
T2V-Bench: Benchmarking Text-to-Video Models

Angel Bujalance¹ Amina Izbassar¹ Jakub Tomaszewski¹ Despoina Touska¹

Abstract

In this paper we introduce a novel, AI-based method of assessing the performance of generative text-to-video models. Despite the existence of numerous benchmarks for text-to-image generative models, the domain of evaluating text-to-video models lacks reliable automated evaluation methods. To this end, we propose a comprehensive method of evaluating the synthetic videos on general alignment with the provided text prompt, as well as through a meticulous analysis of the individual objects within the generated video frames. We conduct experiments with seven different text-to-video models and evaluate their performance using our proposed benchmark across 5 different tasks. The results indicate a frequent inability of T2V models to correctly generate a particular number of objects specified in the prompt. One the other hand, those models demonstrate a satisfying performance in the tasks of color identification and attribution. The official code base developed through this research is available on our [GitHub repository](#).

1. Introduction

The recent emergence of generative models has marked a significant milestone in the field of machine learning. Numerous state-of-the-art Large Language Models (LLMs) became successful at tackling complex, everyday tasks, due to their ability to adhere to prior instructions. Specifically, GPT-4 was found to outperform most humans on academic and professional tasks ([OpenAI, 2024](#)). Therefore, since those models are becoming accessible to the general public, their societal impact intensifies. [Eloundou et al. \(2023\)](#) suspects that 19% of jobs have at least a 50% chance of being affected by LLM given current capabilities and anticipated ones. Given this scenario, the outputs generated from these models are required to be truthful and reliable to be deployed in real-world applications. According to [Ji](#)

[et al. \(2023\)](#), most LLMs suffer from hallucinations, which give the user an impression of a model being fluent, knowledgeable, and providing him with a natural answer, that may not necessarily be accurate or correct. Furthermore, detecting most hallucinations is considered a demanding or on some cases an impossible task. This poses numerous challenges, particularly for domains like healthcare, in which such inaccurate generations may be detrimental and have far-reaching consequences. All the aforementioned highlight the critical need for reliable and realistic outputs across all generative models, including the rapidly developing field of Large Vision-Language Models (LVLMs).

Drawing inspiration from the advancements in LLMs, LVLMs have also seen substantial progress. Pioneering works like GPT-4v ([Achiam et al., 2023](#)), Gemini-Pro-V ([Team et al., 2023](#)), and LLaVA ([Liu et al., 2024](#)) have pushed the boundaries of image content recognition and reasoning, outperforming their predecessors. This trend continues with Text-to-Image (T2I) and Text-to-Video (T2V) models. These models, such as DALL-E ([Betker et al., 2023](#)) and CogVideo ([Hong et al., 2022](#)), have not only enriched the capabilities of generating high-fidelity images and videos from textual descriptions but have also opened new avenues for creative and practical applications. For this reason, ensuring their robustness and reliability through evaluation is crucial for both the research and business community.

While significant research has addressed the performance of LLMs ([Li et al., 2023a; Lin et al., 2021](#)), multimodal LLMs ([Li et al., 2023b; Guan et al., 2024; Tong et al., 2024](#)), and T2I models ([Lee et al., 2023; Ghosh et al., 2023](#)), a critical gap exists in the evaluation of T2V models. This paper aims to bridge this gap by providing a comprehensive evaluation of T2V model performance, thereby contributing to the enhancement of their reliability and the expansion of their application spectrum.

We decided to evaluate the models using an automated evaluation method. This approach ensures consistency, as it eliminates the intrinsic differences between human evaluators and saves time for researchers. Additionally, human evaluation can be influenced by the evaluator's expectations ([Holman et al., 2015](#)) and is often biased by previous beliefs or political views ([MacCoun, 1998; Nickerson, 1998](#)).

¹University of Amsterdam, Netherlands..



Figure 1. T2V models suffer from different types of hallucinations. From left to right: (1) Given the prompt "A cat playfully chases a pink ball of yarn across a blue living room rug", there is no cat on the rug. (2) "A family of four enjoys a movie night, sitting on a brown leather couch with a bowl of popcorn", the depicted people are more than four. (3) "A woman in a yellow raincoat walks a white dog through a forest", there is more than one dog. (4) "A baker with a white apron carefully decorates a chocolate cake with red frosting", the model attributes the red color to many other objects besides the frosting. (5) "A cat playfully chases a pink ball of yarn across a blue living room rug", the model attributes the pink color to many other objects besides the ball of yarn.

Moreover, a study conducted by [Clark et al. \(2021\)](#) found that human evaluators struggled to distinguishing between text generated by humans, GPT2 and GPT3. Even with training, their performance did not significantly improve. These results suggest that humans may be unable to reliably differentiate between outputs from state-of-the-art models and may make arbitrary judgments when evaluating them.

Addressing the limitations of current evaluation methods, namely the lack of automated evaluation metrics in temporal consistency of generated videos, we propose T2V-Bench, a systematically designed objective evaluation benchmark designed to assess various capabilities of Text-to-Video (T2V) models ([Liu et al., 2023](#)). T2V-Bench employs a two-stage pipeline including evaluation of five different abilities, focusing on perception and reasoning. In the first stage, perception is evaluated by verifying whether the generated videos satisfy the semantic requirements specified in the initial prompts used to create them. On the other hand, the second stage assesses the model’s reasoning capabilities including object count and color identification as well as attribution. This stage examines if specific video features, such as object color and quantity, align with the prompts initially provided to the model. Each ability dimension is evaluated on 50 video outputs generated by each of the seven chosen T2V models. We also offer a performance ranking that allows direct comparison between models and provides valuable insights for future optimization.

Our main contributions can be summarized as follows:

- **Two-stage evaluation framework:** We introduce T2V-Bench, a systematic two-stage evaluation framework designed to assess the capabilities of T2V models on different levels of granularity.
- **Evaluation on a comprehensive dataset:** We experiment with an extensive dataset specifically designed for T2V model evaluation. The data (Figure 1) contains

700 videos and covers several input prompt variations.

