

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

Vipula Rawte<sup>1\*</sup>, Sarthak Jain<sup>2†</sup>, Aarush Sinha<sup>3†</sup>, Gav Kaushik<sup>4†</sup>, Aman Bansal<sup>5†</sup>,  
 Prathiksha Rumale Vishwanath<sup>5†</sup>, Samyak Rajesh Jain<sup>6</sup>, Aishwarya Naresh Reganti<sup>7§</sup>,  
 Vinija Jain<sup>8§</sup>, Aman Chadha<sup>9§</sup>, Amit Sheth<sup>1</sup>, Amitava Das<sup>1</sup>

<sup>1</sup>AI Institute, University of South Carolina, USA

<sup>2</sup>Guru Gobind Singh Indraprastha University, India

<sup>3</sup>Vellore Institute of Technology, India

<sup>4</sup>Indian Institute of Technology (BHU), India

<sup>5</sup>University of Massachusetts Amherst, USA

<sup>6</sup>University of California, Santa Cruz, USA

<sup>7</sup>Amazon Web Services, USA

<sup>8</sup>Meta, USA, <sup>9</sup>Amazon GenAI, USA

{vrawte}@mailbox.sc.edu

<https://vibe-t2v-bench.github.io/>

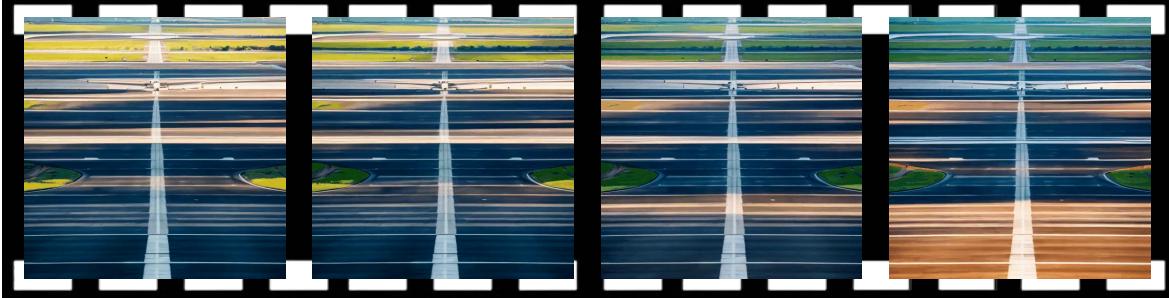


Figure 1. **Prompt:** An airport runway with a large plane and cars parked on one side. **Hallucination:** In the output video, the runway appears devoid of any aircraft or cars on the side, resulting in the absence of the “plane” and “cars” elements described in the prompt. This suggests that the model has failed to incorporate the requested objects - namely, the “large plane” and “cars” into the generated content, thus omitting a critical part of the input specification. We categorize this type of hallucination as an **Omission Error**.

## Abstract

Latest developments in Large Multimodal Models (LMMs) have broadened their capabilities to include video understanding. Specifically, Text-to-video (T2V) models have made significant progress in quality, comprehension, and duration, excelling at creating videos from simple textual prompts. Yet, they still frequently produce hallucinated content that clearly signals the video is AI-generated. We

introduce **ViBe**: a large-scale Text-to-Video Benchmark of hallucinated videos from T2V models. We identify five major types of hallucination: Vanishing Subject, Numeric Variability, Temporal Dysmorphia, Omission Error, and Physical Incongruity. Using 10 open-source T2V models, we developed the first large-scale dataset of hallucinated videos, comprising 3,782 videos annotated by humans into these five categories. The dataset was created by prompting T2V models with MS COCO captions and manually categorizing the results by hallucination type. **ViBe** offers a unique resource for evaluating the reliability of T2V models and provides a foundation

\*Corresponding author.

†Equal contribution.

§Work independent of the position.

