# 𝕍𝕚𝔹𝕖: A Text-to-Video Benchmark for Evaluating Hallucination in Large Multimodal Models

**Vipula Rawte [1*], Sarthak Jain[2†], Aarush Sinha[3†], Garv Kaushik[4†], Aman Bansal[5†], Prathiksha Rumale Vishwanath[5†], Samyak Rajesh Jain[6], Aishwarya Naresh Reganti[7‡], Vinija Jain[8‡], Aman Chadha[9‡], Amit Sheth[1], Amitava Das[1]**

[1]AI Institute, University of South Carolina, USA
[2]Guru Gobind Singh Indraprastha University, India
[3]Vellore Institute of Technology, India, [4]Indian Institute of Technology (BHU), India
[5]University of Massachusetts Amherst, USA, [6]University of California Santa Cruz, USA
[7]Amazon Web Services, [8]Meta, [9]Amazon GenAI

## Abstract

Recent advances in Large Multimodal Models (LMMs) have expanded their capabilities to video understanding, with Text-to-Video (T2V) models excelling in generating videos from textual prompts. However, they still frequently produce hallucinated content, revealing AI-generated inconsistencies. We introduce 𝕍𝕚𝔹𝕖[*]: a large-scale dataset of hallucinated videos from open-source T2V models. We identify five major hallucination types: Vanishing Subject, Omission Error, Numeric Variability, Subject Dysmorphia, and Visual Incongruity. Using ten T2V models, we generated and manually annotated 3,782 videos from 837 diverse MS COCO captions. Our proposed benchmark includes a dataset of hallucinated videos and a classification framework using video embeddings. 𝕍𝕚𝔹𝕖 serves as a critical resource for evaluating T2V reliability and advancing hallucination detection. We establish classification as a baseline, with the TimeSFormer + CNN ensemble achieving the best performance (0.345 accuracy, 0.342 F1 score). While initial baselines proposed achieve modest accuracy, this highlights the difficulty of automated hallucination detection and the need for improved methods. Our research aims to drive the development of more robust T2V models and evaluate their outputs based on user preferences.

## 1 Introduction

Text-to-video (T2V) models have advanced significantly, enabling the generation of coherent and visually detailed videos from textual prompts. These models have improved in capturing intricate visual elements that align with input text, yet a persistent challenge remains - the generation of hallucinated content. Hallucinations introduce visual discrepancies where elements either misalign with or distort the intended scene, compromising the realism and reliability of T2V outputs. This issue is particularly critical in applications that demand high fidelity to input prompts, such as content creation, education, and simulation systems.

To address this challenge, we introduce 𝕍𝕚𝔹𝕖, a comprehensive large-scale dataset designed to systematically analyze and categorize hallucinations in T2V models. Our dataset was constructed using 837 diverse captions from the MS COCO dataset, which were used to prompt 10 leading open-source T2V models, including HotShot-XL, MagicTime, AnimateDiff-MotionAdapter, and Zeroscope V2 XL. The resulting dataset consists of 3,782 videos, each manually annotated to identify common hallucination types, including disappearing subjects, missing scene components, numerical inconsistencies, and visual distortions.

𝕍𝕚𝔹𝕖 serves as a valuable resource for evaluating the limitations of T2V models and facilitating the development of improved hallucination detection techniques. To complement the dataset, we propose a classification benchmark that leverages video embeddings from TimeSFormer and VideoMAE as inputs for hallucination classification. This benchmark establishes a structured evaluation pipeline, offering baseline performance results and highlighting the challenges of hallucination detection.

In summary, our key contributions are:

- **A large-scale dataset for hallucination analysis in T2V models**: We introduce 𝕍𝕚𝔹𝕖, the first dataset focused on systematically categorizing hallucinations in generated videos. This dataset provides a foundation for studying errors in T2V generation and improving model fidelity.
- **A structured framework for quantifying hallucinations**: We define five major hal-

[*]Corresponding Author
[†]Equal Contribution
[‡]Worked independent of the position
[*]https://vibe-t2v-bench.github.io/

*Proceedings of the 5th Workshop on Trustworthy NLP (TrustNLP 2025)*, pages 232–246
May 3, 2025 ©2025 Association for Computational Linguistics

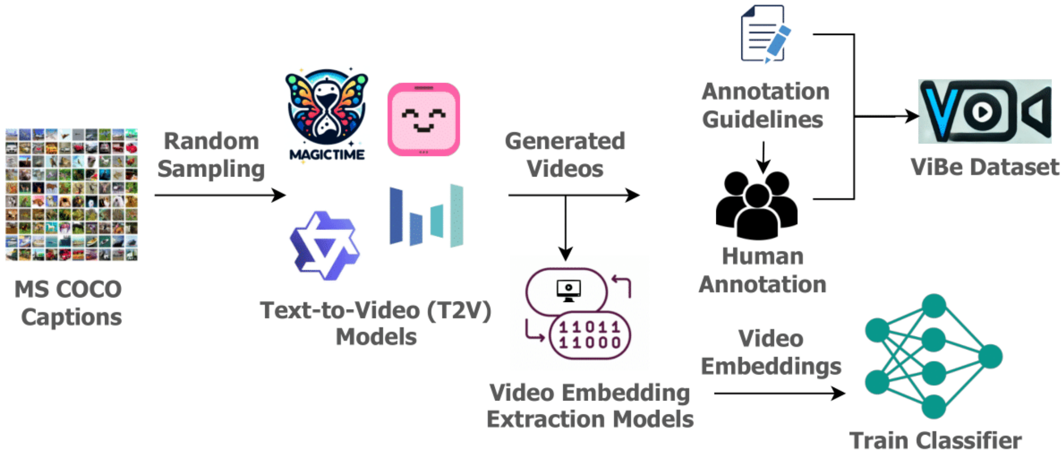


Figure 1: To generate the videos, we utilized randomly sampled image captions from the MS COCO dataset as textual inputs for the video generation models. The resulting videos were then manually annotated by human annotators to construct the ViBe dataset. Following annotation, the videos were processed into feature-rich video embeddings using advanced embedding techniques. These embeddings along with human annotated hallucination labels were subsequently input into various classifier models, which were trained to identify and categorize different types of video hallucinations, enabling the detection of discrepancies between the expected and generated content.

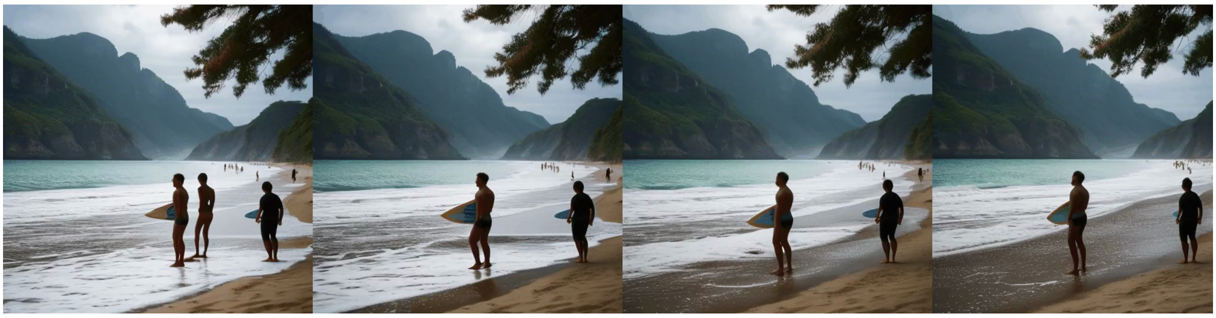


Figure 2: **Prompt:** three guys are standing on a beach next to surfboards. **Vanishing Subject:** The prompt mentions that there are three guys on a beach with surfboards. In the initial frame, we see 3 guys on the beach with surfboards, but in the last frame, we find only two guys remaining. The third guy seems to have *vanished*.