- **Valuable Insights:** We conduct a comprehensive evaluation of the T2V models. The resulting data and observations provide valuable insights for researchers, helping to guide future advancements in T2V model development.

2. Related Work

Previous research has extensively focused on evaluating Large Vision Language Models (LVLMs), which aim to enhance vision-language models by incorporating Large Language Models (LLMs) instead of traditional language encoders ([Li et al., 2023b](#)). While LVLMs outperform traditional models in visual reasoning, they are prone to hallucinations, a trait inherited from LLMs ([Guan et al., 2024](#)).

[Guan et al. \(2024\)](#) evaluate these shortcomings using the Hallusionbench benchmark, which includes 455 visual-control pairs, 346 images, and 1129 handcrafted question-answer pairs related to various types of visual information (e.g., charts, maps, figures, tables). This work analyzes the importance of visual information over textual information, addressing a primary concern: the conspicuous effect of language bias. For this purpose the visual-control pairs assess LVLM responses with and without visual information. The study evaluates 13 LVLMs by presenting images and questions, then assessing the correctness of binary (yes/no) answers. It was found that when models had prior knowledge related to the questions, language bias was the predominant source of information. In the absence of prior knowledge, LVLMs remained prone to hallucinations.

Similarly, [Li et al. \(2023b\)](#) focus on hallucinations arising from visual instructions in their POPE benchmark. They generate ground truth objects in images using human annotations or a segmentation tool (SEEM) and prompt models with questions about these objects and their top-k co-

occurring objects. Their findings indicate that LVLMs tend to hallucinate objects that frequently co-occur in visual instructions during training. Additionally, Tong et al. (2024) evaluate LVLMs’ question-answering abilities using CLIP-blind pairs of images with different DINOv2 embeddings, further emphasizing the visual aspect of LVLM performance.

While the aforementioned studies provide a holistic evaluation of LVLMs, other works offer a more fine-grained approach. The SEED-Bench benchmark by Li et al. (2023a) evaluates model performance across various evaluation dimensions such as scene understanding, instance identity, and spatial relations. This benchmark reduces human supervision by using a pipeline where foundation models describe image aspects with text, and a GPT-4 model generates multiple-choice questions based on these descriptions. In turn, the generated questions along with the images, are used to evaluate LVLMs. The paper also proposes three evaluation dimensions for video evaluations: action recognition, action prediction, and procedure understanding, though these are less fine-grained compared to evaluation dimensions pertaining to image understanding.

Lastly, Lee et al. (2023) and Ghosh et al. (2023) focus on evaluating text-to-image models compared to the previous papers that evaluated LVLMs. While the Gen-Eval benchmark introduced by Lee et al. (2023) is more object oriented, assessing the generated image based on the presence of the objects mentioned in the prompt, correct assignment of their attributes and their position, the HEIM benchmark proposed by Lee et al. (2023) evaluates text-to-image models in a more holistic manner, focusing on aspects such as image quality, originality, bias, and overall alignment with the prompt.

This body of work is significant to our project as the metrics used to evaluate text and image understanding in LVLMs and MLLMs can inform the metrics necessary for evaluating the alignment of videos generated by text-to-video models with the text prompts used to generate them.

3. Methodology

In this section, we present the details of the recipe for building T2V-Bench, including dataset creation, selection of T2V models, and the two-stage evaluation framework. A visualization of the pipeline followed is depicted in Figure 2.

3.1. Dataset Preparation

To facilitate robust evaluation of text-to-video models, we constructed a comprehensive dataset, tailored to a wide range of assessment tasks targeting the generative capabilities of these models.

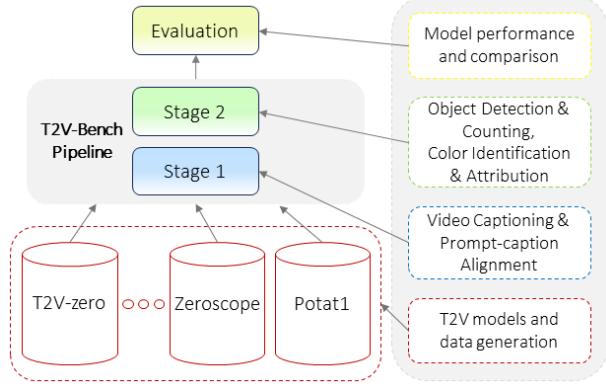


Figure 2. T2V-Bench pipeline



Figure 3. An example of a prompt along with its metadata for color and foreground objects

The dataset consists of a total of 700 two and five second videos, generated across various scenarios. These videos were created from seven distinct T2V models by presenting them with 50 meticulously selected prompts designed to evaluate different aspects of the T2V models. Each prompt in the dataset is augmented with detailed metadata, crucial for tasks such as object detection and color identification. For instance, the example prompt "A woman with long black hair reads a red book on a cosy couch" (see Figure 3) includes metadata specifying the number and type of foreground objects (person, book, couch), as well as the color information for the required objects (red book, black hair). This rich metadata enables an extensive and more targeted evaluation of state-of-the-art T2V models.

3.2. Video Generators

Multiple generators based on the Stable Diffusion model and a zero-shot text-to-video model, were employed in the process of generating the dataset videos. The selection was specifically comprised of the seven most popular and publicly available video diffusion models. This in-

cludes the Text-to-Video-synthesis¹ (T2V-ms-1.7b) (Wang et al., 2023a) model being a T2V generation model that uses spatio-temporal blocks to conduct frame generation with temporal consistency and smooth motions. The model consists of three sub-networks: text feature extraction model, text feature-to-video latent space diffusion model, and video latent space-to-video visual space model. Following a similar generation pipeline, zeroscope² and potat1³ consider more fine-tuning strategies and other generation configurations, such as spatial resolutions. Specifically, zeroscope is a watermark-free Modelscope-based model and has been trained on specific sizes such as 576x320 and 1024x576. For our experiments, we utilized the lower-resolution checkpoint. Furthermore, a zero-shot T2V⁴ generation method (Khachatryan et al., 2023) is also considered, which utilizes the pre-trained image diffusion model and keeps the temporal consistency with motion dynamics. This adaptation allows these models to be utilized within the video domain. As the baselines for the generation results we adopted the Stable Diffusion v1.5 (SD-1.5) (Rombach et al., 2022) and XL 1.0-base (SDXL-1.0b) (Podell et al., 2023) models.