*for improving hallucination detection and mitigation in video generation. We establish classification as a baseline and present various ensemble classifier configurations, with the TimeSFormer + CNN combination yielding the best performance, achieving 0.345 accuracy and 0.342 F1 score. This benchmark aims to drive the development of robust T2V models that produce videos more accurately aligned with input prompts.*

## 1. Introduction

Text-to-video models have made significant strides in recent years, enabling the generation of video content from textual prompts with impressive coherence and visual fidelity. These models have progressively improved in producing high-quality videos that effectively capture intricate visual details, corresponding to the semantics of the input text. However, despite these advancements, one of the most pressing challenges in the field remains the generation of hallucinated content—visual elements that either misalign with or distort the intended scene described by the text prompt. Hallucinations compromise the realism and dependability of T2V outputs, posing a critical issue for applications where precise adherence to the input text is paramount, such as in content creation, education, or simulation systems.

In response to this problem, we introduce **ViBe**, a comprehensive large-scale dataset aimed at systematically investigating and categorizing hallucinations within T2V models. This dataset was developed by curating 700 randomly selected captions from the MS-COCO dataset and utilizing them to prompt ten leading open-source T2V models, including MS1.7B, MagicTime, AnimateDiff-MotionAdapter, and Zeroscope V2 XL. The resulting dataset consists of 3,782 videos, each human-annotated to identify various types of hallucinations frequently encountered in T2V generation. These include errors such as omission of key scene components, discrepancies in subject count, temporal inconsistencies, physical incongruities, and subjects that vanish unexpectedly.

**ViBe** serves as a valuable resource for assessing and advancing hallucination detection in T2V models. The dataset is meticulously annotated to enable detailed analysis, providing researchers with the tools to evaluate the limitations of current T2V systems and explore methodologies for reducing these errors. By offering a standardized framework for categorizing hallucinations and establishing benchmarks, **ViBe** paves the way for the development of more accurate and reliable T2V models that better reflect the intended semantic content of the input text.

In summary, our key contributions are:

- Introducing **ViBe**, a novel benchmark for assessing Text-to-Visual hallucination phenomena. This bench-

mark is designed to rigorously evaluate the ability of models to generate visual content from textual input, specifically focusing on the accuracy, consistency, and fidelity of the generated visuals in relation to the provided textual descriptions (cf. Sec. 3).

- By providing a standardized framework for quantifying hallucinations—instances where the generated visuals deviate from or misrepresent the input text—**ViBe** aims to advance the understanding and mitigation of errors in T2V models, fostering improvements in their alignment and reliability (cf. Sec. 3).
- Conducting a comprehensive benchmark evaluation of various classification models, assessing their performance across key metrics such as accuracy and F1-score (cf. Sec. 4).

## 2. Related Work

The phenomenon of hallucination in generative models has been extensively examined across a variety of modalities, including text, images, and videos [22]. In the domain of text generation, LLMs such as GPT-3 [3] have demonstrated the ability to generate content that, although syntactically plausible, may lack factual accuracy or exhibit inconsistencies with the input prompt. This issue of hallucination has been systematically addressed through the development of specialized benchmarks, including the Hallucinations Leaderboard [11], which provides a framework for evaluating LLMs on tasks that involve hallucinated content.

**Image generation:** Text-to-image models such as DALL-E [21] and Imagen [23] have showcased advanced capabilities in producing highly realistic images based on textual descriptions. Nevertheless, these models are not immune to generating artifacts or producing visual elements that are inconsistent with the input description. To address this issue, datasets like the Hallucination Detection dataset (HADES) [15] have been introduced, providing benchmarks for token-level, reference-free hallucination detection in free-form text-to-image generation.

**Video generation:** The challenge of hallucination becomes more complex in video generation, where temporal consistency must be maintained across a sequence of frames. Recent advancements in the field have sought to mitigate this issue. For instance, the Sora Detector [5] presents a unified framework for detecting hallucinations in large T2V models. This approach incorporates techniques such as keyframe extraction and knowledge graph construction to identify inconsistencies both within individual frames and across the temporal dimension of a video sequence. Additionally, the VideoHallucer benchmark [30] provides a detailed evaluation of hallucinations in video-to-text models by categorizing them into various types, such as object-relation, temporal, semantic detail, extrinsic factual, and extrinsic non-factual hallucinations.