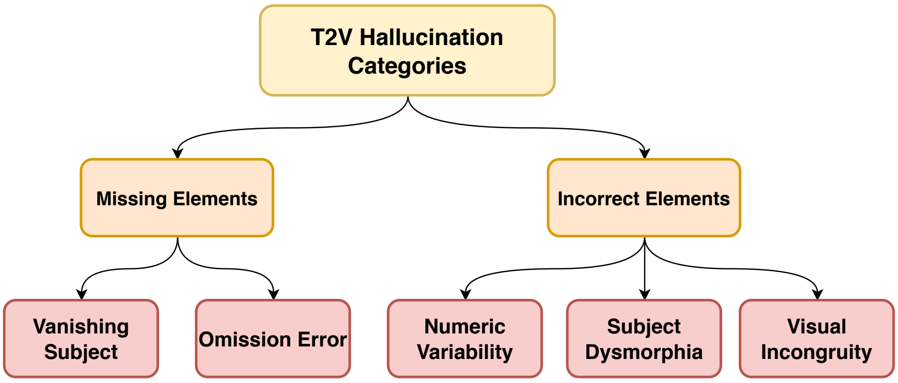


Figure 3: Hierarchy of hallucination categories in ViBe.

lucination categories and provide human-annotated labels, enabling researchers to analyze and mitigate common errors in T2V outputs.

- **A benchmark for hallucination classification**: We propose an evaluation framework using video embeddings and classification models to establish baseline performance on hallucination detection. Our results highlight the difficulty of this task and provide a reference for future improvements.

## 2 Related Work

The phenomenon of hallucination in generative models has been widely studied across different types of media, including text, images, and videos. In text generation, large language models (LLMs) such as GPT-3 (Brown et al., 2020) often produce responses that appear coherent but contain factual inaccuracies. This issue has motivated the development of evaluation benchmarks, such as the Hallucinations Leaderboard (Hong et al., 2024), which aim to measure how frequently and severely these models generate misleading or incorrect content.

In the case of image generation, models like DALL-E (Ramesh et al., 2022) and Imagen (Saharia et al., 2022) have demonstrated impressive abilities in creating high-quality images from textual descriptions. However, these models sometimes generate artifacts that do not align with the provided input text, leading to unrealistic or misleading outputs. To address this problem, datasets such as the HAllucination DEtection dataSet (HADES) (Liu et al., 2022) have been introduced. These datasets provide tools for assessing hallucination in text-to-image models by focusing on specific tokens and offering reference-free evaluation methods.

Video generation models face even greater challenges due to the added complexity of maintaining consistency across multiple frames. Errors in this context can manifest as unrealistic motion, sudden changes in object appearance, or scenes that contradict real-world physics. Recent efforts have aimed to detect and quantify hallucinations in text-to-video models (T2V). The Sora Detector (Chu et al., 2024a) is an example of a framework designed to identify hallucinations in video generation by analyzing key frames and comparing them against knowledge graphs. Similarly, VideoHallucer (Wang et al., 2024b) introduces benchmarks to evaluate hallucinations by distinguishing between errors that originate from the model itself and those that arise due to external inconsistencies. Additionally, VBench (Huang et al., 2024) provides a broad set of evaluation tools to assess the overall quality of generated videos.

Despite these advancements, a major limitation in current research is the lack of a large-scale, human-annotated dataset specifically designed to study hallucinations in text-to-video generation models. 𝕍𝕚𝔹𝕖 addresses this gap by introducing a structured large-scale dataset that categorizes different types of hallucinations observed in generated videos. This dataset includes a diverse collection of human-annotated videos sourced from ten publicly available T2V generative models. By providing detailed annotations, ViBe serves as a valuable resource for developing and testing new methods that aim to detect and reduce hallucinations in text-to-video models.

## 3 Dataset Construction

### 3.1 Dataset Prompt Diversity

To construct the 𝕍𝕚𝔹𝕖 dataset, we carefully selected 837 diverse captions from the MS COCO dataset (Lin et al., 2015), ensuring a balanced representation of real-world scenarios. These captions were used as prompts to generate 3,782 videos, making 𝕍𝕚𝔹𝕖 a valuable resource for evaluating text-to-video (`T2V`) models.

For structured evaluation, the dataset is organized into five distinct thematic categories:

- **Sports**: This category includes prompts describing various athletic activities. An example caption is: *"A baseball hitter stands in position to hit the ball."* These videos capture dynamic motion, human-object interactions, and fast-paced events.

- **Animals**: This category focuses on different species and their behaviors in natural and domestic settings. A sample prompt is: *"Cows strain their necks for hay in between posts of a fence."* These videos challenge models to generate realistic animal motion and interactions with the environment.

- **Objects**: Prompts in this category describe static and dynamic objects in various contexts. For instance, *"Two electrical boxes and signs sit on a street pole."* Evaluating this category helps analyze how well models capture object shapes, textures, and placements.

- **Environment and Settings**: This category includes prompts related to landscapes, weather conditions, and urban or rural scenes. An example caption is: *"Two people in the distance on a beach with surfboards."* This set challenges models to generate coherent spatial layouts and realistic environmental details.

- **Human Activities**: This category involves prompts describing various actions performed by individuals or groups. For example, *"Women are playing WII video games in a big room."* The complexity of human movement, interactions, and physical realism is critical in evaluating these videos.

This structured approach ensures 𝕍𝕚𝔹𝕖 covers diverse real-world scenarios, spanning natural and urban environments, various human activities, and intricate object interactions. It enhances the dataset's utility for evaluating the coherence and fidelity of generated videos while also serving as a foundation for benchmarking improvements in T2V model development.

### 3.2 Models Used for Dataset Creation

We used a subset of 837 captions as input prompts for ten T2V models, representing diverse architectures, sizes, and training paradigms: (i) MS1.7B (ali vilab, 2023), (ii) MagicTime (Yuan et al., 2024a), (iii) AnimateDiff-MotionAdapter (Guo, 2023), (iv) zeroscope_v2_576w (Sterling, 2023a), (v) zeroscope_v2_XL (Sterling, 2023b), (vi) AnimateLCM (Wang et al., 2024a), (vii) HotShotXL (Mullan et al., 2023), (viii) AnimateDiff Lightning (Lin and Yang, 2024), (ix) Show1 (Zhang et al., 2023), and (x) MORA (Yuan et al., 2024b).

Most models generated 1-second videos, except Show1, which produced 2-second videos. Despite their brevity, the hallucination artifacts we define are highly discernible, enabling effective identification and analysis. Table 1 provides a detailed breakdown of video duration across models, highlighting variability in generated outputs.

Videos were systematically analyzed to identify and quantify hallucinations, revealing their widespread occurrence across various open-source T2V systems. Our dataset generation and classification benchmark pipeline are illustrated in Figure 1.

### 3.3 Hallucination Definitions

Hallucination categories were designed based on observed inconsistencies in generated videos rather than technical classifications like those in Sora Detector (Chu et al., 2024a). These inconsistencies broadly fall into subject omissions or incorrect renderings, often exhibiting recurring patterns. We identified five distinct categories, which, while sometimes overlapping, are treated separately due to their frequent occurrence. This framework captures common hallucination patterns in T2V outputs, as detailed in the following section:

| T2V Model | Duration |
|---|---|
| **AnimateLCM** (Wang et al., 2024a) | 1 |
| **zeroscope_v2_XL** (Sterling, 2023b) | 2 |
| **Show1** (Zhang et al., 2023) | 2 |
| **MORA** (Yuan et al., 2024b) | 1 |
| **AnimateDiff Lightning** (Lin and Yang, 2024) | 1 |
| **AnimateDiff-MotionAdapter** (Guo, 2023) | 1 |
| **MagicTime** (Yuan et al., 2024a) | 1 |
| **zeroscope_v2_576w** (Sterling, 2023a) | 2 |
| **MS1.7B** (ali vilab, 2023) | 1 |
| **HotShotXL** (Mullan et al., 2023) | 1 |

Table 1: Video duration per model varies as follows: with the exception of the Show1 and ZeroscopeV2XL model, which generates videos with a duration of 2 seconds, all other models produce videos that are 1 second in length.

