VideoUFO: A Million-Scale User-Focused Dataset for Text-to-Video Generation

Wenhao Wang
University of Technology Sydney
wangwenhao0716@gmail.com

Yi Yang*
Zhejiang University
yangyics@zju.edu.cn

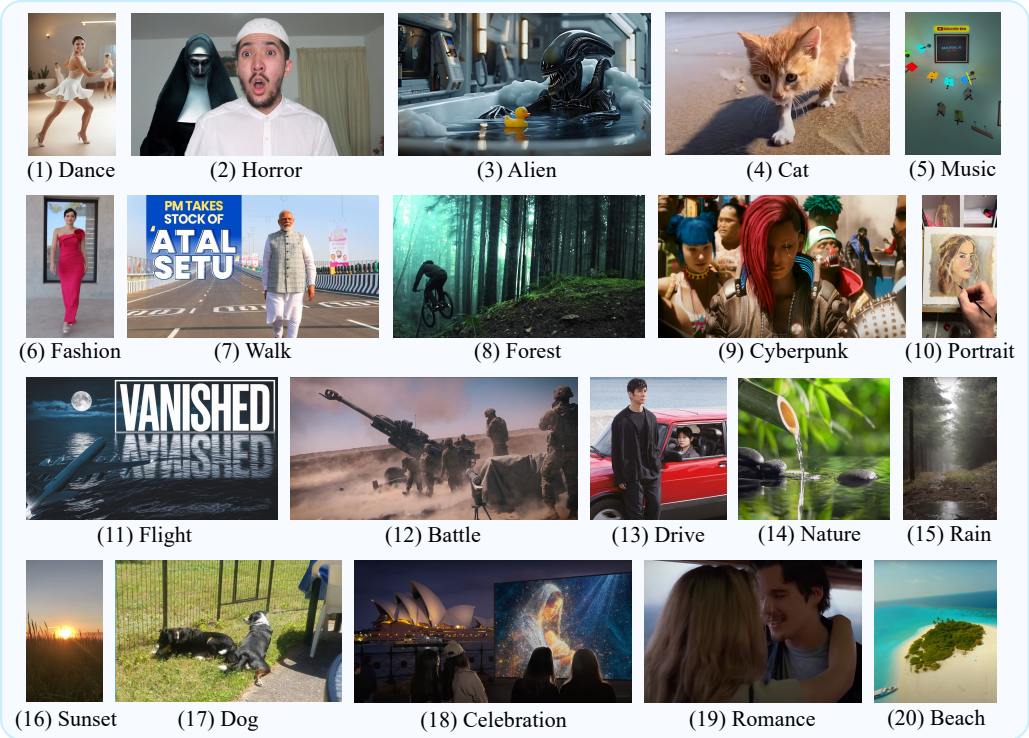


Figure 1: **VideoUFO** is the first dataset curated in alignment with real-world users' focused topics for text-to-video generation. Specifically, the dataset comprises over 1.09 million video clips spanning 1,291 topics. Here, we select the top 20 most popular topics for illustration. Researchers can use our VideoUFO to train or fine-tune their text-to-video generative models to better meet users' needs.

Abstract

Text-to-video generative models convert textual prompts into dynamic visual content, offering wide-ranging applications in film production, gaming, and education. However, their real-world performance often falls short of user expectations. One key reason is that these models have not been trained on videos related to some topics users want to create. In this paper, we propose **VideoUFO**, the first **Video** dataset specifically curated to align with **U**sers' **FO**cus in real-world scenarios. Beyond this, our VideoUFO also features: (1) minimal (0.29%) overlap with existing

*Corresponding Author.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

video datasets, and (2) videos searched exclusively via YouTube's official API under the Creative Commons license. These two attributes provide future researchers with greater freedom to broaden their training sources. The VideoUFO comprises over 1.09 million video clips, each paired with both a brief and a detailed caption (description). Specifically, through clustering, we first identify 1,291 user-focused topics from the million-scale real text-to-video prompt dataset, VidProM. Then, we use these topics to retrieve videos from YouTube, split the retrieved videos into clips, and generate both brief and detailed captions for each clip. After verifying the clips with specified topics, we are left with about 1.09 million video clips. Our experiments reveal that (1) current 16 text-to-video models do not achieve consistent performance across all user-focused topics; and (2) a simple model trained on VideoUFO outperforms others on worst-performing topics. The dataset and code are publicly available here and here under the CC BY 4.0 License.

1 Introduction

Text-to-video generation aims to convert textual descriptions into dynamic visual content. Its applications are extensive and transformative, covering areas from creative media [1] and entertainment [2] to practical domains such as education [3], advertising [4], and assistance [5].

Despite their popularity and usefulness, current text-to-video models often fail to meet users' expectations in real-world applications. For example, when we ask Sora [6] to generate a video using the prompt "*A firefly is glowing on a grass's leaf on a serene summer night*", it fails to capture the concept of a glowing firefly while successfully generating grass and a summer night, as shown in Fig. 2 (a). From the data perspective, we infer this is mainly because Sora [6] has not been trained on *firefly*-related topics, while it has been trained on *grass* and *night*. Furthermore, if Sora [6] has seen the video shown in Fig. 2 (b), it will understand what a glowing firefly should look like.

(a) Video generated by Sora

(b) Real video

Figure 2: The *glowing firefly*: (a) generated by Sora [6] and (b) captured in a real video. The generated firefly is noticeably different from its real-life counterpart and thus unsatisfying. We attribute this primarily to a lack of exposure to such topics.

To this end, this paper presents **VideoUFO**, the first **Video** dataset curated specifically based on real-world **U**sers' **FO**cus in text-to-video generation. Such a dataset can help improve the alignment between text-to-video models and actual user needs. Specifically, we focus on: **(1)** curating VideoUFO by analyzing real users' interests and scraping relevant videos; **(2)** comparing VideoUFO with other datasets to highlight differences; and **(3)** showing how VideoUFO benefits video generation.

The first dataset that aligns with real-world users' focus in text-to-video generation. Our key idea is to analyze user-focused topics from user-provided prompts and then search for videos related to these topics. As shown in Fig. 1, the resulted VideoUFO comprises more than **1.09** million video clips spanning **1,291** user-focused topics. Specifically, we initiate by analyzing user-focused topics, which involves embedding all 1.67 million user-provided prompts from VidProM [7], clustering these embeddings using K-means, and generating a topic for each cluster with GPT-4o [8], followed by combining similar topics. After obtaining these topics, we (1) search for these topics on YouTube, (2) segment the fetched videos into multiple semantically consistent short clips, (3) generate both brief and detailed captions for each clip, (4) filter out clips that do not contain the specific topics, and (5) assign each clip video quality scores that align with human perception. Note that although our current VideoUFO comprises about **one million** videos, it can be easily scaled up to **ten million** or more by sourcing additional videos for each topic. The VideoUFO can also be easily extended to the **image-to-video** domain by using the text and image prompts in TIP-I2V [9].