Despite these advancements, a significant gap remains in the availability of large-scale, human-annotated datasets specifically focused on hallucinations in T2V models. The **ViBe** dataset is designed to address this gap by offering a comprehensive resource for the systematic study and evaluation of hallucinations in T2V models. By categorizing hallucinations into distinct types and providing a substantial volume of annotated video data, **ViBe** serves as a critical benchmark for the development and evaluation of methods aimed at detecting and mitigating hallucinations in T2V models.

While notable progress has been made in understanding and mitigating hallucinations across different modalities, the **ViBe** dataset represents a crucial step forward in the specific context of T2V models. It equips researchers and practitioners with the necessary tools to develop more accurate and reliable video generation systems, ultimately improving the fidelity and applicability of T2V technologies.

### 3. Dataset

#### 3.1. Dataset construction

To construct the **ViBe** dataset, we selected 700 random captions from the MS COCO dataset [14], which is known for its diverse and descriptive textual prompts, making it an ideal resource for evaluating the generative performance of T2V models. These captions were then used as input for **ten** distinct open-source T2V models, chosen to represent a variety of architectures, model sizes, and training paradigms. The specific models included in the study were: (i) MS1.7B [1], (ii) MagicTime [33], (iii) AnimateDiff-MotionAdapter [9], (iv) zeroscope\_v2.576w [24], (v) zeroscope\_v2\_XL [25], (vi) AnimateLCM [29], (vii) HotShotXL [19], (viii) AnimateDiff Lightning [13], (ix) Show1 [35], and (x) MORA [34].

These models generated video outputs based on the MS COCO captions, which were systematically analyzed to identify the presence and frequency of hallucinations. In addition to these open-source models, we also produced approximately 40-50 videos using two closed-source, state-of-the-art models: Runway [8] and Luma [17]. The generated videos from both open and closed-source models were rigorously examined to highlight examples of hallucinations, further underscoring the prevalence of such artifacts across both model categories. This analysis provides evidence that hallucinations are widespread across a diverse range of T2V systems, regardless of whether they are open-source or closed-source. This pipeline is described in Fig. 3.

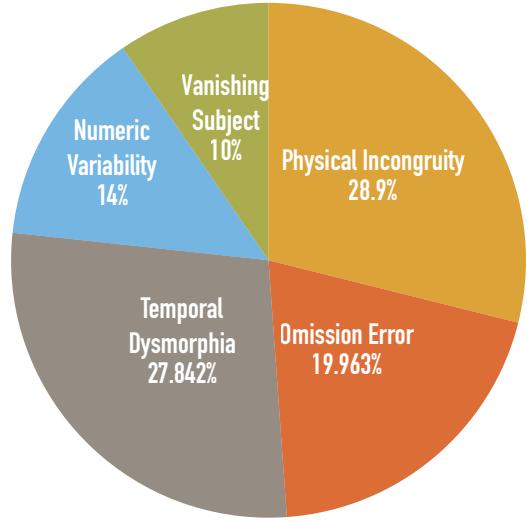


Figure 2. Dataset statistics for various categories of T2V hallucinations. The five categories include: Vanishing Subject, Numeric Variability, Temporal Dysmorphia, Omission Error, and Physical Incongruity. The numbers represent the distribution of hallucinated videos within each respective category.

#### 3.2. Hallucination Categories

To systematically categorize the various types of hallucinations observed, we established five distinct categories in Fig. 2 that collectively encompass the majority of the common hallucinations present in T2V outputs.

- Vanishing Subject (VS):** The subject, or a portion thereof, in the generated video intermittently disappears at arbitrary points within the video's duration (see Fig. 4).
- Numeric Variability (NV):** In a given prompt, if the subject count is specified, the generated video either increases or decreases the number of instances of the subject (see Fig. 5).
- Temporal Dysmorphia (TD):** Objects rendered within the video exhibit continuous temporal deformation, undergoing gradual or intermittent transformations in shape, scale, or orientation over the duration of the sequence (see Fig. 6).
- Omission Error (OE):** The generated video omits essential components of the initial prompt (see Fig. 7) - except in cases involving specified subject counts—resulting in an incomplete or inaccurate portrayal, or introduces unscripted actions or behaviors, it leads to a misrepresentation of the intended scene.
- Physical Incongruity (PI):** The generated video violates fundamental physical laws or juxtaposes incongruent elements (see Fig. 8), leading to perceptual inconsistencies or cognitive dissonance for the viewer.