1. **Vanishing Subject (VS):** A subject or part of a subject unpredictably disappears during the video. This is often observed in dynamic scenes where subjects fail to persist visually as seen in Figure 2.

2. **Omission Error (OE):** The video fails to render key elements explicitly described in the input prompt as seen in Figure 9.

3. **Numeric Variability (NV):** The video alters the specified number of subjects, either increasing or decreasing their count as seen in Figure 4.

4. **Subject Dysmorphia (SD):** Subjects in the video exhibit unnatural or distorted shapes, scales, or orientation changes, violating expected physical consistency during the course of the video as seen in Figure 8.

5. **Visual Incongruity (VI):** Logically incompatible or physically impossible elements are combined, creating perceptual inconsistencies or violating natural laws as seen in Figure 5.

### 3.4 Human Annotation Details

Table 2 presents the distribution of hallucinated videos across models and categories. Five anno-

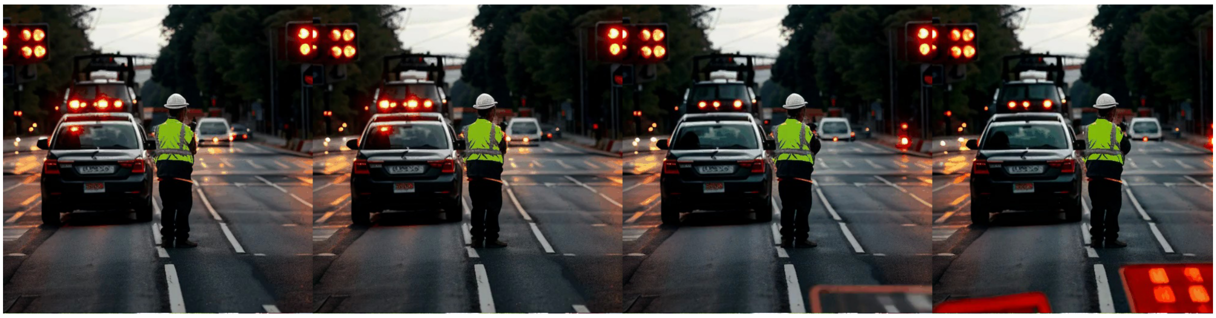


Figure 4: **Prompt:** Two road workers are standing by a red light with a sign. **Numeric Variability:** The prompt explicitly mentions two road workers. However, while the system accurately incorporates elements like the red light and depicts one road worker standing, it fails to generate the second road worker as specified in the prompt. The system modifies the specified number of subjects, decreasing their count, which deviates from the original instructions.

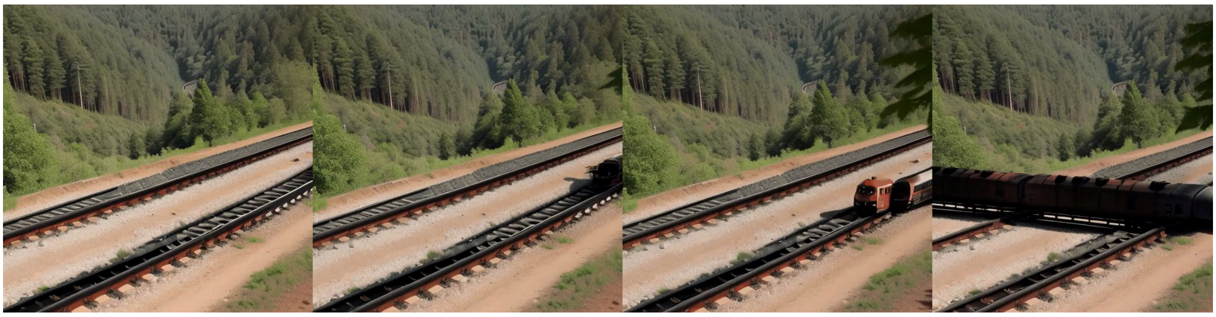


Figure 5: **Prompt:** A train heading for a curve in the track. **Visual Incongruity:** The scenario presents multiple logical and physical impossibilities in its temporal sequence. Initially, no train is visible in the first two frames, violating conservation of mass and the principle of object permanence. In the third frame, the train suddenly materializes on the track without a clear point of origin. In the final frame, the train inexplicably rotates to become perpendicular to the track, an action that defies both the mechanical constraints of train wheels on rails and basic laws of motion. This instantaneous 90-degree rotation would be physically impossible given a train's fixed wheel assembly and its momentum-governed movement along rails.

tators manually categorized 3,782 videos, assigning each to the most prominent hallucination type based on a predefined taxonomy. To ensure consistency, they followed a hierarchical classification approach, prioritizing specific sub-categories before broader ones. Figure 3 visually represents this hierarchy. Additional details on dataset annotation are provided in the appendix A.

### 3.5 Implementation Details

For embedding extraction and classifier training, the process utilized a system with 8 CPU cores, each equipped with 32 GB of memory. This hardware configuration provided the necessary computational resources to efficiently handle data processing and model training. For video generation tasks, an NVIDIA A100 GPU (Jack et al., 2025) was employed, taking advantage of its high-performance capabilities for accelerated computation and rendering of complex video content.

The total duration per model refers to the cumulative time spent annotating all videos associated with that specific model, as shown in 6. 1 provides a detailed report on the video length for each model, allowing for an analysis of how video duration may impact processing times or model performance during annotation tasks.

### 3.6 Inter-Annotator Scores

Two annotators were given 100 common videos to assess inter-annotator agreement, compared against the dataset's gold-standard annotations. Cohen's Kappa scores (Table 3) show the highest agreement for Visual Incongruity (0.8737) and the lowest for Omission Error (0.7474). Cohen's Kappa is calculated as:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{1}$$

where:

- $p_o$ is the observed agreement between the

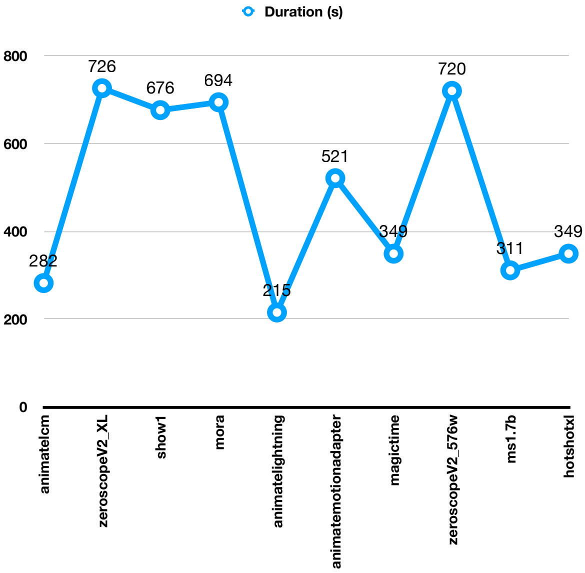


Figure 6: The total duration per model represents the cumulative duration of all videos associated with that model. For instance, **magictime** has a cumulative video duration of 349 seconds. The total duration for **zeroscopeV2_XL** has the longest time, with a duration of 726 seconds, followed by **zeroscopeV2_576w** at 720 seconds. In contrast, the shortest time was recorded for **animatelightning**, which took 215 seconds.

