See the appendix for more results.

to videos generated by Google's Veo[1], a state-of-the-art text to video model *that was not represented in our training data*. Qualitative results are shown in Figure 9. Our pipeline successfully identifies and provides coherent descriptions for typical artifacts in Veo's outputs, such as unnatural movement – a common issue across various generative models. This strong zero-shot performance indicates that our dataset encourages the learning of a general representation of video artifacts, making it a robust tool for evaluating novel generative models.

6. Limitations and Future Work

Repeated annotations. While conducting annotations under a constrained budget, we faced a key trade-off between the number of unique videos and the number of repeated evaluations per video. For GeneVA, we prioritized maximizing the diversity of videos to capture a broader range of language-conditioned contexts. As a result, the current dataset does not include multiple repeated annotations for the same video. Future extensions incorporating repeated evaluations would enable analysis of inter-observer variability and provide a probabilistic understanding of artifact detectability. Moreover, demographic information was not collected from crowd-sourced participants to maintain privacy. Additional studies with repeated experiments and optional demographic reporting could also extend data diversity based on randomly sampled participants.

Number of artifacts annotated. A key challenge in crowdsourced studies is ensuring data quality through unambiguous task design. In our pilot experiments, we found that participants often had inconsistent interpretations of visual artifacts, particularly given their generative nature. To address this, we instructed users to label no more than five artifacts, encouraging them to focus on the most salient ones, with the order indicating perceived importance. In the future, we plan to develop more adaptive strategies by leveraging insights from the initial GeneVA dataset to predict the likely number of artifacts dynamically.

Potential societal impacts. AI-generated text-video content inherently leads to ethical issues, such as the spread of misinformation. By collecting a dataset of localized artifact labels, future models could potentially be trained to inpaint these artifacts, thereby further accelerating the quality of text-video generation. We argue that this dataset can be leveraged for good – better artifact detection, driven by real human annotations, can aid in automatic detection of AI-generated content where it may be difficult to identify.

7. Conclusion

We introduced GeneVA, a large-scale, human-annotated dataset to address a critical gap in understanding artifacts in AI-generated video. By combining structured annotations with free-form feedback, GeneVA offers a rich and flexible resource for analyzing and predicting artifact occurrence at multiple levels of detail. Our proof-of-concept experiments demonstrate its application for training effective artifact prediction models. We anticipate that the public release of GeneVA will accelerate research on evaluating, benchmarking, detecting, and potentially mitigating artifacts in next-generation video generation systems.

References

[1] text-embedding-3-large. 3
[2] Pika art. Accessed: July 10, 2025. 2, 3
[3] OpenAI. Video generation models as world simulators. Accessed: July 10, 2025. 2, 3

[1] https://deepmind.google/models/veo/

[4] Bin Cao, Jianhao Yuan, Yexin Liu, Jian Li, Shuyang Sun, Jing Liu, and Bo Zhao. Synartifact: Classifying and alleviating artifacts in synthetic images via vision-language model. *arXiv preprint arXiv:2402.18068*, 2024. 3

[5] Han Chen, Yuezun Li, Dongdong Lin, Bin Li, and Junqiang Wu. Watching the big artifacts: Exposing deepfake videos via bi-granularity artifacts. *Pattern Recognition*, 135:109179, 2023. 1

[6] 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. 2, 3

[7] Yutian Chen, Max Welling, and Alexander J. Smola. Supersamples from kernel herding. *CoRR*, abs/1203.3472, 2012. 3

[8] Zijian Chen, Wei Sun, Haoning Wu, Zicheng Zhang, Jun Jia, Zhongpeng Ji, Fengyu Sun, Shangling Jui, Xiongkuo Min, Guangtao Zhai, et al. Exploring the naturalness of ai-generated images. *arXiv preprint arXiv:2312.05476*, 2023. 3

[9] Zijian Chen, Wei Sun, Yuan Tian, Jun Jia, Zicheng Zhang, Jiarui Wang, Ru Huang, Xiongkuo Min, Guangtao Zhai, and Wenjun Zhang. Gaia: Rethinking action quality assessment for ai-generated videos. In *Advances in Neural Information Processing Systems*, pages 40111–40144. Curran Associates, Inc., 2024. 3