Another category of models we employed leverages the AnimateDiff architecture (Guo et al., 2024). These models empower the creation of videos using existing Stable Diffusion T2I models. It accomplishes this by integrating motion module layers within a pre-trained T2I model (frozen). The motion modules are strategically inserted after the ResNet and Attention blocks within the Stable Diffusion UNet architecture. Specifically, we utilized two variations of AnimateDiff: AnimateDiff-Lightning⁵ and AnimateDiff-MotionAdapter⁶.

3.3. Two-Stage Evaluation Framework

In order to ensure the robustness of our benchmark we first identify the tasks in which most of the state-of-the-art T2V generation models fall short. Those mainly include prompt comprehension and can be divided into five distinct ability dimensions namely, object identification, object counting, color identification, color attribution, and general video consistency. To account for those major shortcomings we design an end-to-end, two-stage evaluation pipeline, that evaluates various aspects of the model’s generation capabilities.

¹http://huggingface.co/docs/diffusers/main/en/api/pipelines/text_to_video

²http://huggingface.co/cerspense/zeroscope_v2_576w

³<http://huggingface.co/camenduru/potat1>

⁴http://huggingface.co/docs/diffusers/api/pipelines/text_to_video_zero

⁵<http://huggingface.co/ByteDance/AnimateDiff-Lightning>

⁶<http://huggingface.co/guoyww/animatediff-motion-adapter-v1-5-2>

3.3.1. FIRST-STAGE EVALUATION

The first stage of our evaluation framework is a holistic evaluation of the general video-prompt consistency. This stage consists of two steps: generating a caption from the synthetic, input video and measuring how aligned the generated caption is with the initial prompt (see Figure 4). The intuition behind this AI-based approach, is in the assumption that if an AI model is capable of captioning the video in a way that an LLM identifies the caption as similar to the prompt used for video generation, the generated video contains the general information from the prompt.

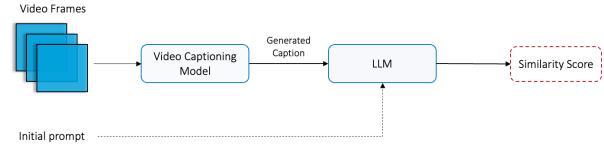


Figure 4. First stage of the evaluation pipeline

Video Captioning: Captions for each video are generated by feeding the video frame sequence to a generative image-to-text model. Subsequently, these captions, together with the prompts, are fed into an LLM, which assesses their mutual similarity in terms of the appearance of distinct objects and their characteristics, as well as the general message the captions convey. Naturally, the generated captions would deviate from the generative prompts, however as mentioned before, if they can be recognized as having a similar meaning by the LLM used for prompt-caption alignment, the example will be considered as positive.

Prompt-caption Alignment: One of the shortcomings of the Video Captioning step of this stage is that the generated captions were shorter than the initial prompt due to the generated videos being less detailed than described in the prompt, as well as the limitations of the captioning model. To account for this, the general prompt-caption alignment was evaluated by an LLM instead of the initial sentence-transformer. By utilizing an LLM for evaluating alignment, we also aim to increase the significance of caption content compared to caption length. In contrast to the sentence transformer, which produced a similarity score, the LLM was prompted to output a binary value (either 0 or 1). Though this is not as fine-grained as a similarity score for each prompt-caption pair, it is still representative of the alignment of the prompt with the caption, and it avoids the scenario where the LLM generates inconsistent similarity scores for different prompt-caption pairs.

3.3.2. SECOND-STAGE EVALUATION

The second stage of the pipeline evaluates the synthetic videos on a frame-based level. It comprises four reasoning

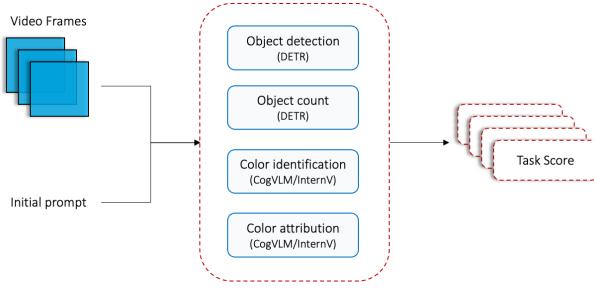


Figure 5. Second stage of the evaluation pipeline

tasks, including examining the number of objects appearing in each video frame and identifying their color. Such an approach strives to address the possible model hallucinations in the generated videos, as depicted in Figure 1. To this end, we employ the specific tasks of object detection and counting as well as color identification and attribution in our benchmark pipeline (see Figure 5).

Object Detection: One of the aspects that we considered that T2V models had to perform correctly was the ability to generate the objects appearing in the input prompt. To detect the different objects in the T2V-generated videos, we applied an object detector model to each of the frames in a video. We used the DETR (End-to-End Object Detection) model with ResNet-50 backbone available in the Hugging Face repository (Carion et al., 2020). Once the object detector model found all the objects in an image frame, we checked if all the objects mentioned in the prompt were present. For instance, given the prompt: “A woman in a yellow raincoat walks a white dog through a forest”, we want person and dog objects to be detected in the T2V-generated frame. If the object detector detects both objects, it will get the maximum accuracy for the frame. The formula used to compute the object detection score is:

$$ObjectDetection = \frac{\text{num objects detected}}{\text{total number of objects}} \quad (1)$$

where total number of objects is the number of objects in the metadata. The final score is the average of this score for all the frames.