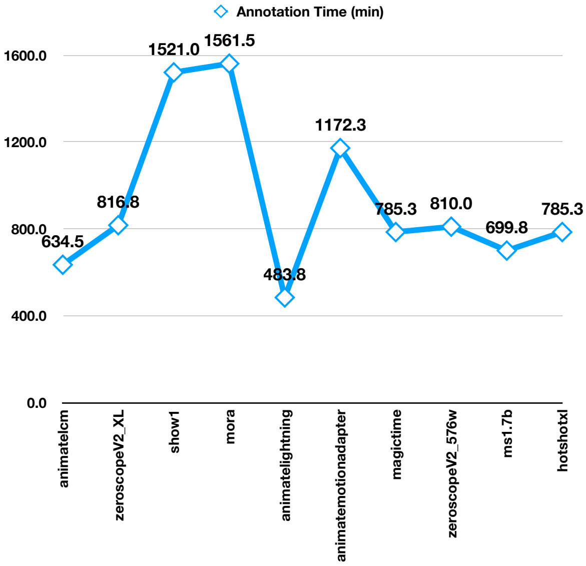


Figure 7: The longest annotation time was recorded for **mora**, taking 1561.5 minutes, followed by **show1** at 1521.0 minutes. Conversely, the shortest annotation time was observed for **animatelightning**, which required 483.75 minutes.

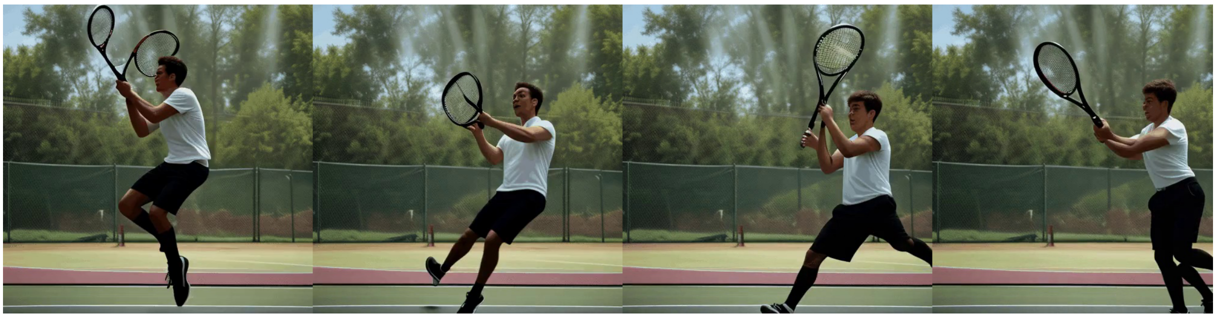


Figure 8: **Prompt:** A man in athletic wear swings a tennis racket through the air. **Subject Dysmorphia:** Throughout the video, both the man and the racket undergo visually inconsistent distortions, resulting in temporal and spatial anomalies. The system-generated artifacts introduce irregularities in the man's form and the racket's structure as they move, causing fluctuations in shape, scale, and position that disrupt the continuity of the intended action.

raters.

- $p_e$ is the expected agreement by chance.

## 4 Classification

Given the growing challenge of video hallucinations, addressing this issue is crucial. Currently, the literature includes only one T2V hallucination benchmark, T2VHaluBench (Chu et al., 2024a), which consists of just 50 videos, limiting its utility for robust evaluation. To overcome this, we propose a large dataset to drive further research, along with several classical classification baselines to support hallucination category prediction. We expect this work to be a key resource for advancing research in this domain.

### 4.1 T2V Hallucination Classification

We evaluate our ViBe dataset using a variety of classification models. We also present a novel task for classifying hallucinations in a text-to-video generation. The first step involves extracting video embeddings from two pre-trained models: VideoMAE (Video Masked Autoencoders for Data-Efficient Pretraining) (Tong et al., 2022) and TimeSFormer (Time-Space Attention Network for Video Un-

| T2V Model | VS | NV | SD | OE | VI | Total |
|---|---|---|---|---|---|---|
| **AnimateLCM** | 2 | 70 | 70 | 70 | 70 | 282 |
| **zeroscope_v2_XL** | 18 | 0 | 37 | 109 | 199 | 363 |
| **Show1** | 13 | 71 | 88 | 111 | 55 | 338 |
| **MORA** | 82 | 96 | 99 | 202 | 215 | 694 |
| **AnimateDiff Lightning** | 11 | 33 | 52 | 56 | 63 | 215 |
| **AnimateDiff-MotionAdapter** | 28 | 59 | 158 | 182 | 94 | 521 |
| **MagicTime** | 70 | 70 | 70 | 69 | 70 | 349 |
| **zeroscope_v2_576w** | 17 | 0 | 41 | 115 | 187 | 360 |
| **MS1.7B** | 51 | 50 | 70 | 70 | 70 | 311 |
| **HotShotXL** | 70 | 70 | 70 | 69 | 70 | 349 |
| **Total** | 362 | 519 | 755 | 1053 | 1093 | 3782 |

Table 2: This table shows the distribution of hallucinated videos produced by ten different text-to-video models, classified into five types of hallucinations. The dataset includes 3,782 videos, each assessed for the occurrence of these hallucination types.

derstanding) (Bertasius et al., 2021). These extracted embeddings are subsequently used as feature representations for seven distinct classification algorithms: Long Short-Term Memory (LSTM) (Sutskever et al., 2014), Transformer (Vaswani et al., 2017), Convolutional Neural Network (CNN) (Krizhevsky et al., 2012), Gated Recurrent Unit (GRU) (Chung et al., 2014), Recurrent Neural Network (RNN) (Mikolov et al., 2010), Random Forest (RF) (Ho, 1995), and Support Vector Machine (SVM) (Cortes and Vapnik, 1995). This comprehensive evaluation across different model architectures allows for a thorough comparison of performance in classifying the given video dataset.

| T2V Hallucination Benchmark | # Videos |
|---|---|
| **T2VHaluBench** (Chu et al., 2024b) | 50 |
| ViBe | 3,782 |

Table 4: The current T2V Hallucination Benchmark, T2VHaluBench, is limited by a small sample size in its dataset. In contrast, our dataset significantly outpaces it, comprising a substantial collection of 3,782 videos, offering a more comprehensive and robust foundation for evaluating T2V hallucination phenomena.

## 4.2 Experimental Setup

The dataset was partitioned into 80% for training and 20% for testing, and the Adam/AdamW optimizer was used (Loshchilov and Hutter, 2019)..