[10] Iya Chivileva, Philip Lynch, Tomas E Ward, and Alan F Smeaton. Measuring the quality of text-to-video model outputs: Metrics and dataset. *arXiv preprint arXiv:2309.08009*, 2023. 3

[11] Guian Fang, Wenbiao Yan, Yuanfan Guo, Jianhua Han, Zutao Jiang, Hang Xu, Shengcai Liao, and Xiaodan Liang. Human-refiner: Benchmarking abnormal human generation and refining with coarse-to-fine pose-reversible guidance. In *European Conference on Computer Vision*, pages 201–217. Springer, 2024. 3

[12] Songwei Ge, Aniruddha Mahapatra, Gaurav Parmar, Jun-Yan Zhu, and Jia-Bin Huang. On the content bias in fréchet video distance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7277–7288, 2024. 1

[13] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 7

[14] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *The Twelfth International Conference on Learning Representations*. 1

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778. IEEE, 2016. 7

[16] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narsupalli, Rongqi Fan, Zhiheng Lyu, Yuchen Lin, and Wenhu Chen. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. *ArXiv*, abs/2406.15252, 2024. 3

[17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 1

[18] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022. 1

[19] Vlad Hondru, Eduard Hogea, Darian Onchis, and Radu Tudor Ionescu. Exddv: A new dataset for explainable deepfake detection in video. *arXiv preprint arXiv:2503.14421*, 2025. 3

[20] Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, et al. Vbench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21807–21818, 2024. 3

[21] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15954–15964, 2023. 1

[22] Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements, 2024. 7

[23] Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems*, 36:36652–36663, 2023. 3

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 7

[25] Jie Liu, Gongye Liu, Jiajun Liang, Ziyang Yuan, Xiaokun Liu, Mingwu Zheng, Xiele Wu, Qiulin Wang, Wenyu Qin, Menghan Xia, et al. Improving video generation with human feedback. *arXiv preprint arXiv:2501.13918*, 2025. 3

[26] Xiaohong Liu, Xiongkuo Min, Guangtao Zhai, Chunyi Li, Tengchuan Kou, Wei Sun, Haoning Wu, Yixuan Gao, Yuqin Cao, Zicheng Zhang, et al. Ntire 2024 quality assessment of ai-generated content challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6337–6362, 2024. 3

[27] Yaofang Liu, Xiaodong Cun, Xuebo Liu, Xintao Wang, Yong Zhang, Haoxin Chen, Yang Liu, Tieyong Zeng, Raymond Chan, and Ying Shan. Evalcrafter: Benchmarking and evaluating large video generation models. 2023. 3

[28] Yuanxin Liu, Lei Li, Shuhuai Ren, Rundong Gao, Shicheng Li, Sishuo Chen, Xu Sun, and Lu Hou. Fetv: A benchmark for fine-grained evaluation of open-domain text-to-video generation. *Advances in Neural Information Processing Systems*, 36:62352–62387, 2023. 3

[29] Wenyu Lv, Yian Zhao, Qinyao Chang, Kui Huang, Guanzhong Wang, and Yi Liu. Rt-detrv2: Improved baseline with bag-of-freebies for real-time detection transformer, 2024. 7

[30] Ekta Prashnani, Michael Goebel, and BS Manjunath. Generalizable deepfake detection with phase-based motion analysis. *IEEE Transactions on Image Processing*, 2024. 3

[31] Md Shohel Rana, Mohammad Nur Nobi, Beddhu Murali, and Andrew H Sung. Deepfake detection: A systematic literature review. *IEEE access*, 10:25494–25513, 2022. 3

[32] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016. 7

[33] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 7

[34] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Fvd: A new metric for video generation. 1

[35] Thomas Unterthiner, Sjoerd Van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 1

[36] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*, 2022. 7

[37] Kaihong Wang, Lingzhi Zhang, and Jianming Zhang. Detecting human artifacts from text-to-image models. *arXiv preprint arXiv:2411.13842*, 2024. 3

[38] Wenhao Wang and Yi Yang. Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. 2024. 1, 3

[39] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 1

[40] Zeqing Wang, Qingyang Ma, Wentao Wan, Haojie Li, Keze Wang, and Yonghong Tian. Is this generated person existed in real-world? fine-grained detecting and calibrating abnormal human-body. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21226–21237, 2025. 3