Object Counting: When considering object counting, we wanted to test the ability of T2V models to generate the exact number of objects of a class. In this task, we want the model not only to generate the correct objects but also to ensure that each object appears the exact number of times specified in the prompt. For instance, if we ask a T2V model to generate “A group of five children fly red kites in a field of wildflowers.”, we will expect the generated video to have 5 kites and 5 children in all the frames. To check if the models can achieve this task, we counted the number of

times an object was detected by the DETR model (Carion et al., 2020). If the count of an object in the frame matched the number in the metadata of the prompt, we considered that the model was correct for this object. Finally, to get the Object Counting score we used the following formula:

$$ObjectCount = \frac{\text{num objects with correct count}}{\text{total number of objects}} \quad (2)$$

As in the previous case, the final score is the average for all the frames.

Color Identification: Another important ability that the T2V models should possess is the ability to identify the colors mentioned in the prompt and apply these colors to the specified objects. This dimension evaluates whether a T2V model can reason based on text information. It requires the model to fully understand the text and utilize its common-sense knowledge to correctly generate the specified colors in the objects within the video. The evaluation is conducted using LVLMs by asking three different questions related to color identification: (1) “What is the color of the **object**?", (2) “Find the color of the **object** in the image.” and (3) “If there is an **object** in the image, what is its color?”. This process is applied to every frame of a video. The accuracy is then calculated for all the colored objects in the video, averaged for the video, and subsequently, the mean accuracy is taken for each T2V model separately.

Color Attribution: A common behavior of T2V models is to apply the color included in the prompt to more objects than required. While this cannot be strictly labeled as a mistake, it suggests that the model may be uncertain about where to place the color. This dimension evaluates whether a T2V model colors more objects than required with the specified color in the prompt. The evaluation is conducted using LVLMs by asking the question: “Are there any **color** objects in the image other than the **object**. Include Yes or No in your answer and a small explanation”. This process is applied to every frame of a video. The accuracy is then calculated similarly to the color identification task.

4. Evaluation Results

4.1. Experimental Setup

All experiments were conducted on an NVIDIA A100 GPU with 40 GB HBM2 memory. For the color identification and attribution experiments, we used two LVLMs: the CogVLM (Wang et al., 2023b) and InternVL (Chen et al., 2023) models. The GIT (Wang et al., 2022) model by Microsoft and Llama-3-8b-instruct (AI@Meta, 2024) were used for generating video captions and evaluating prompt-caption similarity, respectively.

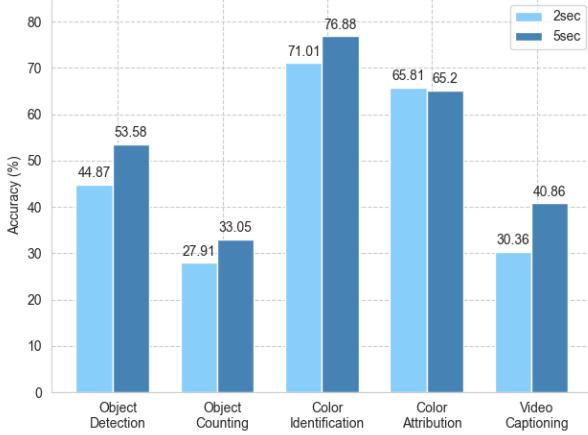


Figure 6. Accuracy for each task across the two different video durations

4.2. Main Results

Figure 6 presents the task performance averaged for all seven models. The plot includes different bars for the results of videos with durations of 2 and 5 seconds. Among the five tasks, object counting shows the lowest performance with only 27.91% accuracy for videos with a 2-second duration, indicating that T2V models struggle with generating the correct number of objects from the prompt. Similarly, while the average model performance in the object detection task appears to be substantially better, the overall results are still mediocre, reaching an accuracy score of around 50% for videos with a 5-second duration. On the other hand, all models demonstrate a satisfying and competitive performance for the color identification task, namely 71% for both 2- and over 75% for 5-second video durations. Analogously, the models achieve an acceptable score in the color attribution task (over 65%), suggesting that models are often influenced by a particular color appearing in the prompt to color objects other than those explicitly mentioned. Lastly, video captioning has relatively low performance, which may indicate that the models’ ability to generate high-quality, semantically accurate videos is low.

Furthermore, the generated five-second videos consistently outperform two-second videos across most metrics, with the notable exception of the color attribution task. This pattern likely reflects an inherent bias in Text-to-Video (T2V) models towards processing longer sequences, which could be attributed to the training datasets predominantly consisting of longer videos. Nonetheless, we leave the exploration of these potential patterns and their implications for future research.

4.2.1. ANALYTICAL RESULTS

In our experiments we evaluated seven different state-of-the-art T2V models on five different tasks. All the employed models were publicly available in the Hugging Face library. The results are depicted in Table 2, as well as in Figure 7, and Figure 8.

Video Captioning. The video captioning measures the similarity of the video with the prompt. We saw that accuracies for the 2- and 5- second videos, ranged between 6% and 52%. Only zeroscope² and T2V-ms-1.7b (Wang et al., 2023a) had scores above 50% in the 5-second duration videos. This poor accuracy highlights the difficulties of current models in consistently generating quality videos. See Table 2 for details.

Object Detection. The results for the object detection task are depicted in Figure 7 and in Figure 8. We note AnimateDiff-MotionAdapter (Guo et al., 2024) as the best-performing model, achieving a performance of approximately 70% for both 2- and 5-second duration videos. Additionally, Figures 12 and 13 in the Appendix section illustrate examples where the evaluated TV2 models correctly and incorrectly generated the objects in the prompt, respectively.