For classification, video embeddings were extracted using the TimeSformer and VideoMAE models, which process individual frames to generate meaningful feature representations. However, despite these models operating on a per-frame basis, the classification task itself did not strictly follow a frame-by-frame approach. Instead, the classification was performed at a higher level, incorporating aggregated representations of the extracted embeddings.

| Hallucination Categories | Cohen's Kappa |
|---|---|
| **Vanishing Subject** | 0.7660 |
| **Omission Error** | 0.7474 |
| **Numeric Variability** | 0.8500 |
| **Subject Dysmorphia** | 0.8173 |
| **Visual Incongruity** | 0.8737 |

Table 3: This table presents Cohen's Kappa Score for Evaluating Inter-Annotator Agreement. The score ranges from **-1 to 1**: **1** represents perfect agreement between annotators. **0** implies that the agreement is no better than random chance. **Negative values** indicate stronger disagreement than expected by chance, suggesting systematic annotation inconsistencies.

| | Hyperparameters | | | |
|---|---|---|---|---|
| **Model** | **# Epochs** | **Batch size** | **Optimizer** | **Loss** |
| **GRU** | 30 | 32 | AdamW | categorical_crossentropy |
| **LSTM** | 120 | 128 | Adam | categorical_crossentropy |
| **Transformer** | 100 | 128 | Adam | categorical_crossentropy |
| **CNN** | 100 | 128 | Adam | categorical_crossentropy |
| **RNN** | 120 | 128 | Adam | categorical_crossentropy |
| **RF** | | | N/A | |
| **SVM** | | | N/A | |

Table 5: Specifications of the model hyperparameters employed during the classifier training process: for both RF and SVM classifiers, default settings from scikit-learn (Pedregosa et al., 2011) were applied.

| Model | Accuracy ↑ | F1 Score ↑ |
|---|---|---|
| **VideoMAE + GRU** | 0.268 | 0.190 |
| **VideoMAE + LSTM** | 0.302 | 0.299 |
| **VideoMAE + Transformer** | 0.284 | 0.254 |
| **VideoMAE + CNN** | 0.303 | 0.290 |
| **VideoMAE + RNN** | 0.289 | 0.289 |
| **VideoMAE + RF** | 0.331 | 0.279 |
| **VideoMAE + SVM** | 0.277 | 0.282 |
| **TimeSFormer + GRU** | 0.325 | 0.279 |
| **TimeSFormer + LSTM** | 0.337 | 0.334 |
| **TimeSFormer + Transformer** | 0.322 | 0.284 |
| **TimeSFormer + CNN** | **0.345** | **0.342** |
| **TimeSFormer + RNN** | 0.299 | 0.299 |
| **TimeSFormer + RF** | 0.341 | 0.282 |
| **TimeSFormer + SVM** | 0.270 | 0.274 |

Table 6: A detailed comparison of model accuracy and F1 score is presented for various combinations of models utilizing VideoMAE and TimeSFormer embeddings. The model yielding the highest performance is denoted in **green** for easy identification. This analysis aims to assess the effectiveness of different embedding strategies in optimizing both classification accuracy and the balance between precision and recall, as captured by the F1 score.

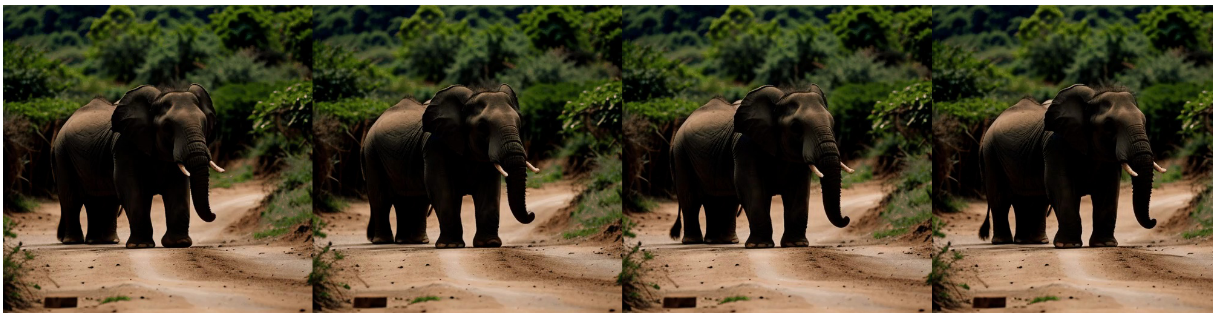


Figure 9: **Prompt:** a baby elephant walking behind a large one **Omission Error:** The generated output fails to render a critical component explicitly specified in the input prompt *the larger one*. While the baby elephant is depicted, the absence of the larger elephant represents a significant deviation from the prompt requirements. This omission fundamentally alters the intended relationship and scale reference that was meant to be portrayed through the presence of both elephants, demonstrating incomplete prompt adherence.

### 4.3 Results and Analysis

Table 6 presents a comprehensive comparison of the performance metrics, namely accuracy and F1 score, for each model across two distinct feature sets: VideoMAE and TimeSFormer embeddings.

For the models trained with VideoMAE embeddings, the RF model demonstrated the highest accuracy, achieving a value of 0.331. However, the LSTM model excelled in the F1 score, recording the highest value of 0.299. On the other hand, the GRU model exhibited the lowest performance, with an accuracy of 0.268 and an F1 score of 0.190, indicating a significant drop in both metrics compared to the other models in this category.

When the TimeSFormer embeddings were utilized, the CNN model outperformed all other models, attaining both the highest accuracy (0.345) and F1 score (0.342). The LSTM model also performed competitively, yielding an accuracy of 0.337 and an F1 score of 0.334. In contrast, the SVM model was the least effective, with an accuracy of 0.270 and an F1 score of 0.274, which were notably lower than those of other models.

Overall, TimeSFormer embeddings consistently outperformed VideoMAE embeddings across most models, showing superior accuracy and F1 scores. The combination of TimeSFormer embeddings with the CNN model delivered the optimal performance in terms of both accuracy and F1 score, making it the most effective configuration in this study.

## 5 Conclusion and Future Work

In this paper, we present ViBe, a large-scale dataset of 3,782 manually annotated videos, surpassing prior benchmarks like T2VHaluBench by 75 times in scale. It provides a robust foundation for evaluating hallucination, ensuring prompt adherence, and improving video generation quality across diverse scenarios across T2V models. We introduce a five-category hallucination taxonomy, enabling systematic analysis and benchmarking of T2V models.

Future research directions encompass several key areas of improvement. First, expanding the existing taxonomy will provide a more comprehensive framework for categorizing and understanding various aspects of video generation. Additionally, evaluating longer-duration videos will help assess the scalability and temporal coherence of the models over extended sequences. Another critical focus is the development of automated classification techniques, which will enhance the efficiency and accuracy of video analysis by reducing reliance on manual annotation. Finally, an essential step forward involves training T2V models using RLHF. This approach aims to refine the alignment of generated videos with human preferences, improving the synthesized content's relevance and quality.

## 6 Limitations

ViBe, while robust, has some limitations. Videos are classified into a single hallucination category for streamlined annotation, which may overlook multi-category overlaps. The dataset is also limited to short video durations due to constraints in open-source T2V models and annotation feasibility. Future work could address these limitations by incorporating multi-category annotations and extending video durations as computational and automatic annotation methods improve.

## 7 Ethics Statement

Our research on the video hallucinations benchmark aims to advance the understanding and evaluation of generative models, ensuring transparency and accountability in their development. We acknowledge the ethical concerns surrounding potential misuse, particularly in creating highly realistic, doctored videos that could contribute to misinformation, fraud, or manipulation. To mitigate these risks, we emphasize responsible disclosure, promote the use of our benchmark for detection and mitigation efforts, and advocate for ethical AI development practices.

## References

ali vilab. 2023. ali-vilab/text-to-video-ms-1.7b · hugging face. https://huggingface.co/ali-vilab/text-to-video-ms-1.7b. (Accessed on 10/28/2024).

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is space-time attention all you need for video understanding? *Preprint*, arXiv:2102.05095.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Zhixuan Chu, Lei Zhang, Yichen Sun, Siqiao Xue, Zhibo Wang, Zhan Qin, and Kui Ren. 2024a. Sora detector: A unified hallucination detection for large text-to-video models. *Preprint*, arXiv:2405.04180.

Zhixuan Chu, Lei Zhang, Yichen Sun, Siqiao Xue, Zhibo Wang, Zhan Qin, and Kui Ren. 2024b. Sora detector: A unified hallucination detection for large text-to-video models. *arXiv preprint arXiv:2405.04180*.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. Cite arxiv:1412.3555Comment: Presented in NIPS 2014 Deep Learning and Representation Learning Workshop.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Yuwei Guo. 2023. guoyww/animatediff-motion-adapter-v1-5-2 · hugging face. https://huggingface.co/guoyww/animatediff-motion-adapter-v1-5-2. (Accessed on 10/28/2024).