Differences between VideoUFO and other recent video datasets. Recently, several video datasets have been released, including OpenVid-1M [10], HD-VILA-100M [11], InternVid [12], Koala-36M [13], LVD-2M [14], MiraData [15], Panda-70M [16], VidGen-1M [17], and WebVid-10M [18].

While inheriting their excellent attributes – such as large scale, accurate captions, and high resolution – our VideoUFO explores some novel directions: (1) Guided by real user focus. Specifically, whereas recent datasets are typically gathered from open-domain sources, we concentrate on topics that are focused by text-to-video users. This feature enables text-to-video models trained on our dataset to better cater to users, while avoiding unnecessary expansion of the dataset and wasting resources. (2) Introducing new data. Although these recent papers claim to contribute new datasets, most of them primarily introduce a new data pipeline – that is, they reprocess HD-VILA-100M [11] and are subsets of it. While their contributions are useful and meaningful, in theory, a generative model that has already been fully fitted on HD-VILA-100M [11] would not gain new video knowledge from them. In contrast, we collect new data from YouTube, with only 0.29% of the videos overlapping with existing datasets. (3) Data compliance. When curating VideoUFO, we retrieve videos using YouTube's official API and select only those with a Creative Commons license. In contrast, most recent datasets do not explicitly address the regulatory compliance of their data collection process. This feature grants researchers greater flexibility in using our data.

Benchmarking current text-to-video models on user-focused topics and demonstrating the effectiveness of our VideoUFO. We present a new benchmark to quantify the performance of text-to-video models on user-focused topics. Specifically, our process involves: (1) selecting 10 user-provided prompts per topic; (2) generating a video for each prompt; (3) using a multimodal large language model to describe each video; and (4) calculating the similarity between the generated descriptions and the original prompts. For each topic, we calculate the average similarity between the 10 prompts and their corresponding descriptions. By sorting these averages, we can identify the worst-performing and best-performing topics for each model. Using this benchmark, we evaluate the performance of current text-to-video models and a newly trained text-to-video model on our VideoUFO. Experimental results indicate that (1) current 16 text-to-video models have some poor-performing topics, and (2) our model achieves the highest similarities on worst-performing topics while maintaining performance on the best-performing ones.

In conclusion, our key contributions are as follows:

1. We present VideoUFO, the first video dataset curated based on the focus of real text-to-video users. This dataset comprises over 1.09 million clips spanning $1,291$ user-focused topics.
2. We compare VideoUFO with recent video datasets, highlighting their differences in both fundamental attributes and topics coverage, thereby emphasizing the necessity of our dataset. We also follow best practices in their curation processes to ensure the quality of our dataset.
3. We evaluate current text-to-video models on user-focused topics and observe that a simple model trained on our VideoUFO outperforms competing models on worst-performing topics.

2 Related Works

Text-to-Video Generation. The introduction of Sora [6] has sparked significant research interest in text-to-video generation. Commercial models such as Movie Gen [19], Veo [20], Kling [21], Pixel-Dance [22], and Seaweed [22] demonstrate strong performance and are increasingly being integrated into various industries. Meanwhile, open-source models like HunyuanVideo [23], LTX-Video [24], CogVideoX [25], Mochi-1 [26], and Pyramidal [27] empower researchers to experiment, customize, and enhance existing frameworks. Although they are powerful, their training data determines the upper limit of generated videos' quality. This paper proposes a dataset, VideoUFO, which has the potential to help these models better cater to users' preferences in real-world applications.

Text-Video Datasets. A text-video dataset consists of video clips paired with corresponding textual descriptions or captions. Many text-video datasets, such as OpenVid-1M [10], HD-VILA-100M [11], InternVid [12], Koala-36M [13], LVD-2M [14], MiraData [15], Panda-70M [16], VidGen-1M [17], and WebVid-10M [18], have been proposed to advance multimodal understanding and generation. However, these datasets are not designed to align with the preferences of text-to-video users, creating a gap between training data topics and real-world applications. In contrast, our dataset VideoUFO is collected from a user-focus perspective. To better cater users, future researchers can fine-tune text-to-video models with the VideoUFO or integrate it with existing datasets to develop new models.

Preference Alignment in Diffusion Models. While diffusion models have demonstrated impressive capabilities, there is an increasing demand to ensure their outputs align with human preferences.

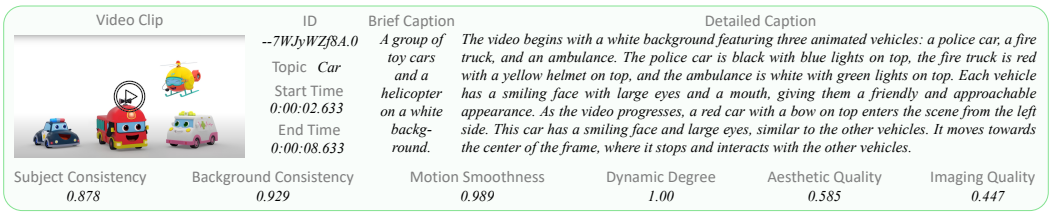


Figure 3: Each data point in our **VideoUFO** includes a video clip, an ID, a topic, start and end times, a brief caption, and a detailed caption. Beyond that, we evaluate each clip with six different video quality scores from VBench [28].

Direct Preference Optimization (DPO) [29], originally proposed to align large language models (LLMs) with human preferences, can also be effectively applied to diffusion models. For example, DiffusionDPO [30] fine-tunes text-to-image models using human comparison data to improve image generation quality. Additionally, VideoDPO [31] introduces a comprehensive preference scoring system that evaluates both visual quality and semantic alignment to enhance text-to-video generation. Unlike other approaches that align text-to-image or text-to-video models with attributes such as aesthetics, motion, consistency, and visual appeal, our goal is to improve the performance of text-to-video models on real-world users' focused topics.

3 Curating VideoUFO

Fig. 3 illustrates a data point from our million-scale VideoUFO, which comprises: (1) a video clip along with its ID; (2) the topic of the clip; (3) the start and end times within its original video; (4) the corresponding brief and detailed captions; and (5) the quality of the clip. The following provides an explanation of how we curate the VideoUFO.

Analyzing text-to-video users' focused topics. We aim to analyze the topics that real users focus on or prefer when generating videos from texts. Here, we use VidProM [7] as it is the only publicly available, million-scale text-to-video prompt dataset written by real users. **F**irst, we embed all 1.67 million prompts from VidProM into 384-dimensional vectors using `SentenceTransformers` [33]. **N**ext, we cluster these vectors with K-means. Note that here we pre-set the number of clusters to a relatively large value, *i.e.*, $2,000$, and merge similar clusters in the next step. **F**inally, for each cluster, we ask GPT-4o [8] to conclude a topic ($1 \sim 2$ words). The prompt is shown in the Appendix (Section A). After merging singular and plural forms, restoring verbs to their base forms, removing duplicates, and manually verifying the topics – removing overly broad ones such as '*animation*', '*scene*', '*movement*', and '*film*' – we finally obtained $\mathbf{1,291}$ topics. The semantic distribution of these topics is shown in Fig. 4.