Physical Incongruity and Temporal Dysmorphia

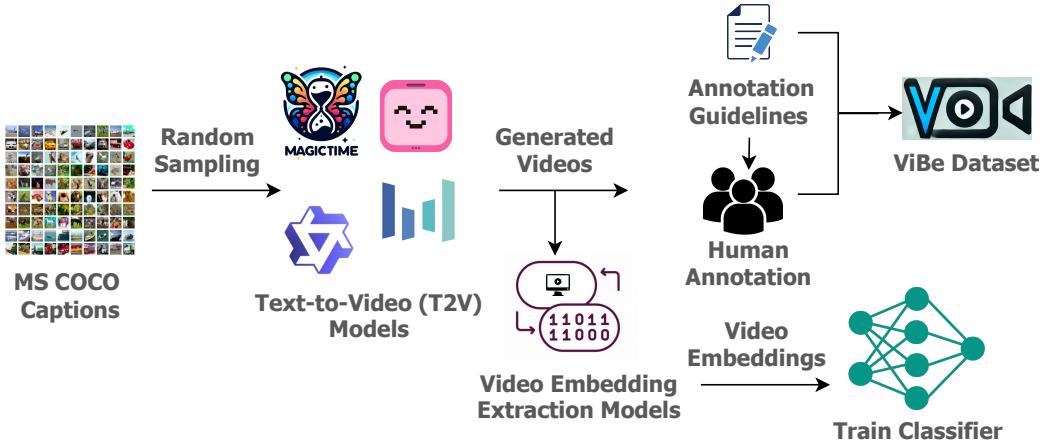


Figure 3. To generate the videos, we utilized randomly sampled image captions from the MS COCO dataset as textual inputs for the video generation models. The resulting videos were then manually annotated by human annotators to construct the **ViBe** dataset. Following annotation, the videos were processed 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 4. **Prompt:** A man scooping food into a pan. **Vanishing Subject:** A man is observed transferring food into a pan, but upon closer analysis, a visual anomaly occurs towards the end of the sequence. The food in the man's hand is not consistently rendered as it nears the pan, effectively disappearing from the visual frame. This anomaly is indicative of hallucination artifacts, where the model fails to maintain object permanence and continuity in a spatial-temporal context, resulting in the disappearance of the object before the action is completed.



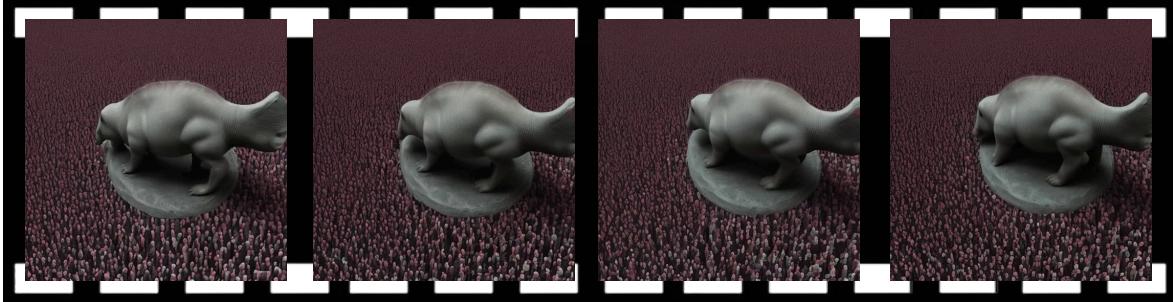
Figure 5. **Prompt:** Six people with their hands on an outdoor kiln. **Numeric Variability:** The prompt specifies an outdoor scene with six individuals interacting with a kiln. However, in the generated content, only two individuals are depicted. This is an instance of hallucination, where the model deviates from the original input, misrepresenting critical details, such as the specified number of participants.



**Figure 6. Prompt:** A man in athletic wear swings a tennis racket through the air. **Temporal 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.