Tin Kam Ho. 1995. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE.

Giwon Hong, Aryo Pradipta Gema, Rohit Saxena, Xiaotang Du, Ping Nie, Yu Zhao, Laura Perez-Beltrachini, Max Ryabinin, Xuanli He, Clémentine Fourrier, and Pasquale Minervini. 2024. The hallucinations leaderboard - an open effort to measure hallucinations in large language models. *CoRR*, abs/2404.05904.

Ziqi Huang, Yinan He, Jiashuo Yu, Fan Zhang, Chenyang Si, Yuming Jiang, Yuanhan Zhang, Tianxing Wu, Qingyang Jin, Nattapol Chanpaisit, Yaohui Wang, Xinyuan Chen, Limin Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. 2024. VBench: Comprehensive benchmark suite for video generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Wishwesh Choquette Jack, Gandhi Olivier, Giroux Nick, Stam Ronny, and Krashinsky. 2025. Ieee xplore full-text pdf. https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9361255&tag=1. [Online; accessed 2025-02-06].

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.

Shanchuan Lin and Xiao Yang. 2024. Animatediff-lightning: Cross-model diffusion distillation. *Preprint*, arXiv:2403.12706.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context. *Preprint*, arXiv:1405.0312.

Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2022. A token-level reference-free hallucination detection benchmark for free-form text generation. *Preprint*, arXiv:2104.08704.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černocký, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Interspeech 2010*, pages 1045–1048.

John Mullan, Duncan Crawbuck, and Aakash Sastry. 2023. Hotshot-XL.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *Preprint*, arXiv:2204.06125.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Preprint*, arXiv:2205.11487.

Spencer Sterling. 2023a. cerspense/zeroscope_v2_576w · hugging face. https://huggingface.co/cerspense/zeroscope_v2_576w. (Accessed on 10/28/2024).

Spencer Sterling. 2023b. cerspense/zeroscope_v2_xl · hugging face. https://huggingface.co/cerspense/zeroscope_v2_XL. (Accessed on 10/28/2024).

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. 2022. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Preprint*, arXiv:2203.12602.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Fu-Yun Wang, Zhaoyang Huang, Weikang Bian, Xiaoyu Shi, Keqiang Sun, Guanglu Song, Yu Liu, and Hongsheng Li. 2024a. Animatelcm: Computation-efficient personalized style video generation without personalized video data. *Preprint*, arXiv:2402.00769.

Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. 2024b. Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *Preprint*, arXiv:2406.16338.

Shenghai Yuan, Jinfa Huang, Yujun Shi, Yongqi Xu, Ruijie Zhu, Bin Lin, Xinhua Cheng, Li Yuan, and Jiebo Luo. 2024a. Magictime: Time-lapse video generation models as metamorphic simulators. *Preprint*, arXiv:2404.05014.

Zhengqing Yuan, Yixin Liu, Yihan Cao, Weixiang Sun, Haolong Jia, Ruoxi Chen, Zhaoxu Li, Bin Lin, Li Yuan, Lifang He, Chi Wang, Yanfang Ye, and Lichao Sun. 2024b. Mora: Enabling generalist video generation via a multi-agent framework. *Preprint*, arXiv:2403.13248.

David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. 2023. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *Preprint*, arXiv:2309.15818.

## A Appendix

This section offers supplementary material, including additional examples, implementation details, and more, to enhance the reader's understanding of the concepts discussed in this work. We also present additional details of the annotation process in Section B.

## B Annotation Details

The objective of this annotation task is to detect and classify hallucinations in videos produced by `T2V` models. The annotated data will be utilized to assess the model's adherence to input prompts and contribute to improving hallucination detection and mitigation.

1 **Understanding Hallucination Categories** Annotators will be trained to recognize the five predefined categories of `T2V` hallucination: `Vanishing Subject`, `Omission Error`, `Numeric Variability`, `Subject Dysmorphia`, = and `Visual Incongruity`.

2 **Training and Evaluation Protocol**

a. **Training:** Annotators will receive example videos for each hallucination category, along with justifications for category assignments.

b. **Evaluation:** Annotators will classify five test videos, each corresponding to a unique hallucination category. A minimum agreement score of 60% (correct classification of at least 3 out of 5 videos) is required to proceed to the annotation phase.

c. **Feedback Loop:** Annotators who do not meet the agreement threshold will receive targeted feedback and additional training.

3 **Annotation Process**

a. **Video Evaluation:** Annotators will carefully review the assigned video, comparing the visual content to the input text prompt to identify inconsistencies.

b. **Hallucination Categorization:** Annotators will assign the most prominent hallucination category. If multiple hallucinations exist, the most visibly apparent one will be selected.

c. **Annotation Tool:** The identified category will be entered into the annotation tool (see 10, 11). Supplementary notes can be added for clarification.

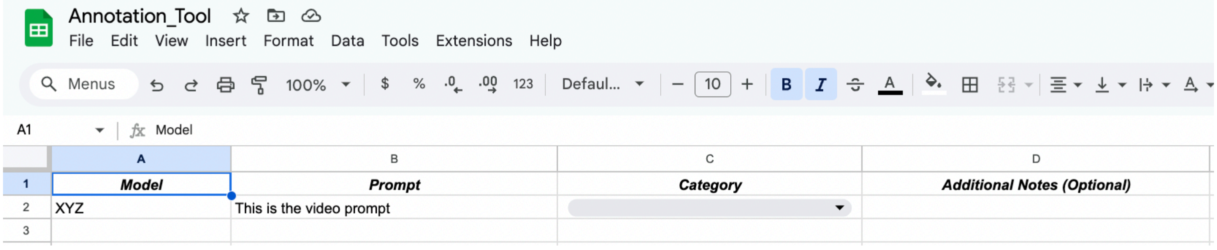
d. **Annotation Time:** The average annotation time was recorded at 2.25 seconds per video (see 7).

## C Dataset

The five categories of hallucination have been previously defined, with examples provided for each. In this section, we will present additional examples to further illustrate these categories.

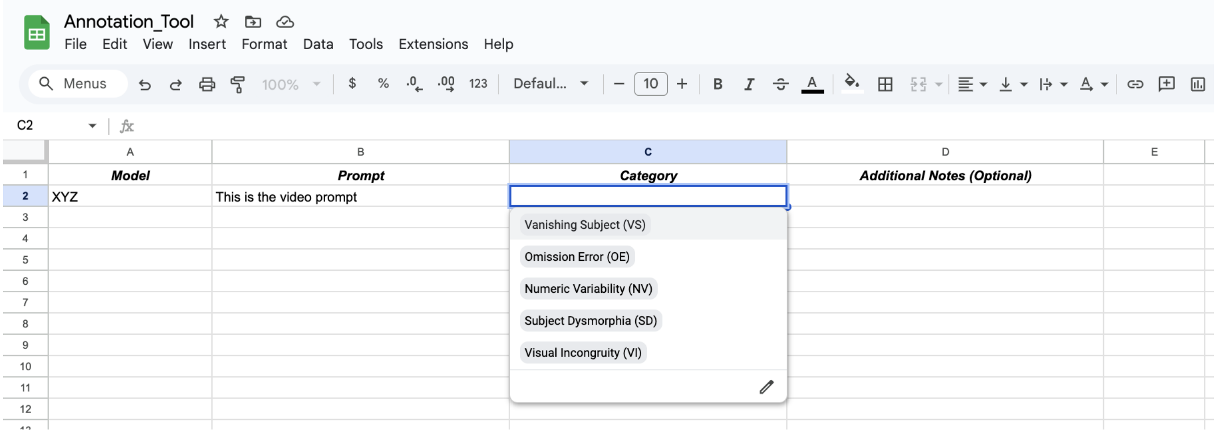
### C.1 Hallucination Categories

1. **Vanishing Subject (VS):** See figs. 12 and 13
2. **Omission Error (OE):** See figs. 14 and 15
3. **Numeric Variability (NV):** See figs. 16 and 17
4. **Subject Dysmorphia (SD):** See figs. 18 and 19
5. **Visual Incongruity (VI):** See figs. 20 and 21



| | A | B | C | D |
|---|---|---|---|---|
| 1 | *Model* | *Prompt* | *Category* | *Additional Notes (Optional)* |
| 2 | XYZ | This is the video prompt | | |
| 3 | | | | |