Object Counting. When evaluated if the model generated the exact number of objects as in the metadata of the prompt, the model’s accuracy decreased. No model surpassed 50% accuracy in this task. The model with the best performance was the SD-1.5 (Rombach et al., 2022) for the 2-second duration videos and the AnimateDiff-Lightning (Guo et al., 2024) for the 5-second duration videos.

Color Identification and Attribution Results In Table 1, the results for color identification and attribution are presented for each of the two LVLMs used in the experiments: CogVLM and InternVL. Using two LVLMs helps make the results more robust. According to the results, the models show similar scores for the color identification task. However, there is a significant difference in the answers provided by these two models in the color attribution task. Specifically, when CogVLM is asked to respond to the color attribution question, “Are there any **color** objects in the image other than the **object**? Include Yes or No in your answer and a small explanation,” it tends to find small details in the image rather than objects that have the desired color. One example of this behavior is presented in the Figures 14 in the Appendix section. This tendency results in very high scores for this task. On the other hand, InternVL is more accurate in identifying the desired color in objects rather than in background details, resulting in lower scores for this task. Lastly, the general trend that models perform better when generating 5-second videos is apparent here as well.

Our results indicate that model performance is highly task-specific. For instance, Zeroscope exhibited superior per-

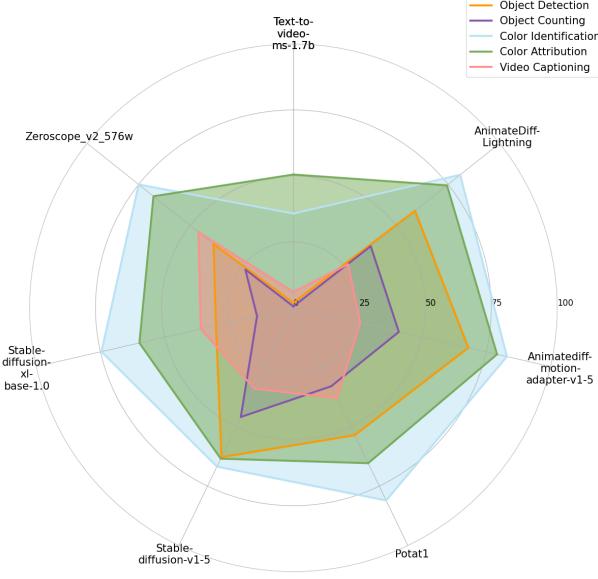


Figure 7. Evaluation results for videos with 2sec duration

formance in Video Captioning but was outperformed by AnimateDiff-MotionAdapter (Guo et al., 2024) in the Object Detection task. Similarly, AnimateDiff-Lightning (Guo et al., 2024) showed competitive results in the Object Counting of 5-second videos but did not excel in other tasks. These variations highlight the absence of a single model that consistently leads across all evaluated tasks.

T2V Model	Sec.	CogVLM		InternVL	
		Color Id.	Color Attr.	Color Id.	Color Attr.
Text-to-video-ms_1.7b	2	38.32	97.92	33.14	3.12
AnimateDiff_Light	2	80.42	100.0	81.33	48.92
AnimateDiff_ma_v1.5	2	83.29	99.6	82.73	58.77
Potat1	2	82.04	100.0	80.04	30.71
Sd_v1.5	2	70.76	95.75	62.83	31.17
Sd_xl_base_1.0	2	77.29	97.71	71.93	22.0
Zeroscope_v2_576w	2	75.83	100.0	74.17	35.67
Text-to-video-ms_1.7b	5	76.62	94.6	76.15	18.8
AnimateDiff_Light	5	76.64	97.35	77.97	30.37
AnimateDiff_ma_v1.5	5	83.23	99.5	82.91	58.42
Potat1	5	78.43	99.0	77.15	32.48
Sd_v1.5	5	76.24	94.55	71.65	25.28
Sd_xl_base_1.0	5	80.6	97.95	79.68	29.12
Zeroscope_v2_576w	5	71.49	99.75	67.55	35.58

Table 1. CogVLM vs. InternVL performance in color identification and attribution

5. Conclusion

In this work, we introduced T2V-Bench a 2-stage method for evaluating text-to-video generative models. Our benchmark evaluates T2V model capabilities across a number of object-oriented tasks using Vision-Language Models for object detection, color classification and color assignment to verify fine-grained object attributes. In addition, we mea-

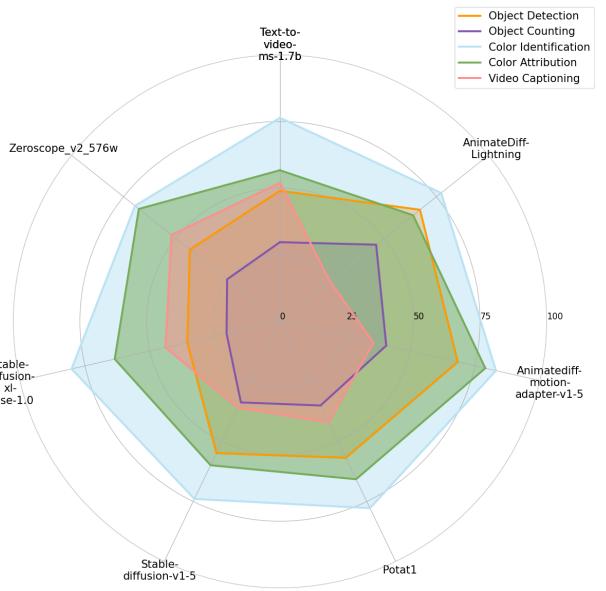


Figure 8. Evaluation results for videos with 5sec duration

sured coarse video quality by evaluating the similarity of the caption generated from the video with the original prompt. Our evaluation experiments conducted on 7 text-to-video models reveal that the video quality, considering the dimensions we evaluate on, does not decrease with the length of the generated video, despite the presence of a higher number of frames and thus a higher chance of error.