[41] Zijie J. Wang, Fred Hohman, and Duen Horng Chau. Wizmap: Scalable interactive visualization for exploring large machine learning embeddings. *Association for Computational Linguistics*, 3:516–523, 2023. 4

[42] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score: Better aligning text-to-image models with human preference. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2096–2105, 2023. 3

[43] Zhen Xing, Qijun Feng, Haoran Chen, Qi Dai, Han Hu, Hang Xu, Zuxuan Wu, and Yu-Gang Jiang. A survey on video diffusion models. *ACM Computing Surveys*, 57(2):1–42, 2024. 1

[44] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:15903–15935, 2023. 3

[45] Xiaomeng Yang, Zhiyu Tan, and Hao Li. Ipo: Iterative preference optimization for text-to-video generation. *arXiv preprint arXiv:2502.02088*, 2025. 3

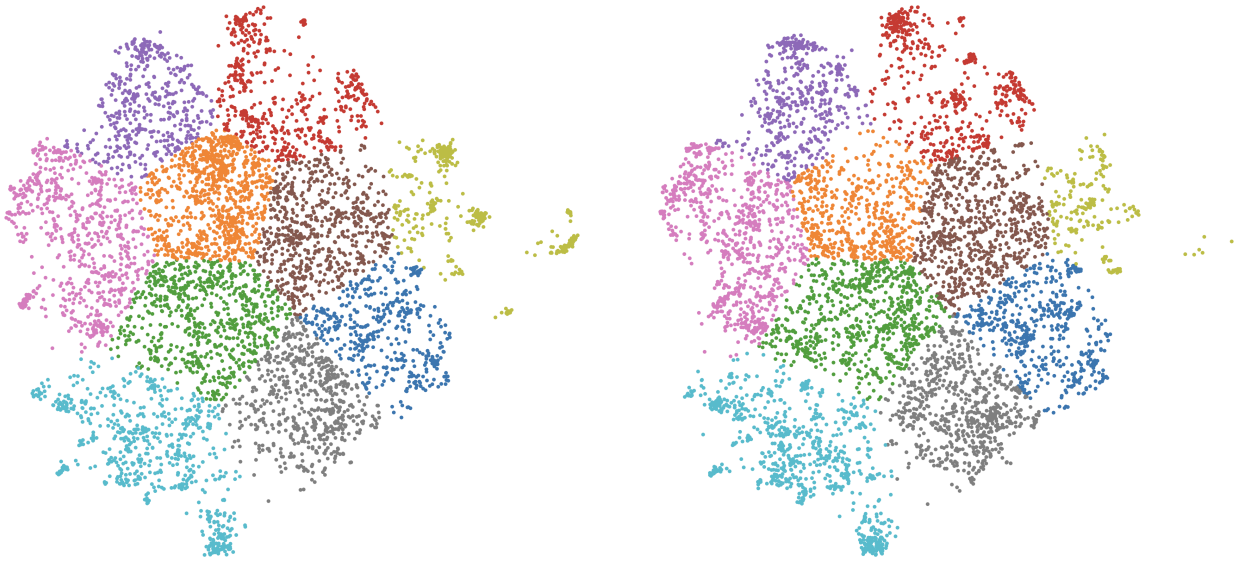
[46] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 1