Figure 10: This figure illustrates the annotation tool employed to label various video samples. The tool comprises four columns:
**Model:** Represents the specific T2V model.
**Prompt:** Contains the image caption text derived from the MS COCO dataset.
**Category:** Indicates one of the five predefined hallucination categories.
**Additional Notes:** An optional column for supplementary annotations.



| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | *Model* | *Prompt* | *Category* | *Additional Notes (Optional)* | |
| 2 | XYZ | This is the video prompt | | | |

- Vanishing Subject (VS)
- Omission Error (OE)
- Numeric Variability (NV)
- Subject Dysmorphia (SD)
- Visual Incongruity (VI)

Figure 11: Using this annotation tool, annotators can classify the generated videos into one of the five predefined hallucination categories.

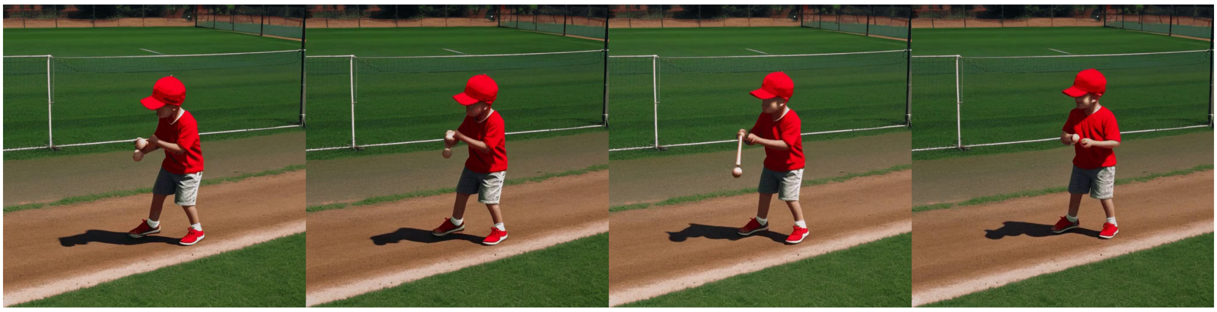


Figure 12: **Prompt:** A boy in a red hat playing with tee ball set. **Vanishing Subject:** The visual content depicts a boy wearing a red hat engaged in play with a tee-ball set. However, a hallucination occurs within the generated scene, where the tee-ball set, initially present, inexplicably disappears during the sequence.

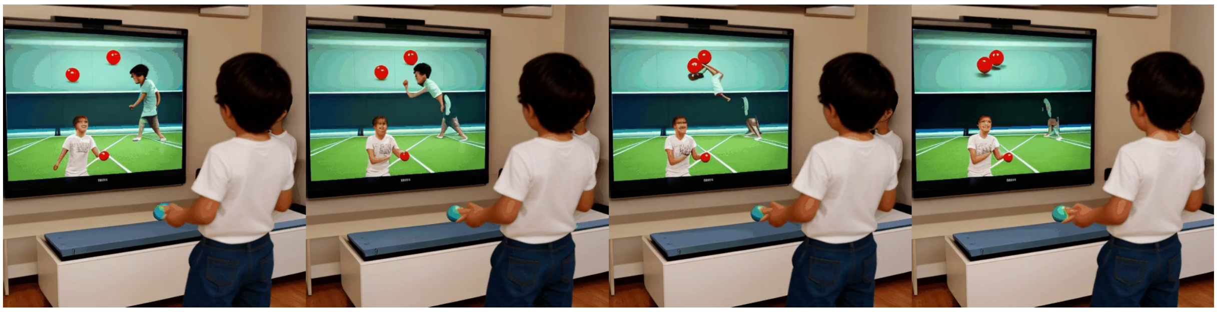


Figure 13: **Prompt:** Two young boys playing Wii bowling on a large television screen **Vanishing Subject:** In the video frames, the TV initially displays two boys. However, as the video progresses, subtle changes occur. By the final frame, one of the boys on the TV has mysteriously vanished, leaving only the other behind.

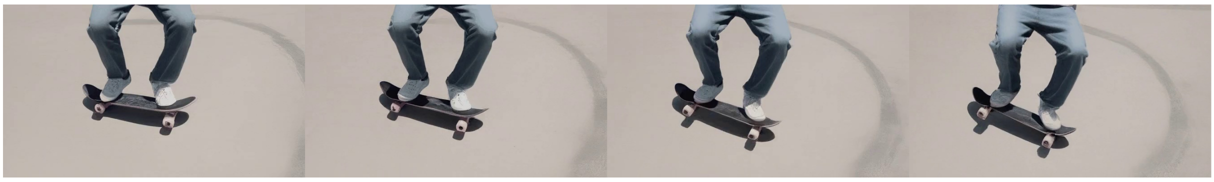


Figure 14: **Prompt:** A person on a skateboard with his arms in the air. **Omission Error:** The prompt describes a scene featuring a person on a skateboard with their arms raised in the air. However, this description exhibits a hallucination, as the video does not depict the individual's arms at all.

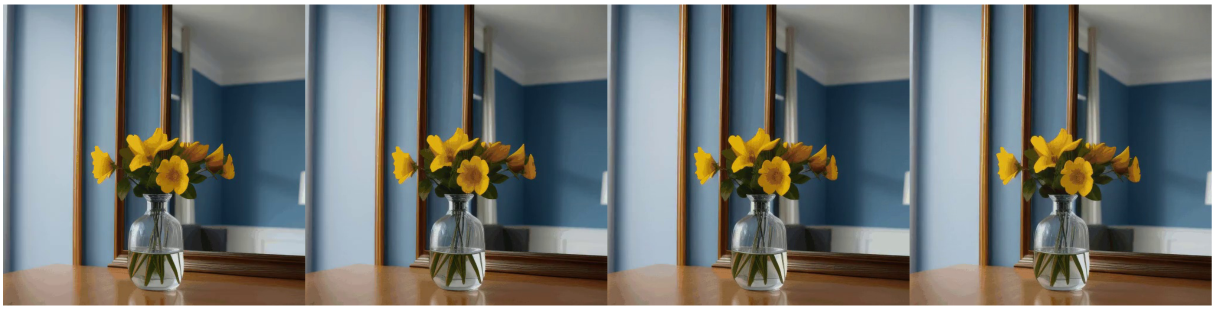


Figure 15: **Prompt:** Blue and yellow flowers in a glass vase near a mirror. **Omission Error:** The video lacks any blue flowers, despite their explicit mention in the prompt. This discrepancy highlights a failure of the model to accurately represent key visual elements specified in the input.