5.1. Limitations

Despite the robust framework of our text-to-video model benchmark, several limitations remain that could impact the generalizability and applicability of our findings. These limitations are outlined below and detailed further in our appendices.

Limited Object Classes in DETR. A fundamental constraint in our evaluation pipeline arises from the use of the DETR object detection model, as outlined in Appendix D. DETR, trained exclusively on the COCO dataset, is inherently limited to detecting only the object classes included in that dataset. This restricts the ability of our benchmark to evaluate scenarios involving objects outside the set of predefined classes, such as specific types of animals or unusual artifacts. Furthermore, DETR’s performance degrades with increased instances of objects within a frame, leading to missed detections. This limits the scope of videos the evaluation framework can be applied to.

Errors in Model Performance. As detailed in Appendix C, Appendix D, and Appendix E, each of the models integrated into our benchmark may sporadically produce errors, which can significantly affect the reliability of the evaluations. For

instance, in the color identification tasks discussed in Appendix E, the models occasionally failed to accurately identify or attribute colors in video frames. Such inaccuracies can undermine the utility of the benchmark in applications where precise color recognition is critical. Similarly, in Appendix C, we identified rare cases in which the utilized video captioning yielded an erroneous result.

Limited Number of Evaluation Tasks. The scope of evaluation tasks currently integrated into our benchmark is another limitation. While the modularity of our codebase allows for the potential expansion of evaluation tasks, at present, the benchmark only supports a limited, yet relevant, set of tasks tailored to the capabilities of existing models like DETR.

5.2. Future work

Future work should focus on integrating additional models that surpass the class limitations of DETR, enhancing error correction mechanisms, and expanding the variety and complexity of evaluation tasks to better meet the evolving demands of video content analysis. We encourage other researchers from the community to build upon our work and tailor the benchmark to their needs by extending the number of evaluation tasks in the pipeline.

References

- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- AI@Meta. Llama 3 model card. 2024. URL https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.
- Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al. Improving image generation with better captions. *Computer Science*. <https://cdn.openai.com/papers/dall-e-3.pdf>, 2(3):8, 2023.
- Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. End-to-end object detection with transformers. *CoRR*, abs/2005.12872, 2020. URL <https://arxiv.org/abs/2005.12872>.
- Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Muyan, Z., Zhang, Q., Zhu, X., Lu, L., et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.
- Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., and Smith, N. A. All that's 'human' is not gold: Evaluating human evaluation of generated text, 2021.
- Eloundou, T., Manning, S., Mishkin, P., and Rock, D. Gpts are gpts: An early look at the labor market impact potential of large language models, 2023.
- Ghosh, D., Hajishirzi, H., and Schmidt, L. Geneval: An object-focused framework for evaluating text-to-image alignment, 2023.
- Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., Manocha, D., and Zhou, T. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models, 2024.
- Guo, Y., Yang, C., Rao, A., Liang, Z., Wang, Y., Qiao, Y., Agrawala, M., Lin, D., and Dai, B. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning, 2024.
- Holman, L., Head, M. L., Lanfear, R., and Jennions, M. D. Evidence of experimental bias in the life sciences: Why we need blind data recording. *PLOS Biology*, 13(7):1–12, 07 2015. doi: 10.1371/journal.pbio.1002190. URL <https://doi.org/10.1371/journal.pbio.1002190>.
- Hong, W., Ding, M., Zheng, W., Liu, X., and Tang, J. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022.
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12), mar 2023. ISSN 0360-0300. doi: 10.1145/3571730. URL <https://doi.org/10.1145/3571730>.
- Khachatryan, L., Movsisyan, A., Tadevosyan, V., Henschel, R., Wang, Z., Navasardyan, S., and Shi, H. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15954–15964, 2023.
- Lee, T., Yasunaga, M., Meng, C., Mai, Y., Park, J. S., Gupta, A., Zhang, Y., Narayanan, D., Teufel, H. B., Bellagente, M., Kang, M., Park, T., Leskovec, J., Zhu, J.-Y., Fei-Fei, L., Wu, J., Ermon, S., and Liang, P. Holistic evaluation of text-to-image models, 2023.
- Li, B., Wang, R., Wang, G., Ge, Y., Ge, Y., and Shan, Y. Seed-bench: Benchmarking multimodal llms with generative comprehension, 2023a.

- Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., and Wen, J.-R. Evaluating object hallucination in large vision-language models, 2023b.
- Lin, S., Hilton, J., and Evans, O. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*, 2021.
- Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.
- Liu, Y., Li, L., Ren, S., Gao, R., Li, S., Chen, S., Sun, X., and Hou, L. Fetv: A benchmark for fine-grained evaluation of open-domain text-to-video generation, 2023.
- MacCoun, R. J. Biases in the interpretation and use of research results. *Annual review of psychology*, 49(1): 259–287, 1998.
- Nickerson, R. S. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175–220, 1998. doi: 10.1037/1089-2680.2.2.175. URL <https://doi.org/10.1037/1089-2680.2.2.175>.
- OpenAI. Gpt-4 technical report, 2024.
- Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023.
- Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695, June 2022.
- Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.
- Tong, S., Liu, Z., Zhai, Y., Ma, Y., LeCun, Y., and Xie, S. Eyes wide shut? exploring the visual shortcomings of multimodal llms, 2024.
- Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., Liu, Z., Liu, C., and Wang, L. Git: A generative image-to-text transformer for vision and language. Technical report, Microsoft, May 2022.
- Wang, J., Yuan, H., Chen, D., Zhang, Y., Wang, X., and Zhang, S. Modelscope text-to-video technical report. *arXiv preprint arXiv:2308.06571*, 2023a.
- Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., et al. Cogvlm: Visual expert for pretrained language models. *arXiv preprint arXiv:2311.03079*, 2023b.

A. Models Performance

In Table 2, the accuracy scores for each task for each model are presented in detail for both 2-second and 5-second video durations. These results are the same as those depicted in Figures 7 and 8, but in this case the reader can have a better view of the actual values.