[47] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection, 2023. 7

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

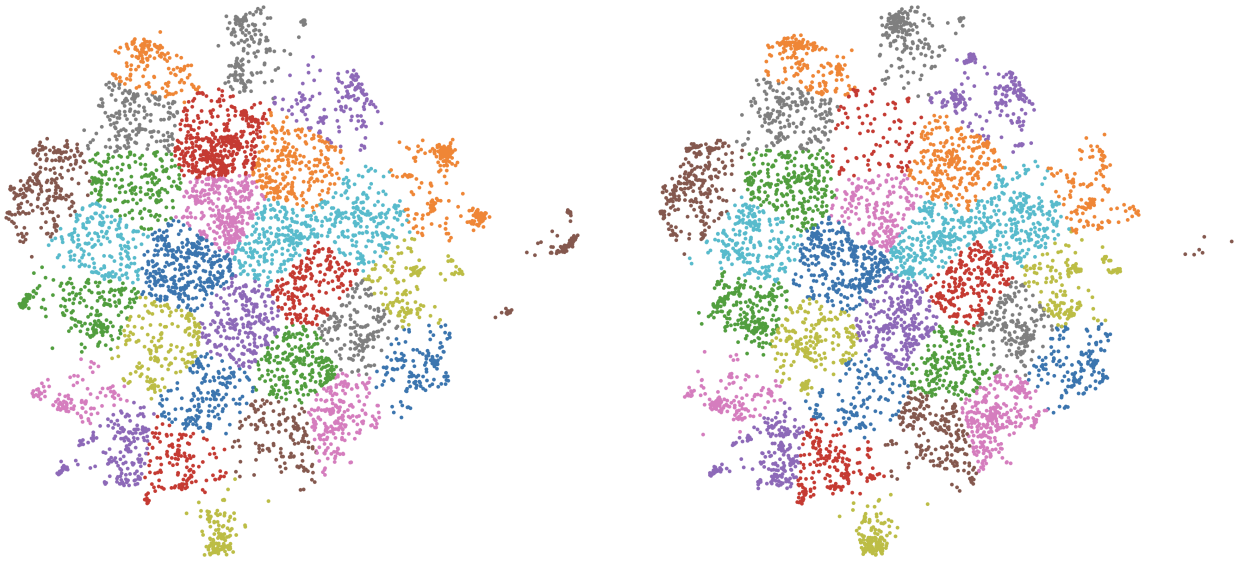
Figure 1. We show the clustered original uncleaned (left) and cleaned (right) prompts for 10 clusters.