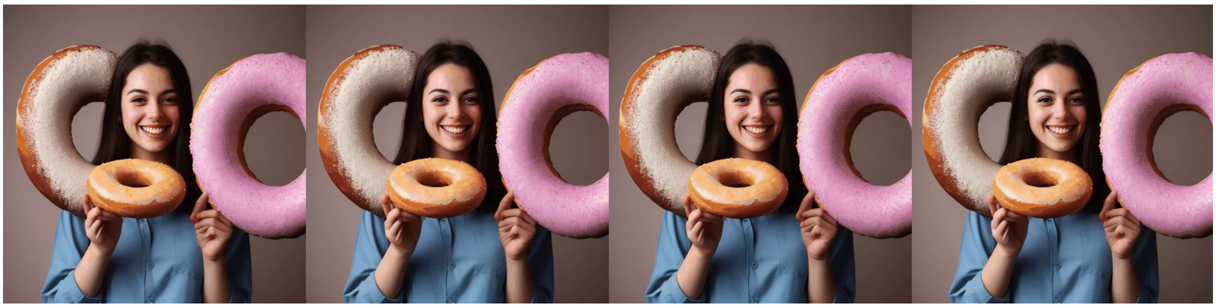


Figure 16: **Prompt:** A happy adult holding two large donuts. **Numeric Variability:** The description depicts a content scenario where a happy adult is holding two large donuts. However, a hallucination occurs within the video, where the depicted woman is shown holding three donuts instead of two.

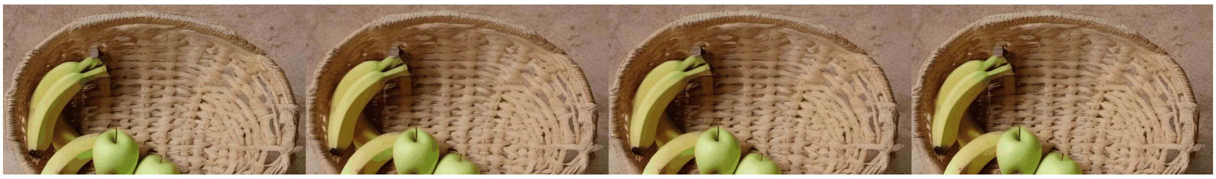


Figure 17: **Prompt:** A banana and a yellow apple in a woven basket. **Numeric Variability:** The visual scene consists of a woven basket containing one banana and one yellow apple. However, the generative output exhibits a hallucination, inaccurately depicting two bananas and two apples within the basket.

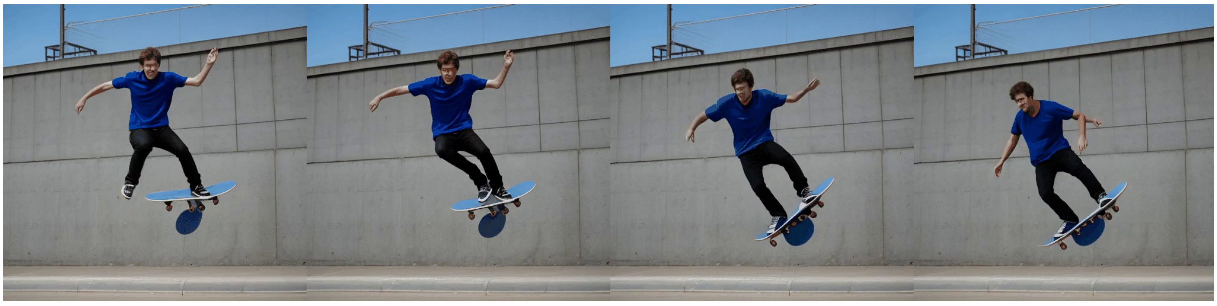


Figure 18: **Prompt:** Skateboarder and blue shirt and black jeans jumping on his board **Subject Dysmorphia:** The video depicts a person riding a skateboard. Throughout the frames, the wheels of the skateboard keep morphing, fluctuating in number as they increase and decrease. Additionally, the skateboarder's arms undergo a similar distortion, gradually shifting in shape over time.

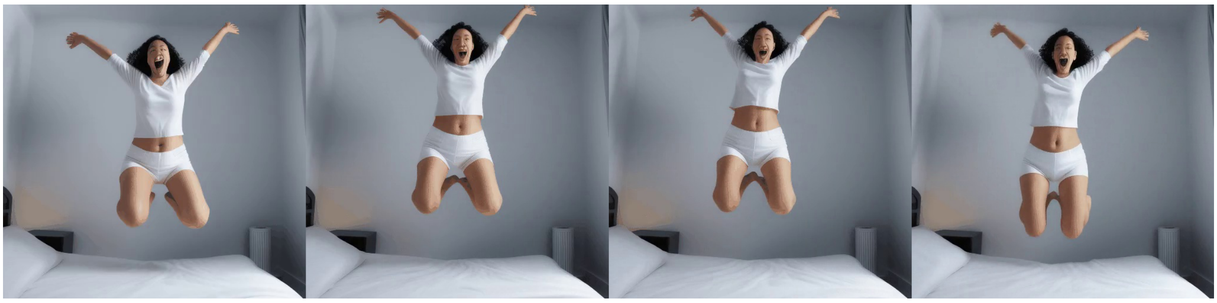


Figure 19: **Prompt:** A woman is jumping on a white bed. **Subject Dysmorphia:** The video depicts a woman jumping on a white bed. Over time, a hallucination effect manifests, leading to a dysmorphic transformation of the woman's face within the video.

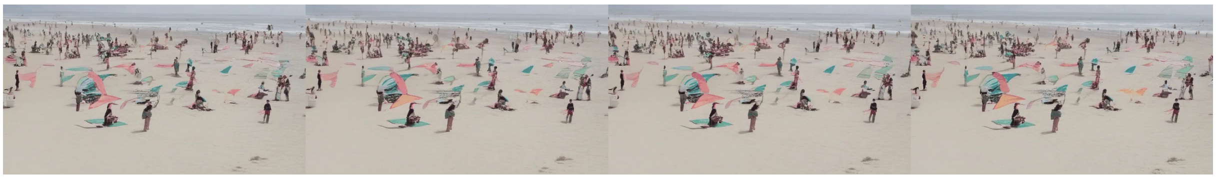


Figure 20: **Prompt:** A crowd of people standing on a beach flying kites. **Visual Incongruity:** Instead of being depicted in the sky as expected, the kites appear visually inconsistent, resembling objects embedded in the sand.

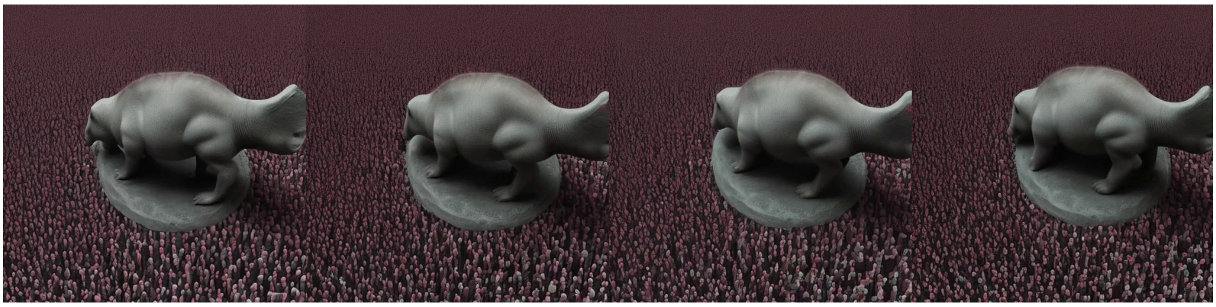


Figure 21: **Prompt:** a animal that is walking in a crowd of people **Visual Incongruity:** In the generated video, a stone statue of an animal is seen moving atop a vast crowd that appears to be composed of human heads. The statue's movement contrasts with its rigid, lifeless material, creating an unsettling effect. The generated video blurs the line between the inanimate and the living.