Model	Vid. Captioning		Obj. Detect.		Obj. Count.		Color Id.		Color Attr.	
	2sec	5sec	2sec	5sec	2sec	5sec	2sec	5sec	2sec	5sec
Text-to-video-ms_1.7b (Wang et al., 2023a)	6.0	52.0	1.92	49.0	0.54	29.72	35.73	76.39	50.52	56.7
AnimateDiff_Light (Guo et al., 2024)	26.53	24.0	58.96	67.15	37.5	46.13	80.88	77.3	74.46	63.86
AnimateDiff_ma_v1.5 (Guo et al., 2024)	26.0	36.0	68.03	68.38	41.0	40.87	83.01	83.07	79.18	78.96
Potat1 ³	38.0	42.0	53.58	56.83	33.0	35.03	81.04	77.79	65.35	65.74
Sd.v1.5 (Rombach et al., 2022)	34.0	36.0	62.83	54.87	45.92	33.75	66.8	73.94	63.46	59.92
Sd.xl_base_1.0 (Podell et al., 2023)	36.0	44.0	30.08	35.63	14.12	20.53	74.61	80.14	59.85	63.53
Zeroscope_v2_576w ²	46.0	52.0	38.71	43.18	23.29	25.32	75.0	69.52	67.83	67.67

Table 2. Accuracy results for each task analytically

B. Dataset

In Figure 9, there are four more examples from the dataset that we created. For each example, the prompt used as input to a T2V model to generate the video is provided, along with some metadata regarding the color and the foreground objects.

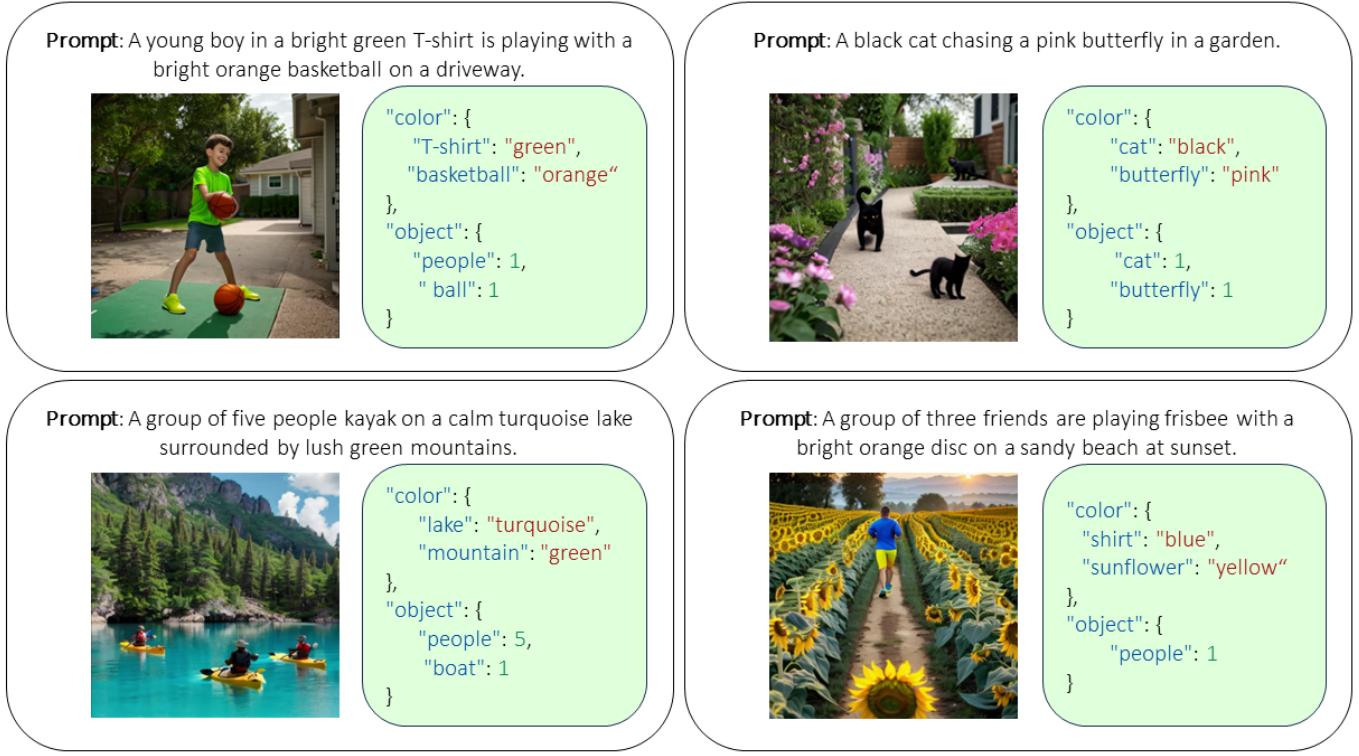


Figure 9. Examples of prompts and metadata along with the frames from the generated videos using the AnimateDiff motion adapter v1.5 model

C. Examples of first-stage evaluation

Figures 10 and 11 illustrate examples produced by the first stage of our evaluation pipeline, as well as possible errors that may sporadically occur during execution. See section 3.3.1 for the pipeline details.



Figure 10. Example cases where both the Text-to-Video (T2V) and video captioning models produced accurate outputs. In the highlighted examples, the T2V model successfully generated a video that aligned with the provided initial prompt, while the video captioning model produced a descriptive caption that corresponded to the video. The LLM subsequently evaluated the prompt and the resulting caption, confirming their consistency.

Figure 11. Examples of inaccuracies produced either by the Text-to-Video (T2V) or video captioning model. The upper example illustrates a case where the T2V model failed to produce a video corresponding to the given prompt, which was subsequently detected by the LLM. In contrast, the lower example depicts a rare error where the video captioning model generated an incorrect caption, despite the video itself being accurately rendered.