**Figure 7. Prompt:** A baby elephant walking behind a large one. **Omission Error:** The original prompt describes a scene with a baby elephant walking behind a larger elephant. However, the output fails to accurately reflect this description, as no actual representation of a baby elephant appears in the scene. This discrepancy exemplifies a form of hallucination, where the model either misinterprets or fails to include certain elements specified in the input prompt. In this instance, although the prompt references a baby elephant, it is absent from the generated output, leading to the omission of an essential component.



**Figure 8. Prompt:** An animal that is walking in a crowd of people. **Physical Incongruity:** Instead of depicting an animal naturally moving within the crowd, the model introduces an unintended transformation where the animal appears as if it were made of stone or rock, positioned above the crowd. This artifact gives the animal a rigid, statue-like appearance that contradicts the intended dynamic interaction with the people. The crowd also appears like a repetitive pattern causing a sense of dissonance. As a result, the model's output diverges from the prompt by introducing an unnatural texture and an unrealistic spatial positioning, misrepresenting both the visual characteristics and the intended context of the scene.

represent the predominant categories of hallucinations, accounting for more than 50% of the hallucinated content observed in current T2V models. This distribution implies that

these models frequently encounter challenges in ensuring logical coherence between the generated image and the textual input, as well as in faithfully representing all compo-

nents specified in the prompt.

On the other hand, the **Vanishing Subject** category, the least frequent, indicates that T2V models occasionally struggle to consistently depict the primary subject. However, this issue is rarer than physical inconsistencies and omissions, highlighting that subject preservation is less common in T2V hallucinations.

**No Hallucination:** The generated video accurately reflects the given context, with no extraneous or fabricated elements, ensuring no hallucination. The visual output aligns seamlessly with the real-world description in the prompt, maintaining fidelity to the scenario and avoiding implausible elements (see Fig. 9).

### 3.3. Dataset Analysis

The videos generated by the T2V open-source models vary in duration from 1 to 2 seconds, culminating in a total dataset comprising 3,782 individual videos. They exhibit characteristics corresponding to one of the five predefined hallucination categories. This distribution ensures a diverse dataset, facilitating a comprehensive evaluation and analysis of content that exhibits hallucinated elements. Tab. 1 shows the hallucination categories across different video models.

### 3.4. Annotation Details

Each of the 3,782 videos in ViBe was assigned a label corresponding to the most prominent hallucination type identified. While some videos may contain multiple hallucinations, we opted to annotate each video according to its most dominant hallucination category to ensure consistency in the annotation process.

### 3.5. Human Annotation

Although a total of 6,950 videos were generated across 10 T2V models (695 videos per model), human annotation, being resource-intensive, was performed only on a limited sample. The annotation guidelines are provided in the Algorithm 1.

#### 3.5.1. Inter-Annotator Agreement

To assess the consistency and reliability of our annotations, we computed both Cohen’s Kappa ( $\kappa$ ) [31] and Krippendorff’s Alpha ( $\alpha$ ) [32] for each hallucination category. These measures of inter-annotator agreement provide a quantitative assessment of the extent to which different annotators converge on their classifications. The identical inter-annotator agreement scores in Tab. 2 can be attributed to the limited sample size, as only two annotators\* were involved in the video annotation process. This restricts the possibility of differing interpretations or disagreements that could have been observed with a larger group of annotators.

\*Two graduate students

The results demonstrate a high level of agreement across most categories, indicating robust consistency in the annotation process.

Our analysis reveals that the **Physical Incongruity** category exhibited the highest inter-annotator reliability, with both  $\kappa$  and  $\alpha$  reaching a value of 0.87. This suggests that the criteria for identifying this particular type of hallucination are clear and well-defined, leading to consistent judgments among annotators. On the other hand, the **Omission Error** category yielded the lowest agreement scores, with  $\kappa$  and  $\alpha$  at 0.7474 and 0.7487 respectively. This lower consistency can be attributed to the subjective nature of evaluating time-based distortions, which may involve more interpretation and varying thresholds for identification across annotators.