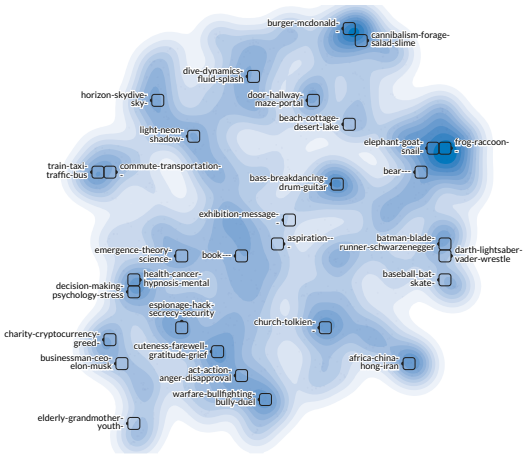


Figure 4: The semantic distribution of users' focused topics. It is visualized by WizMap [32]. Please 🔍 zoom in or visit here to see the details.

Collecting videos from YouTube. For each topic obtained, we use the official YouTube API to search for videos. Specifically, for each topic, we search for approximately 500 videos by relevance, with each video meeting the following criteria: **(1)** it is shorter than 4 minutes, **(2)** it has a resolution of 720p or higher, and **(3)** it is licensed under Creative Commons. These requirements ensure that the collected videos are **suitable** and **freely usable** for training video generation models. In the end, we obtain $586,490$ videos from YouTube. We compare the YouTube IDs of these videos with those in existing datasets, including OpenVid-1M [10], HD-VILA-100M [11], InternVid [12], Koala-36M [13], LVD-2M [14], MiraData [15], Panda-70M [16], VidGen-1M [17], and WebVid-10M [18], and find that only $1,675$ IDs (0.29%) are already present in these datasets. This suggests that our

Table 1: The comparison between the VideoUFO with recent video datasets based on their **fundamental attributes**. Unlike previous collections, our VideoUFO is derived directly from real user-focused topics and also offers a more flexible license, while remaining comparable to these datasets in other aspects. "#" and "¯" are abbreviations for "numbers" and "average", respectively.

Dataset	#Vid.	Len.¯	Words¯	Resolution	Domain	#Topic	License
WebVid-10M [18]	10M	18.0s	14.2	<360p	Open	1,000	Retracted
HD-VILA-100M [11]	103M	13.4s	32.5	720p	Open	648	R-UDA
InternVid [16]	234M	11.7s	17.6	720p	Open	1,051	Apache 2.0
Panda-70M [16]	70M	8.5s	13.2	720p	Open	719	R-UDA
LVD-2M [14]	2M	20.2s	88.7	Diverse	Open	814	R-UDA
MiraData [15]	0.33M	72.1s	318.0	720p	Open	639	GPL 3.0
Koala-36M [13]	36M	17.2s	202.1	720p	Open	724	R-UDA
VidGen-1M [17]	1M	10.6s	89.3	720p	Open	835	R-UDA
OpenVid-1M [10]	1M	7.2s	127.3	Diverse	Open	671	R-UDA
ChronoMagic-Pro [36]	460K	234s	1794.0	720p	Time	792	CC BY
VideoUFO	1M	12.6s	155.5	720p	**Users**	$\mathbf{1,291}$	**CC BY**

VideoUFO introduces novel information or knowledge, which can serve to **expand the range of existing training sources**.

Splitting videos and generating captions. After obtaining the videos, we segment them into multiple semantics-consistent short clips following the steps in curating Panda-70M [16]. The process includes *shot boundary detection*, *stitching*, and *video splitting*, producing a total of $\mathbf{3,181,873}$ clips. To facilitate the training of text-to-video models on our VideoUFO, we generate both brief and detailed captions for each video clip. For the brief captions, we utilize the video captioning model provided by Panda-70M [16]. For the detailed captions, we adopt the pipeline used in `Open-Sora-Plan` v1.3.0 [34], which employs `QWen2-VL-7B` [35] for video annotation. In the Appendix (Fig. 10 (a) and (b)), we present the statistical distribution of the lengths of brief and detailed captions, respectively.

Verifying clips. We notice that, although the videos are searched by topics, not every clip contains our intended topics. A straightforward solution is to use GPT-4o [8] to verify whether a given clip corresponds to a specific topic. However, while effective, this approach would be prohibitively expensive for verifying more than 3 million video clips. As an alternative, we use the **detailed caption** (instead of the original clip) and feed it into **GPT-4o mini** (instead of GPT-4o) to verify whether it belongs to a specific topic. Since (1) GPT-4o mini and GPT-4o have similar capabilities in basic language understanding, and (2) the video understanding model provides detailed descriptions of the videos, we effectively complete the final verification step at a significantly lower cost. After verifying, there are $\mathbf{1,091,712}$ remaining clips. In the Appendix (Fig. 10 (c)), we present the statistical distribution of clips duration.

Video quality assessment. To further support research in text-to-video generation, we evaluate the quality of each clip in VideoUFO. Specifically, we adopt six different video quality assessment metrics from VBench [28], which automatically assess videos and align well with human perception. The evaluation dimensions are: **(1)** *subject consistency*: assesses whether the main subject's identity and appearance remain consistent throughout the video; **(2)** *background consistency*: evaluates the temporal stability and uniformity of the video's background; **(3)** *motion smoothness*: measures the continuity and fluidity of movements within the video; **(4)** *dynamic degree*: assesses the level of activity and variation in motion present in the video; **(5)** *aesthetic quality*: judges the visual appeal and attractiveness of the video; and **(6)** *imaging quality*: examines the clarity, brightness, and color accuracy of the video. The statistical distributions of scores assessed by these six metrics are shown in the Appendix (Fig. 11).

Extension. Future researchers can easily extend our VideoUFO in three aspects: **(1) Scaling up.** Although our VideoUFO has already reached a million-scale level, future researchers may still want to scale it up to 10 million or even 100 million. They can easily achieve this by leveraging our extracted topics to search more videos on platforms such as YouTube and TikTok. **(2) New focusing.** When we curate VideoUFO, the only available text-to-video prompt dataset is VidProM [7], which is used to analyze user focus and preferences. In the future, user focus may change, and other text-to-video prompt datasets may emerge for analyzing it. Future researchers can easily use our topic extraction pipeline to study these new focuses. **(3) Image-to-video.** Our current focus is text-to-video, which is the most common approach in the video generation community. Meanwhile, we also notice that

image-to-video is gaining popularity. Future researchers can use text and image prompts, such as those in TIP-I2V [9], to analyze the focus of image-to-video users and collect corresponding datasets.