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}
"""
```

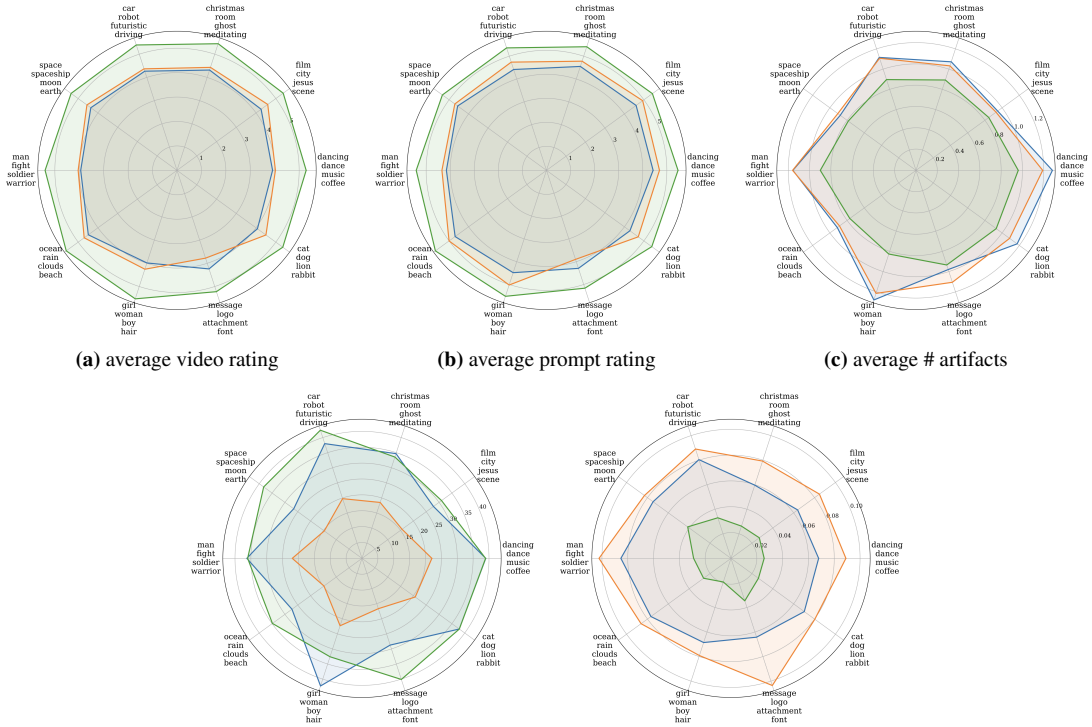
Figure 3a and Figure 3b in the main manuscript show the unclean and clean text prompt embeddings, respectively, projected onto a 2D space; qualitative inspection shows the two align well. In addition to the clustering, we manually confirmed that cleaned text prompts capture the intent of the original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

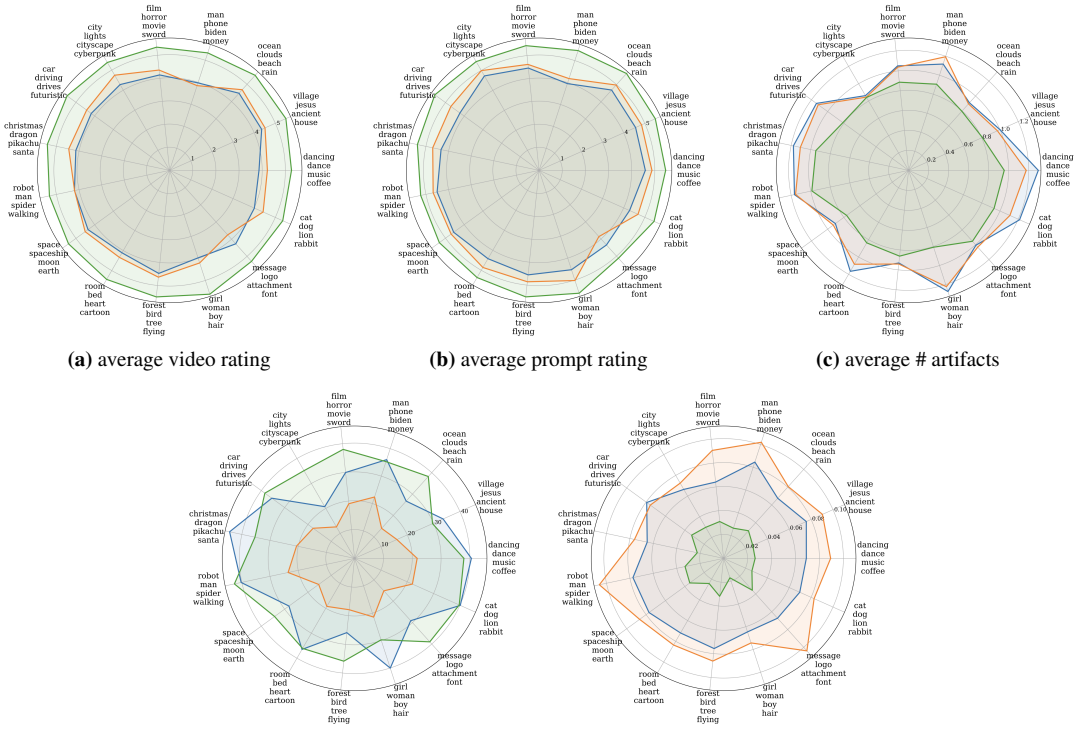
The scale of the VidProM dataset [3] paired with its lack of direct prompt-to-file mappings became a challenge as we sought to extract a specific subset. To address this, we developed a pipeline to