**Inter-annotator challenge:** Multiple hallucinations can occur within a single video, and human cognition may prioritize one over the other. An example, shown in Fig. 10, was selected from videos used to assess inter-annotator agreement. Initially labeled as a **Vanishing subject**, a subsequent annotator categorized it as **Physical Incongruity**. Both interpretations are valid: the frisbee disappears over time (**Vanishing subject**), while the mismatch between the rendered player and camera angle creates cognitive dissonance (**Physical Incongruity**).

### 3.6. Open-source vs. Closed-source T2V models

Closed-source models typically generate videos with durations exceeding 4 seconds, while all 10 open-source videos in our study were limited to a maximum of 2 seconds. While hallucinations are present in closed-source models, their frequency appears lower. Among the 40 videos generated for each model, at least 6-8 videos from each model exhibited no hallucinations. This suggests that closed-source models tend to adhere more faithfully and consistently to the provided prompts compared to open-source models. The video quality and clarity of rendered objects are superior in closed-source models. In contrast, open-source models, particularly in lower-resolution videos, may exhibit issues where it becomes difficult to discern whether an object is undergoing morphing or if temporal dysmorphia is present due to the video’s reduced resolution.

## 4. Benchmark

Given the growing challenge of video hallucinations, addressing this issue is crucial. Currently, the literature includes only one T2V hallucination benchmark, T2VHaluBench [5], which consists of just 50 videos, limiting its utility for robust evaluation (Tab. 3). To overcome this, we propose a more comprehensive benchmark to drive further research, along with several classical classification baselines to support hallucination category prediction. We



Figure 9. **Prompt:** Looking out a train window at the scenery including a mountain. **No Hallucination:** The scenario described involves observing the landscape from a train window, which includes a mountain as part of the visible scenery. As per the initial input, the generated video accurately represents this context without introducing any extraneous or fabricated elements. Therefore, the generated content does not exhibit hallucination, as the visual output aligns directly with the real-world description provided in the prompt. The depiction is faithful to the input, with no deviations or inaccuracies that would characterize a hallucination in the model’s output.

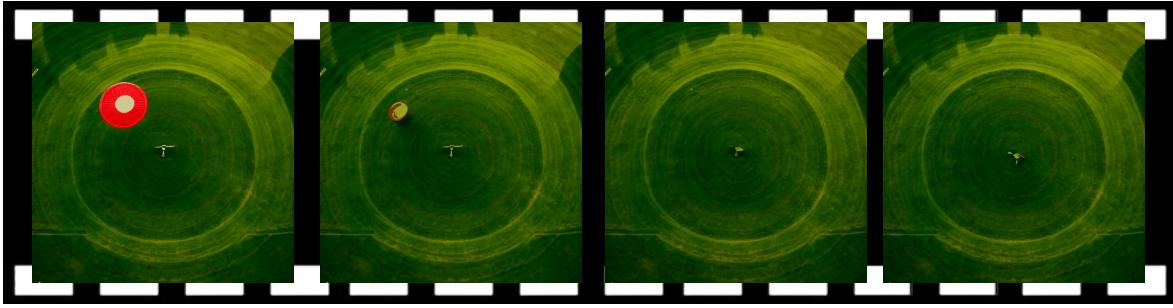


Figure 10. **Prompt:** A frisbee player throws the frisbee across the grass. **Multiple Hallucinations:** This hallucination may manifest as both a **Vanishing Subject** and a **Physical Incongruity**. The vanishing subject is observed when the frisbee disappears over time, while the incongruity arises from an unusual camera angle, such as a top-down view, which distorts the person’s appearance and induces a sense of dissociation.

---

#### Algorithm 1: Annotation Guidelines

---

- 1 Provide annotators with comprehensive guidelines specifying definitions, additional criteria, and examples for each hallucination category.
  - 2 Conduct training by reviewing example videos for each category, then assess annotators’ understanding by having them classify a set of 5 videos—each representing a distinct category.
  - 3 Evaluate annotator performance through a validation process, requiring them to achieve at least 60% agreement (correctly classifying 3 out of 5 videos) with reference annotations to advance.
- 

expect this benchmark to be a key resource for advancing research in this domain.