4 Comparison with Other Video Datasets

This section compares the proposed VideoUFO with other recent video datasets in terms of *fundamental attributes* and *topics coverage*. These differences underscore the necessity of introducing VideoUFO for text-to-video generation.

4.1 Fundamental Attributes

From the Table 1, we draw three conclusions:

- VideoUFO is collected in line with real text-to-video **users' focus** or preferences. In contrast to VideoUFO, other datasets collect videos from the open domain, which may not cover the topics users focus on, and text-to-video models trained on them may fail to meet users' needs.
- VideoUFO is released under a more **flexible license** and introduces **new data**. Specifically, we search for videos on YouTube by ourselves and only select those with a Creative Commons license. In contrast, most recent datasets (including Panda-70M [16], LVD-2M [14], Koala-36M [13], VidGen-1M [17], and OpenVid-1M [10]) directly source data from HD-VILA-100M [11]. As a result: (1) they do not contribute new data but rather introduce a new data processing pipeline; (2) they must adhere to the same license as HD-VILA-100M, *i.e.*, the Research Use of Data Agreement (R-UDA), which restricts commercial use. In addition, one of the most widely used datasets, WebVid-10M [18], has been retracted due to potential copyright infringement.
- VideoUFO **inherits best practices** from other curated datasets. We observe that recent datasets (1) feature detailed captions generated by multimodal large language models (MLLMs), (2) contain millions of video clips, and (3) are high-resolution (*i.e.*, 720p). Therefore, to provide a large-scale and high-quality resource, our VideoUFO inherits these features.

4.2 Topics Coverage

This section analyzes the differences in topics coverage between the VideoUFO and existing datasets.

Calculation process. We calculate the number of user-focused topics covered by the existing datasets as follows:

- Extract topics of recent video datasets. We repeat VidProM's topic extraction process (Step 1 in Section 3) on the video datasets listed in Table 1. The number of topics extracted for each dataset is shown in Table 1 (#Topics).
- Match the extracted topics with users' focused ones. We observe that the same or similar topic may be described using different words. For example, both *cathedral* and *church* refer to a similar topic. Therefore, when comparing user-focused topics with those from each existing dataset, we choose semantic matching rather than a word-to-word approach. Specifically, (1) we first use `SentenceTransformers` [33] to embed the two lists of topics; and (2) then for each user-focused topic, if there exists a topic in the existing dataset with a cosine similarity greater than 0.6, we consider that user-focused topic to be covered by the existing dataset. Note that the threshold of 0.6 is an empirical value, as we observe that most similar topics exhibit a cosine similarity greater than 0.6.

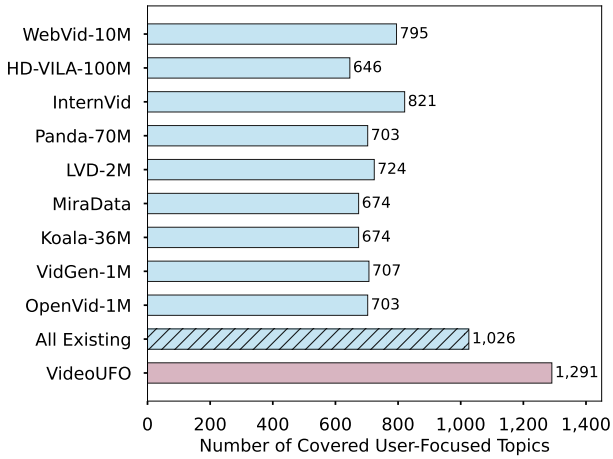


Figure 5: The number of user-focused topics covered by recent video datasets. None successfully includes all user-focused topics.

Observations. The experimental results in Fig. 5 shows:

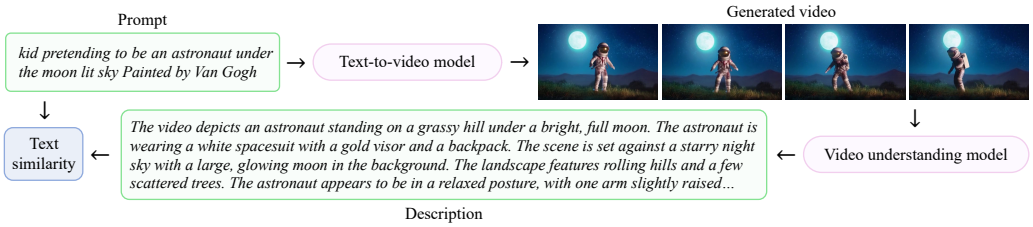


Figure 6: The calculation process of BenchUFO. It is designed to evaluate whether a text-to-video model can effectively generate videos that contain user-focused topics. It comprises 791 concrete noun topics, each paired with 10 real-world user-provided prompts.

- None of the existing datasets cover all user-focused topics in VideoUFO. Specifically, the most comprehensive dataset, InternVid [12], covers 821 topics, accounting for 63.6% of all user-focused topics. Furthermore, if we combine all existing datasets in Table 1 into the "largest" one, the coverage of user-focused topics will increase to 79.5%. However, this will lead to a higher computational burden when training text-to-video models.
- Interestingly, a dataset may cover more user-focused topics than the actual number of topics it contains. For instance, OpenVid-1M [10] contains 671 topics but covers 703 user-focused ones. This is reasonable because two topics in our dataset, VideoUFO, may be similar and thus mapped to a single topic in OpenVid-1M [10]. Nevertheless, this does **not** compromise the quality of our VideoUFO since it only leads to some relative duplicated topics, while the videos in VideoUFO still accurately reflect real users' focus and interests.

Limitation. The above topic analysis is based directly on the captions/descriptions provided by each recent video dataset. This is because downloading and re-captioning all these videos is prohibitively expensive in terms of network bandwidth and computation. The quality of these captions/descriptions may affect the number of extracted topics, as some may not be informative. To address this limitation, in the next section, we demonstrate that text-to-video models trained on these datasets fail to generate satisfactory videos for some topics.

5 VideoUFO Benefits Video Generation

5.1 Benchmark

This section details the proposed benchmark, BenchUFO, for evaluating text-to-video models' performance on user-focused topics. We first introduce the design of the benchmark and then explain why it is reasonable.

Construction. As shown in Fig. 6, the calculation process of our BenchUFO includes: **(1) Selecting prompts.** First, we select 791 concrete nouns from 1,291 user-focused topics. (Please refer to the next section for the rationale behind selecting concrete nouns.) Then, for each chosen topic, we randomly select 10 text prompts from VidProM [7]. All these prompts constitute the prompt set for our benchmark. **(2) Generating videos.** For each prompt, we use a text-to-video model to generate a corresponding video. **(3) Describing videos.** For each video, we use a multimodal large language model (here, we choose `QWen2-VL-7B` [35]) to understand and describe it. The instruction prompt is provided in the Appendix (Section B). **(4) Calculating similarity.** We use a sentence embedding model (in this case, `SentenceTransformers` [33]) to encode both the input prompt and the output description, and then compute the cosine similarity between them. We calculate the average similarity across 10 prompts for each topic. A higher similarity score indicates better performance on that specific topic. Our benchmark computes and considers the 10/50 worst-performing and 10/50 best-performing topics.