retrieve the videos corresponding to the prompts chosen by our kernel herding sampling. This involved downloading 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

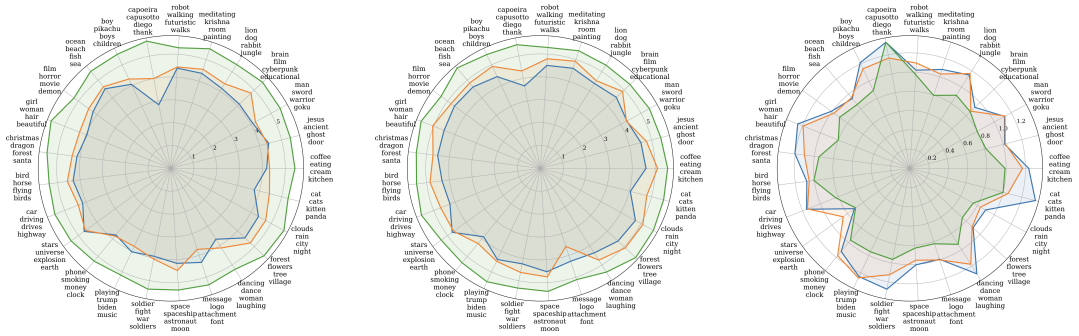
[3] Wenhao Wang and Yi Yang. Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. 2024. 3



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

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

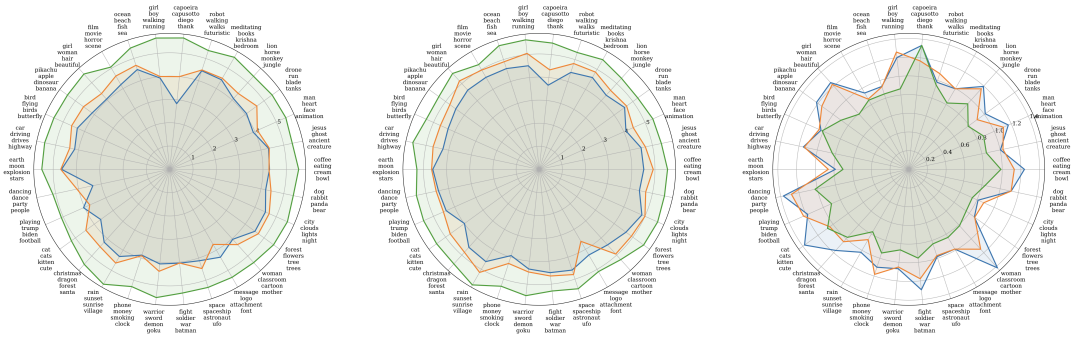
Figure 4. This figure shows results for 15 clusters.



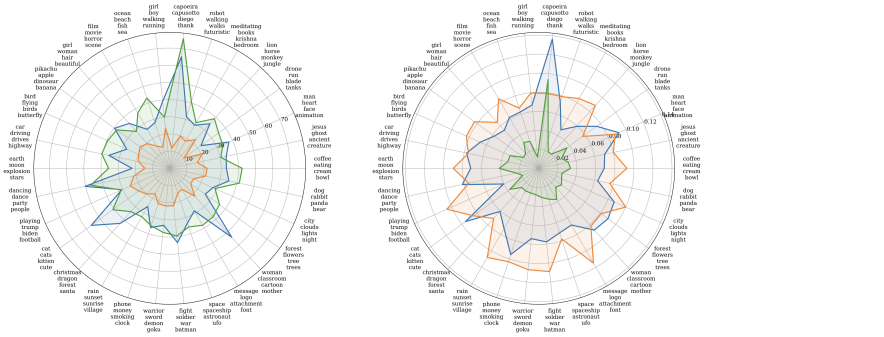
(a) average video rating **(b)** average prompt rating **(c)** average # artifacts

(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.