T2V Hallucination Benchmark	# Videos
T2VHaluBench [4]	50
ViBe	3,782

Table 3. 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.

#### 4.1. 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) [27] and TimeSFormer (Time-Space Attention Network for Video Understanding) [2]. These extracted embeddings are subsequently used as feature representations for seven distinct classification algorithms: Long Short-Term Memory (LSTM) [26], Transformer [28], Convolutional Neural Network (CNN)

T2V Model	VS	NV	TD	OE	PI	Total
AnimateLCM [29]	2	70	70	70	70	282
zeroscope_v2_XL [25]	18	0	37	109	199	363
Show1 [35]	13	71	88	111	55	338
MORA [34]	82	96	99	202	215	694
AnimateDiff Lightning [13]	11	33	52	56	63	215
AnimateDiff-MotionAdapter [9]	28	59	158	182	94	521
MagicTime [33]	70	70	70	69	70	349
zeroscope_v2_576w [24]	17	0	41	115	187	360
MS1.7B [1]	51	50	70	70	70	311
HotShotXL [19]	70	70	70	69	70	349
<b>Total</b>	<b>362</b>	<b>519</b>	<b>755</b>	<b>1053</b>	<b>1093</b>	<b>3782</b>

Table 1. Data Comparison Across Models with Totals for various Hallucination Categories

Hallucination Categories	Cohen's Kappa	Krippendorff's Alpha
<b>Vanishing Subject</b>	0.7660	0.7669
<b>Numeric Variability</b>	0.8500	0.8508
<b>Temporal Dysmorphia</b>	0.8173	0.8181
<b>Omission Error</b>	0.7474	0.7487
<b>Physical Incongruity</b>	0.8737	0.8743

Table 2. A comparative analysis of inter-annotator agreement metrics, focusing on Cohen's Kappa and Krippendorff's Alpha scores, to evaluate their effectiveness in measuring consistency across annotators.

[12], Gated Recurrent Unit (GRU) [6], Recurrent Neural Network (RNN) [18], Random Forest (RF) [10], and Support Vector Machine (SVM) [7]. This comprehensive evaluation across different model architectures allows for a thorough comparison of performance in classifying the given video dataset.

## 4.2. Experimental Setup

The dataset was partitioned into 80% for training and 20% for testing, and the Adam/AdamW optimizer was used [16]. Additional details are provided in Tab. 4.

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 4. Specifications of the model hyperparameters employed during the classifier training process: for both RF and SVM classifiers, default settings from scikit-learn [20] were applied.

Classification was performed using video embeddings extracted by TimeSformer and VideoMAE models, which operate on individual frames. However, the classification task did not explicitly utilize a frame-by-frame approach.

## 4.3. Results and Analysis

Tab. 5 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.

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	<b>0.345</b>	<b>0.342</b>
TimeSFormer + RNN	0.299	0.299
TimeSFormer + RF	0.341	0.282
TimeSFormer + SVM	0.270	0.274

Table 5. 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.

## 5. Conclusion, Limitations, and Future Work

With rapid advancements in generative AI, particularly T2V models, their performance now matches other modalities. However, hallucinations in these models pose significant challenges. To tackle this, we introduce a novel large-scale benchmark for evaluating hallucinations in T2V models, enabling standardized assessments and laying the groundwork for future research, comparative studies, and model improvements. The key contributions of this work are:

- Introduction of ViBe:, a novel large-scale benchmark specifically focused on evaluating hallucinations in T2V models.
- Comprehensive analysis of the dataset, including the establishment of baseline performance for hallucination detection via classifiers.

The limitation of our current work is that we do not address the detection of multiple hallucination categories within a single video, which remains a complex problem. Additionally, the inherent subjectivity of annotations introduces challenges, as individual evaluations may vary in terms of the threshold at which a specific level of hallucination is deemed acceptable or justifiable for exclusion.

Future work will focus on expanding the dataset to include emerging categories of hallucinations and exploring potential techniques for mitigating these errors.