Justification. This section provides justification for the proposed BenchUFO from three perspectives:

- Why choose concrete nouns? This is due to the inconsistency between the input prompt and the output description when using abstract nouns. For instance, for the abstract noun topic *"freedom"* and the prompt *"In a freedom world, ideal and love both exist"*, a model may generate videos containing *"dove"*, *"star"*, or *"heart"*, and the video understanding model will summarize them as *"a dove is playing..."*, *"a star is shining..."*, and *"a heart is beating..."*. However, the embeddings of

these descriptions are not expected to be close to those of the given prompt. This might create a false impression that the model cannot effectively generate content for that topic.

- Why use video understanding model reasonable? This is because the video understanding model has advanced significantly with the advent of large language models, and it is reasonable to assume that these models have been exposed to images/videos containing user-focused topics. Here, we choose `QWen2-VL-7B` [35] as the video understanding model. Trained on 800 billion tokens of visual-related data and 600 billion tokens of text data, it is expected to accurately comprehend these user-focused topics and faithfully summarize a video.
- Why not use established benchmarks? This is because the existing benchmarks fail to reflect real-world scenarios. Specifically, we note that there exists a well-established benchmark, VBench [28], whose "*object class*" and "*multiple objects*" dimensions are similar to our BenchUFO. However, the prompts in VBench have two main drawbacks: **(1)** They cover only a limited number of topics/objects – specifically, just 79. **(2)** They focus exclusively on common topics/objects (e.g., "*bicycle*", "*car*", and "*airplane*"), while many topics that users care about, such as "*forest*", "*sunset*", and "*beach*", are missing.

5.2 Observations

In this section, we evaluate current text-to-video models on this new benchmark and show that, with the help of VideoUFO, we achieve state-of-the-art performance. The quantitative and qualitative results are shown in Table 7 and Fig. 8, respectively. We observe that:

Figure 7: The performance of both publicly available text-to-video models and our trained models on the proposed BenchUFO. The publicly available models are trained on various datasets, including both public and private ones. MVDiT [10] is trained on VidGen [17], OpenVid [10], and VideoUFO, respectively. "Low/Top N" denotes the average score of the worst/best-performing N topics.

Models	Low 10	Low 50	Top 50	Top 10
Mira [37]	0.236	0.282	0.508	0.550
Show-1 [38]	0.266	0.303	0.524	0.564
LTX-Video [24]	0.268	0.310	0.532	0.574
Open-Sora-Plan [34]	0.314	0.361	0.559	0.598
TF-T2V [39]	0.316	0.359	0.560	0.595
Mochi-1 [26]	0.323	0.367	0.580	0.616
HiGen [40]	0.352	0.394	0.589	0.625
Open-Sora [41]	0.363	0.409	0.601	0.639
Pika [42]	0.365	0.404	0.583	0.619
RepVideo [43]	0.368	0.402	0.589	0.619
T2V-Zero [44]	0.375	0.410	0.586	0.616
CogVideoX [25]	0.383	0.419	0.601	0.629
Latte-1 [45]	0.384	0.421	0.592	0.636
HunyuanVideo [23]	0.388	0.427	0.612	0.645
LaVie [46]	0.399	0.426	0.595	0.632
Pyramidal [27]	0.400	0.433	0.606	0.647
MVDiT-VidGen [17]	0.382	0.426	0.594	0.626
MVDiT-OpenVid [10]	0.401	0.437	0.609	0.645
MVDiT-VideoUFO	**0.442**	**0.465**	**0.619**	**0.651**

- Current text-to-video models do not consistently perform well across all user-focused topics. Specifically, there is a score difference ranging from 0.233 to 0.314 between the top-10 and low-10 topics. These models may not effectively understand topics such as *"giant squid"*, *"animal cell"*, *"Van Gogh"*, and *"ancient Egyptian"* due to insufficient training on such videos.
- Current text-to-video models show a certain degree of consistency in their best-performing topics. We discover that most text-to-video models excel at generating videos on animal-related topics, such as *'seagull'*, *'panda'*, *'dolphin'*, *'camel'*, and *'owl'*. We infer that this is partly due to a bias towards animals in current video datasets.
- The proposed VideoUFO helps reduce the gap between the worst-performing and best-performing topics. To demonstrate the effectiveness of the proposed VideoUFO, we train an MVDiT [10] solely on VideoUFO. We find that the trained model achieves the highest low-10 scores (a $+4.2\%$ improvement compared to the current state-of-the-art) while maintaining performance on the top-10 topics. Visually, the trained model successfully generates videos on topics that other models previously could not.
- The proposed VideoUFO outperforms other similar-scale datasets, such as OpenVid [10] and VidGen [17]. To demonstrate that the improvement originates from our VideoUFO rather than

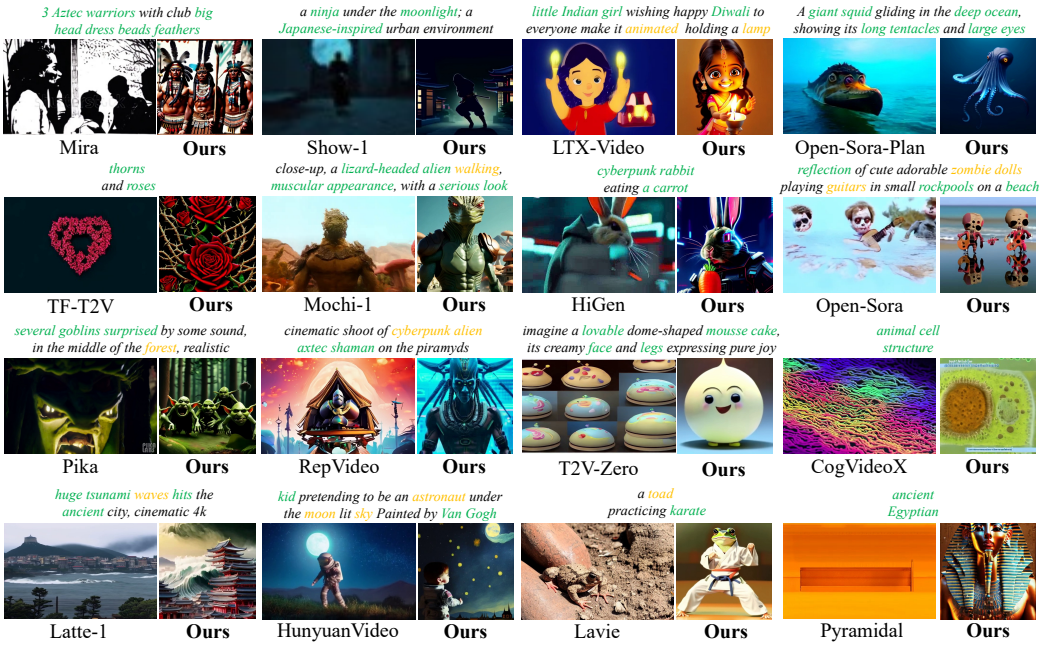


Figure 8: Visual comparisons between our approach (MVDiT-VideoUFO) and other text-to-video models. The model trained on VideoUFO outperforms the alternatives in generating user-focused topics. In the prompts, concepts in green indicate those successfully generated **solely** by our method, whereas those in yellow denote concepts successfully generated by both our model and the competing models. All generated videos are provided in the supplementary materials.