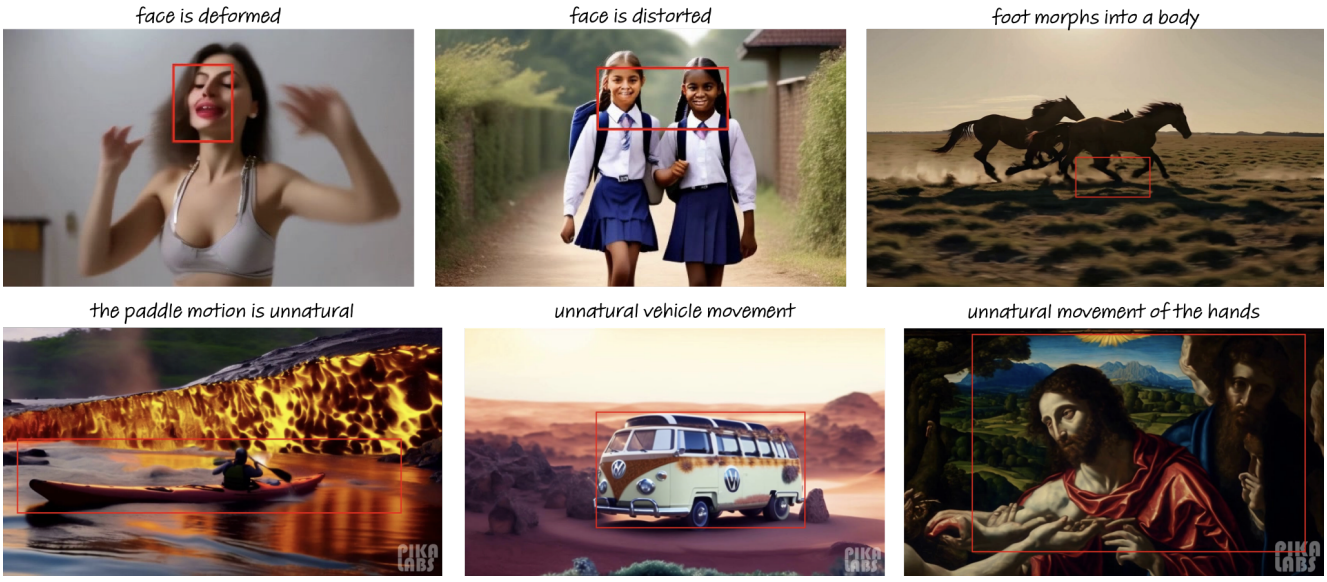


Figure 7. Additional results for the artifact detection model.

"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

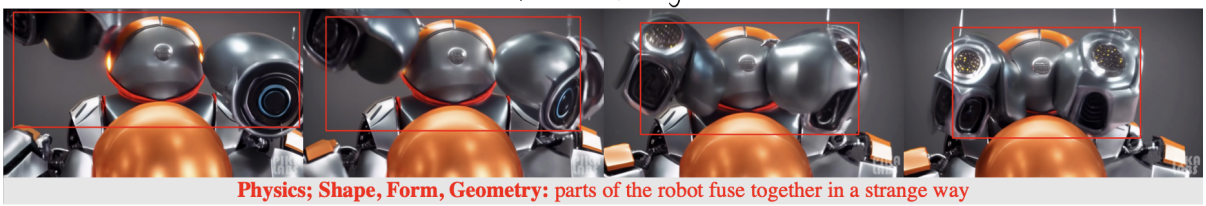
"A cute dog playing with a girl."

Overall: 2/7
Prompt: 6/7

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

"A robot is dancing."

Overall: 2/7
Prompt: 1/7

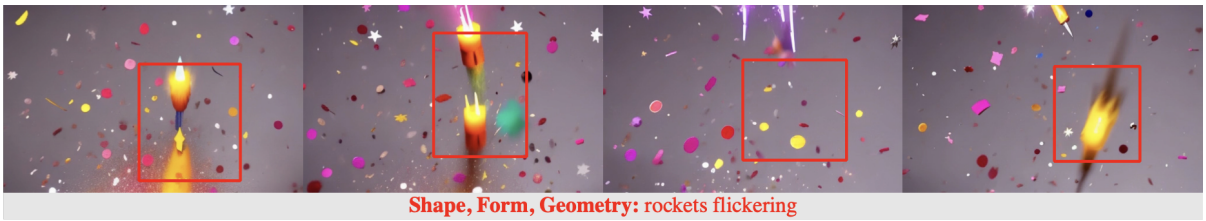


Physics; Shape, Form, Geometry: parts of the robot fuse together in a strange way

Figure 8. Annotations for Pika model.

"Confetti rockets."

Overall: 5/7
Prompt: 7/7



Shape, Form, Geometry: rockets flickering

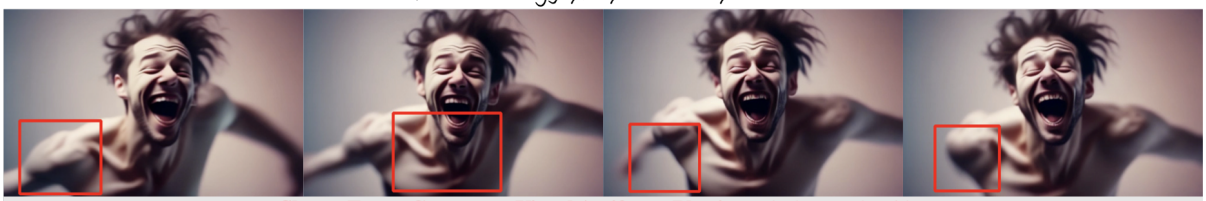