## References

- [1] ali vilab. ali-vilab/text-to-video-ms-1.7b · hugging face. <https://huggingface.co/ali-vilab/text-to-video-ms-1.7b>, 2023. (Accessed on 10/28/2024). 3, 8
- [2] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding?, 2021. 7
- [3] 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. Language models are few-shot learners, 2020. 2
- [4] Zhixuan Chu, Lei Zhang, Yichen Sun, Siqiao Xue, Zhibo Wang, Zhan Qin, and Kui Ren. Sora detector: A unified hallucination detection for large text-to-video models. *arXiv preprint arXiv:2405.04180*, 2024. 7
- [5] Zhixuan Chu, Lei Zhang, Yichen Sun, Siqiao Xue, Zhibo Wang, Zhan Qin, and Kui Ren. Sora detector: A unified hallucination detection for large text-to-video models, 2024. 2, 6
- [6] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling, 2014. cite arxiv:1412.3555Comment: Presented in NIPS 2014 Deep Learning and Representation Learning Workshop. 8
- [7] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995. 8
- [8] Anastasis Germanidis. Gen-2: Generate novel videos with text, images or video clips. 3
- [9] Yuwei Guo. guoyww/animatediff-motion-adapter-v1-5-2 · hugging face. <https://huggingface.co/guoyww/animatediff-motion-adapter-v1-5-2>, 2023. (Accessed on 10/28/2024). 3, 8
- [10] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, pages 278–282. IEEE, 1995. 8
- [11] 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. The hallucinations leaderboard - an open effort to measure hallucinations in large language models. *CoRR*, abs/2404.05904, 2024. 2
- [12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2012. 8
- [13] Shanchuan Lin and Xiao Yang. Animatediff-lightning: Cross-model diffusion distillation, 2024. 3, 8
- [14] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. 3
- [15] Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. A token-level reference-free hallucination detection benchmark for free-form text generation, 2022. 2
- [16] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 8
- [17] Lumalabs. Dream machine. 3
- [18] Tomáš Mikolov, Martin Karafiat, Lukáš Burget, Jan Černocký, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Interspeech 2010*, pages 1045–1048, 2010. 8
- [19] John Mullan, Duncan Crawbuck, and Aakash Sastry. Hotshot-XL, 2023. 3, 8
- [20] 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. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011. 8
- [21] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 2
- [22] Vipula Rawte, Amit Sheth, and Amitava Das. A survey of hallucination in large foundation models. *arXiv preprint arXiv:2309.05922*, 2023. 2
- [23] 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. Photorealistic text-to-image diffusion models with deep language understanding, 2022. 2
- [24] Spencer Sterling. cerspense/zeroscope\_v2\_576w · hugging face. [https://huggingface.co/cerspense/zeroscope\\_v2\\_576w](https://huggingface.co/cerspense/zeroscope_v2_576w), 2023. (Accessed on 10/28/2024). 3, 8
- [25] Spencer Sterling. cerspense/zeroscope\_v2\_xl · hugging face. [https://huggingface.co/cerspense/zeroscope\\_v2\\_XL](https://huggingface.co/cerspense/zeroscope_v2_XL), 2023. (Accessed on 10/28/2024). 3, 8
- [26] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2014. 7
- [27] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training, 2022. 7
- [28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 7
- [29] Fu-Yun Wang, Zhaoyang Huang, Weikang Bian, Xiaoyu Shi, Keqiang Sun, Guanglu Song, Yu Liu, and Hongsheng Li. Animateclm: Computation-efficient personalized style video generation without personalized video data, 2024. 3, 8
- [30] Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models, 2024. 2
- [31] Wikipedia.Cohen's\_Kappa. Cohen's kappa. 6
- [32] Wikipedia.Krippendorff's\_Alpha. Krippendorff's alpha. 6
- [33] Shenghai Yuan, Jinfa Huang, Yujun Shi, Yongqi Xu, Ruijie Zhu, Bin Lin, Xinhua Cheng, Li Yuan, and Jiebo Luo. Magictime: Time-lapse video generation models as metamorphic simulators, 2024. 3, 8
- [34] 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. Mora: Enabling generalist video generation via a multi-agent framework, 2024. 3, 8
- [35] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation, 2023. 3, 8