from the MVDiT architecture [10], we replace our VideoUFO with OpenVid [10] and VidGen [17]. We find that when using the popular OpenVid [10], the low-10 scores are similar to previous state-of-the-art models (0.401 vs. 0.400). Furthermore, when training on VidGen [17], the performance is even lower.

Limitation. Due to limited computational resources, when validating the effectiveness of our dataset, we currently follow the video generation method from the academic community introduced by [10], which already requires 32 A100 GPUs. However, the quality of videos generated by this baseline remains limited. In the future, industry researchers could incorporate our dataset into their large-scale foundational model training pipelines to further explore its potential.

6 Ablation Studies on Curation of VideoUFO

This section performs ablation studies to investigate key components in curating VideoUFO, *i.e.*, examining (1) the impact of the number of clusters in user-focused topic analysis and (2) the methods for clip verification. For more details, please refer to Appendix (Section C).

7 Conclusion

This paper presents a newly collected million-scale dataset for text-to-video generation with a focus on user needs. Beyond this, our dataset exhibits minimal overlap with existing datasets and is released under a more permissive license (*i.e.*, CC BY). We first describe the curation process, which involves analyzing user-focused topics, searching for these topics, segmenting the resulting videos, and applying various post-processing techniques. Then, we highlight the differences between our dataset and existing ones in terms of fundamental attributes and topics coverage. Finally, we build a new benchmark to evaluate text-to-video models on user-focused topics and demonstrate that our dataset helps improve model performance in this regard. We encourage both the research community and industry to use our dataset to further advance the field of text-to-video generation.

Acknowledgments

We sincerely thank OpenAI for their support through the Researcher Access Program. Without their generous contribution, this work would not have been possible.

References

[1] Stuart Marshall Bender. Coexistence and creativity: Screen media education in the age of artificial intelligence content generators. *Media Practice and Education*, 24(4):351–366, 2023.

[2] Tao Yu, Wei Yang, Junping Xu, and Younghwan Pan. Barriers to industry adoption of ai video generation tools: A study based on the perspectives of video production professionals in china. *Applied Sciences*, 14(13):5770, 2024.

[3] Marta Garcia-Sampedro, Susana Agudo Prado, and Antonio Torralba-Burrial. Pre-service teachers' skills development through educational video generation. *European Journal of Teacher Education*, pages 1–19, 2024.

[4] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024.

[5] Chenxin Li, Hengyu Liu, Yifan Liu, Brandon Y Feng, Wuyang Li, Xinyu Liu, Zhen Chen, Jing Shao, and Yixuan Yuan. Endora: Video generation models as endoscopy simulators. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 230–240. Springer, 2024.

[6] Sora: Ai text-to-video model. https://openai.com/index/sora/, 2024.

[7] Wenhao Wang and Yi Yang. Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. *Thirty-eighth Conference on Neural Information Processing Systems*, 2024.

[8] OpenAI. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024.

[9] Wenhao Wang and Yi Yang. Tip-i2v: A million-scale real text and image prompt dataset for image-to-video generation. *arXiv preprint arXiv:2411.04709*, 2024.

[10] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. *International Conference on Learning Representations*, 2025.

[11] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[12] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. *The Twelfth International Conference on Learning Representations*, 2024.

[13] Qiuheng Wang, Yukai Shi, Jiarong Ou, Rui Chen, Ke Lin, Jiahao Wang, Boyuan Jiang, Haotian Yang, Mingwu Zheng, Xin Tao, et al. Koala-36m: A large-scale video dataset improving consistency between fine-grained conditions and video content. *IEEE/CVF conference on computer vision and pattern recognition*, 2025.

[14] Tianwei Xiong, Yuqing Wang, Daquan Zhou, Zhijie Lin, Jiashi Feng, and Xihui Liu. Lvd-2m: A long-take video dataset with temporally dense captions. *arXiv preprint arXiv:2410.10816*, 2024.

[15] Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. *Thirty-eighth Conference on Neural Information Processing Systems*, 2024.

[16] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[17] Zhiyu Tan, Xiaomeng Yang, Luozheng Qin, and Hao Li. Vidgen-1m: A large-scale dataset for text-to-video generation. *arXiv preprint arXiv:2408.02629*, 2024.

[18] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. *IEEE/CVF International Conference on Computer Vision*, 2021.

[19] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, David Yan, Dhruv Choudhary, Dingkang Wang, Geet Sethi, Guan Pang, Haoyu Ma, Ishan Misra, Ji Hou, Jialiang Wang, Kiran Jagadeesh, Kunpeng Li, Luxin Zhang, Mannat Singh, Mary Williamson, Matt Le, Matthew Yu, Mitesh Kumar Singh, Peizhao Zhang, Peter Vajda, Quentin Duval, Rohit Girdhar, Roshan Sumbaly, Sai Saketh Rambhatla, Sam Tsai, Samaneh Azadi, Samyak Datta, Sanyuan Chen, Sean Bell, Sharadh Ramaswamy, Shelly Sheynin, Siddharth Bhattacharya, Simran Motwani, Tao Xu, Tianhe Li, Tingbo Hou, Wei-Ning Hsu, Xi Yin, Xiaoliang Dai, Yaniv Taigman, Yaqiao Luo, Yen-Cheng Liu, Yi-Chiao Wu, Yue Zhao, Yuval Kirstain, Zecheng He, Zijian He, Albert Pumarola, Ali Thabet, Artsiom Sanakoyeu, Arun Mallya, Baishan Guo, Boris Araya, Breena Kerr, Carleigh Wood, Ce Liu, Cen Peng, Dimitry Vengertsev, Edgar Schonfeld, Elliot Blanchard, Felix Juefei-Xu, Fraylie Nord, Jeff Liang, John Hoffman, Jonas Kohler, Kaolin Fire, Karthik Sivakumar, Lawrence Chen, Licheng Yu, Luya Gao, Markos Georgopoulos, Rashel Moritz, Sara K. Sampson, Shikai Li, Simone Parmeggiani, Steve Fine, Tara Fowler, Vladan Petrovic, and Yuming Du. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