"A time-traveling rabbit wizard stands on a time platform in space, next to a rotating clock."

Overall: 6/7
Prompt: 6/7

Semantic Mismatch with Text: Clock does not rotate

"A man dancing joyfully in a dreamy scene."

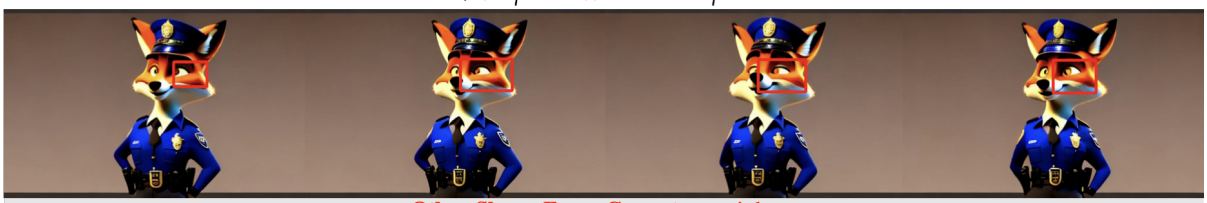
Overall: 3/7
Prompt: 4/7



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."

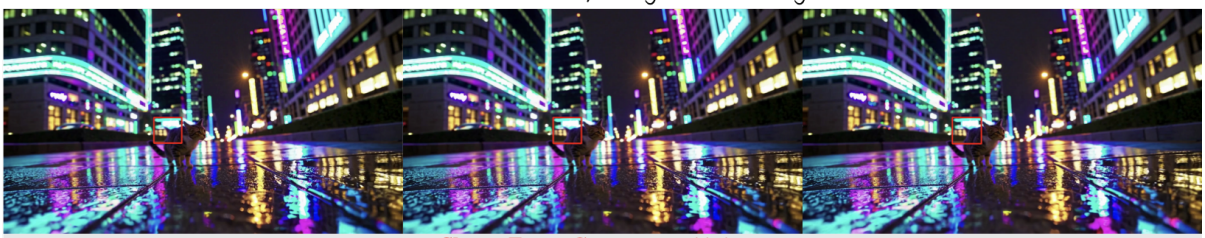
Overall: 4/7
Prompt: 6/7

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."

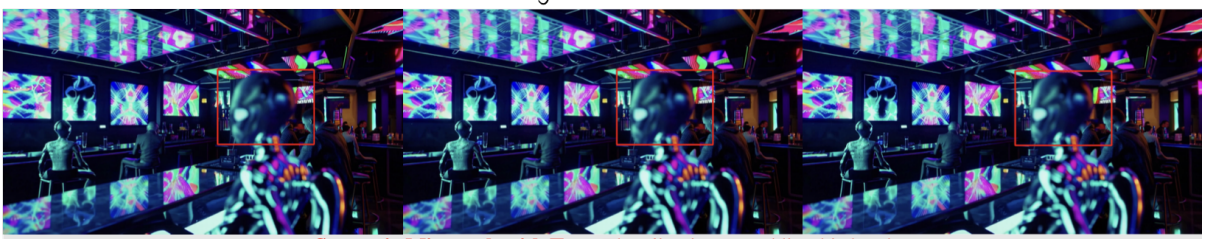
Overall: 7/7
Prompt: 7/7



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

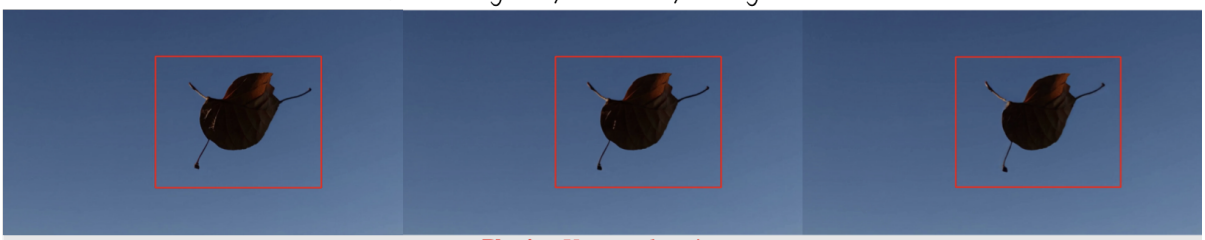
"A technician operating an AOI machine to inspect fabrics."

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.