[20] Abhishek Sharma, Adams Yu, Ali Razavi, Andeep Toor, Andrew Pierson, Ankush Gupta, Austin Waters, Aäron van den Oord, Daniel Tanis, Dumitru Erhan, Eric Lau, Eleni Shaw, Gabe Barth-Maron, Greg Shaw, Han Zhang, Henna Nandwani, Hernan Moraldo, Hyunjik Kim, Irina Blok, Jakob Bauer, Jeff Donahue, Junyoung Chung, Kory Mathewson, Kurtis David, Lasse Espeholt, Marc van Zee, Matt McGill, Medhini Narasimhan, Miaosen Wang, Mikołaj Bińkowski, Mohammad Babaeizadeh, Mohammad Taghi Saffar, Nando de Freitas, Nick Pezzotti, Pieter-Jan Kindermans, Poorva Rane, Rachel Hornung, Robert Riachi, Ruben Villegas, Rui Qian, Sander Dieleman, Serena Zhang, Serkan Cabi, Shixin Luo, Shlomi Fruchter, Signe Nørly, Srivatsan Srinivasan, Tobias Pfaff, Tom Hume, Vikas Verma, Weizhe Hua, William Zhu, Xinchen Yan, Xinyu Wang, Yelin Kim, Yuqing Du, and Yutian Chen. Veo. 2024.

[21] Kling - kuaishou - image-to-video. https://kling.kuaishou.com/en.

[22] ByteDance AI Lab. Pixeldance and seaweed: Advanced ai video generation models. https://pixeldance.io/zh, 2024.

[23] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024.

[24] Lightricks. Ltx-video: Open-source video editing tools, 2024.

[25] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *The Thirteenth International Conference on Learning Representations*, 2025.

[26] Genmo Team. Mochi 1. https://github.com/genmoai/models, 2024.

[27] Yang Jin, Zhicheng Sun, Ningyuan Li, Kun Xu, Hao Jiang, Nan Zhuang, Quzhe Huang, Yang Song, Yadong Mu, and Zhouchen Lin. Pyramidal flow matching for efficient video generative modeling. *International Conference on Learning Representations*, 2025.

[28] 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. *the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[29] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2023.

[30] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[31] Runtao Liu, Haoyu Wu, Zheng Ziqiang, Chen Wei, Yingqing He, Renjie Pi, and Qifeng Chen. Videodpo: Omni-preference alignment for video diffusion generation. *arXiv preprint arXiv:2412.14167*, 2024.

[32] Zijie J Wang, Fred Hohman, and Duen Horng Chau. Wizmap: Scalable interactive visualization for exploring large machine learning embeddings. *Annual Meeting Of The Association For Computational Linguistics*, 2023.

[33] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *Conference on Empirical Methods in Natural Language Processing*, 2019.

[34] Bin Lin, Yunyang Ge, Xinhua Cheng, Zongjian Li, Bin Zhu, Shaodong Wang, Xianyi He, Yang Ye, Shenghai Yuan, Liuhan Chen, et al. Open-sora plan: Open-source large video generation model. *arXiv preprint arXiv:2412.00131*, 2024.

[35] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

[36] Shenghai Yuan, Jinfa Huang, Yongqi Xu, Yaoyang Liu, Shaofeng Zhang, Yujun Shi, Rui-Jie Zhu, Xinhua Cheng, Jiebo Luo, and Li Yuan. Chronomagic-bench: A benchmark for metamorphic evaluation of text-to-time-lapse video generation. *Advances in Neural Information Processing Systems*, 37:21236–21270, 2024.

[37] Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

[38] 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. *International Journal of Computer Vision*, 2024.

[39] Xiang Wang, Shiwei Zhang, Hangjie Yuan, Zhiwu Qing, Biao Gong, Yingya Zhang, Yujun Shen, Changxin Gao, and Nong Sang. A recipe for scaling up text-to-video generation with text-free videos. *the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[40] Zhiwu Qing, Shiwei Zhang, Jiayu Wang, Xiang Wang, Yujie Wei, Yingya Zhang, Changxin Gao, and Nong Sang. Hierarchical spatio-temporal decoupling for text-to-video generation. *the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

[41] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all, March 2024.

[42] Pika Labs. Pika labs: Ai text-to-video generation. https://pika.art/.

[43] Chenyang Si, Weichen Fan, Zhengyao Lv, Ziqi Huang, Yu Qiao, and Ziwei Liu. Repvideo: Rethinking cross-layer representation for video generation. *arXiv preprint arXiv:2501.08994*, 2025.

[44] 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. *the IEEE/CVF International Conference on Computer Vision*, 2023.

[45] Xin Ma, Yaohui Wang, Gengyun Jia, Xinyuan Chen, Ziwei Liu, Yuan-Fang Li, Cunjian Chen, and Yu Qiao. Latte: Latent diffusion transformer for video generation. *arXiv preprint arXiv:2401.03048*, 2024.

[46] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, et al. Lavie: High-quality video generation with cascaded latent diffusion models. *International Journal of Computer Vision*, 2024.

[47] Leland McInnes, John Healy, Steve Astels, et al. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, 2(11):205, 2017.

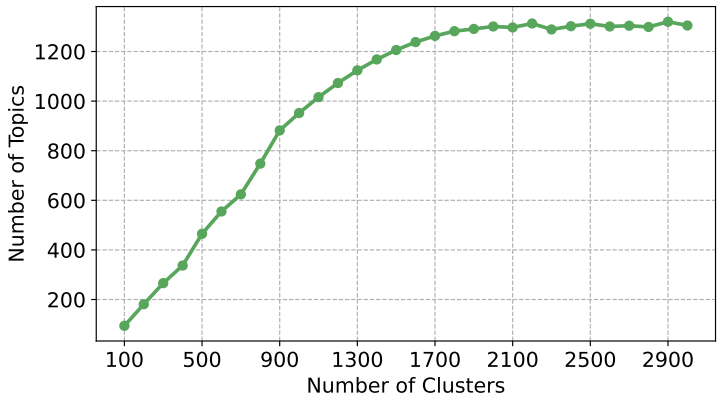


Figure 9: The relationship between pre-set number of clusters for K-means and final number of user-focused topics.

Table 2: The methods for clip verification. We use $1,000$ clips for this ablation study. "Direct GPT-4o" refers to processing videos directly with GPT-4o, and "Overlap" measures the extent to which the predictions of other methods match those of "Direct GPT-4o".

	Method	# Verified	Overlap	API Cost
	Direct GPT-4o	507	–	$162.23
Brief	GPT-4o	174	0.621	$0.04
	GPT-4o-mini	276	0.657	$0.01
Detail	GPT-4o	299	0.770	$0.46
	GPT-4o-mini	393	0.806	$0.03

A The Prompt for Concluding Topics

Prompt: Could you describe the topic in the following short sentences using only 1-2 words? Please return only the topic (1-2 word) as a singular noun or in the base form of the verb.

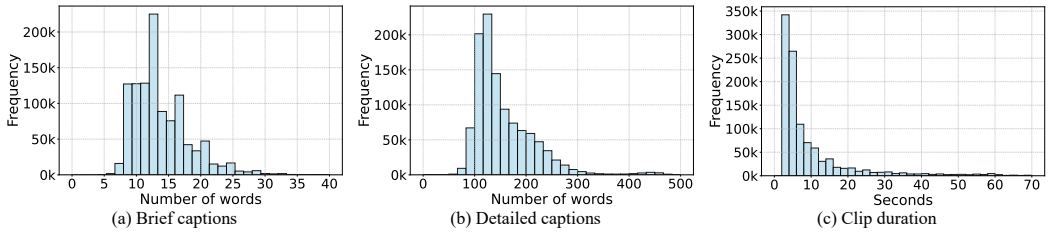
B The Prompt for Describing Videos

Prompt: Please describe the content of this video in as much detail as possible, including the objects, scenery, animals, characters, and camera movements within the video. Do not include \n in your response. Start the description with the video content directly. Describe the content of the video and the changes that occur, in chronological order.

C Ablation Studies on Curation of VideoUFO

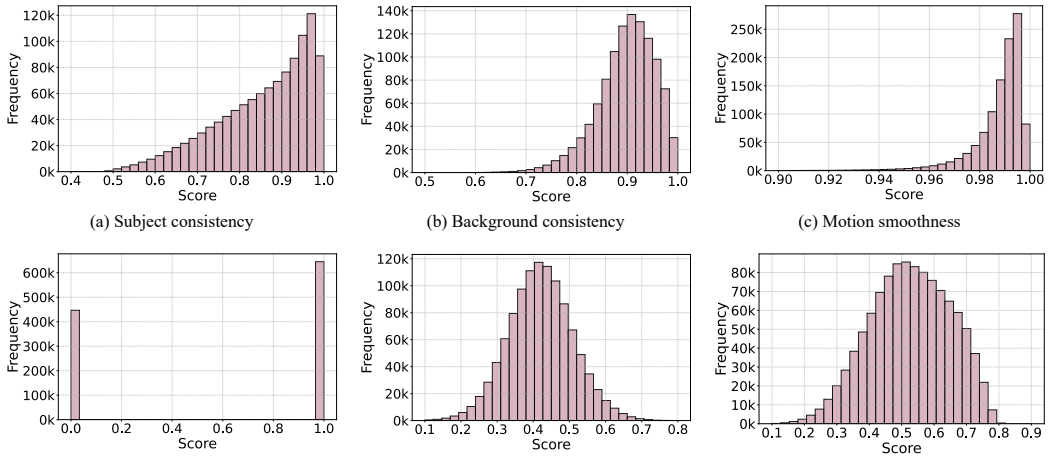
This section performs ablation studies to investigate key components in curating VideoUFO, *i.e.*, examining (1) the impact of the number of clusters used in user-focused topic analysis and (2) the methods used for clip verification.

The number of clusters. As shown in Fig. 9, we vary the number of pre-set clusters from 100 to $3,000$ when analyzing the focused topics of text-to-video users. We observe that initially, as the number of pre-set clusters increases, the number of resulting topics also increases; however, after reaching approximately $2,000$ clusters, this increase stops and begins to fluctuate. Therefore, we finally choose $2,000$ as the number of pre-set clusters. This is reasonable because the number of user-focused topics is limited, and an excessive number of pre-set clusters would eventually merge together. Beyond this, we also experiment with DBSCAN and HDBSCAN [47] to automatically determine the number of clusters; however, we do not obtain reasonable results.



(a) Brief captions (b) Detailed captions (c) Clip duration

Figure 10: The statistical information of captions and video clips in the proposed VideoUFO dataset. The average word count for brief and detailed captions is $\mathbf{13.8}$ and $\mathbf{155.5}$, respectively, while the average clip duration is $\mathbf{12.6}$ seconds.



(a) Subject consistency (b) Background consistency (c) Motion smoothness

(d) Dynamic degree (e) Aesthetic quality (f) Imaging quality

Figure 11: The statistical distributions of scores assessed by these six video quality metrics, which have been aligned with human perception. Researchers can use these scores to filter videos according to their specific training needs.

The method for clip verification. We randomly select $1,000$ video clips to conduct ablation studies on the methods used to verify whether a clip contains a specific topic. Feeding the entire video into GPT-4o (denoted as "Direct GPT-4o") is considered the most accurate approach, and we compare the overlap of other cost-effective methods with it. As shown in Table 2, we observe that: **(1)** Feeding videos directly into GPT-4o is very expensive – processing $1,000$ clips costs $\$162.23$. **(2)** Using detailed captions instead of video clips reduces the API cost by approximately $5,400\times$, while maintaining about 80% prediction overlap. **(3)** Although using brief captions is even cheaper, the overlap drops by 14.9%, which is reasonable since a brief caption cannot capture all the information in a clip. **(4)** Interestingly, we find that GPT-4o-mini achieves slightly better performance than GPT-4o. **In conclusion**, for verifying clips, we choose detailed captions with GPT-4o-mini.

D The Distributions of Caption Lengths and Clip Durations

Fig. 10 presents the statistical distributions of caption lengths (for both brief and detailed captions) and clip durations.

E The Distributions of Video Quality Score

Fig. 11 shows the statistical distributions of scores evaluated in terms of *subject consistency*, *background consistency*, *motion smoothness*, *dynamic degree*, *aesthetic quality*, and *imaging quality*.

F Broader Impacts

The open, topic-rich dataset lowers the barrier for global researchers to experiment with text-to-video generation. Its accessibility can spark new tools for education, creativity, and assistive technologies.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We follow this common practice for writing a paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See the Limitation at the end of Sections 4 and 5.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: This is not a theory paper.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All the information needed to reproduce the main experimental results are provided in the main paper as well as the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The dataset is publicly available at https://huggingface.co/datasets/WenhaoWang/VideoUFO, and the code is publicly available at https://github.com/WangWenhao0716/BenchUFO.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental settings and details are provided in the main paper as well as the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: All the experimental results are significant and stable, and error bars are not reported because it would be too computationally expensive.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We include the experiments compute resources in each related section.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We follow the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: For the positive social impacts, please refer to Section F in the Appendix. The data used in this study is already publicly available on YouTube, and we do not anticipate any negative societal impacts resulting from the release of our dataset.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [Yes]

    Justification: We only scrape data with explicit CC BY 4.0 License to avoid any potential copyright infringement.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite all the assets used in our paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Please see the README file in the released dataset.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve that.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve that.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: We only use LLM for writing, editing, or formatting purposes.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.