D. Examples of second-stage evaluation Object Detention and Counting tasks

The figures below show examples of how the Object Detection and Counting tasks were performed. We selected frames from the videos and compared the ground truth with the objects detected with the DETR model. Figure 12 shows frames where there is a correct match between the ground truth and the objects detected by the object detection model. This indicates that the T2V model successfully generated the proper number of objects.

On the other hand, in Figure 13 we depict examples of errors produced by the text-to-video generation model, being a mismatch between the ground truth prompt and the actual number of generated objects. Specifically:

- In Figure 13 A, one of the hikers is incorrectly generated by the T2V model as having two backpacks.
- In Figure 13 B, the generated video is missing a ball.
- In Figure 13 C, the model T2V generated three people instead of four.
- In Figure 13 D, there is one extra dog in the video frame.

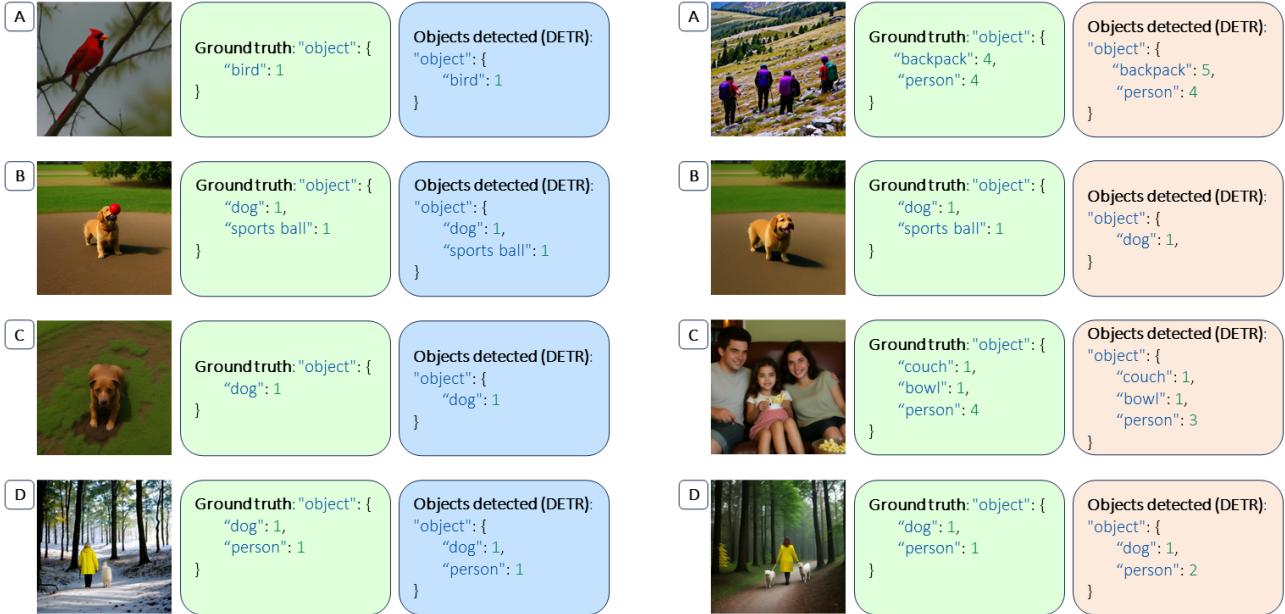


Figure 12. Example cases where the T2V correctly generates all the objects in the prompt with the correct number of instances. In this case, there is a match between the ground truth and the objects detected with the object detector model (DETR). Prompts: A) A bright red cardinal sings a cheerful melody from a branch of a green tree; B) A golden retriever playing fetch with a bright red ball in a sunny park; C) A brown dog is digging a hole in a green garden; D) A woman in a yellow raincoat walks a white dog through a forest.

Figure 13. Example cases where the T2V misses the generation of objects in the prompt. In this case, there is a mismatch between the ground truth and the objects detected with the DETR model. Prompts: A) A group of four hikers with purple colored backpacks climb a rocky mountain path, overlooking a vast green valley; B) A golden retriever playing fetch with a bright red ball in a sunny park; C) A family of four enjoys a movie night, sitting on a brown leather couch with a bowl of popcorn; D) A woman in a yellow raincoat walks a white dog through a forest.

E. Examples of second-stage evaluation: Color Identification and Attribution tasks

The figures below illustrate example frames along with questions prompted to each of the two LVLMs. Figure 14 demonstrates an example of CogVLM failing to identify that the only red object is the red cardinal given the prompt question, "Are there any red objects in the image other than a cardinal? Include Yes or No in your answer and a small explanation". Instead, it appears to detect tiny red details in the image. This is why the scores are very high for CogVLM in color attribution. On the other hand, InternVL correctly identifies only the red bird in the image when prompted with this question. Figure 15 shows an example where a T2V model fails to color the backpacks orange. Instead, it paints them red, which was correctly identified by the two LVLMs.

Prompt: A bright **red** cardinal sings a cheerful melody from a branch of a green tree



Prompt: Please briefly answer: What is the color of the cardinal in the image?
InternVL: The cardinal in the image is red.
Prompt: Please briefly answer: Find the color of the cardinal in the image.
InternVL: Red.
Prompt: Please briefly answer: If there is a cardinal in the image, what is its color?
InternVL: Red.
Prompt: Are there any red objects in the image other than a cardinal. Include Yes or No in your answer and a small explanation.
InternVL: No, there are no other red objects in the image besides the cardinal. The cardinal is the only object with a red hue in the image.

Prompt: A group of four friends hiking through a mountainous terrain wearing bright **orange** backpacks.



Prompt: Please briefly answer: What is the color of the backpack in the image?
InternVL: The backpack in the image is red.
Prompt: Please briefly answer: Find the color of the backpack in the image.
InternVL: Red.
Prompt: Please briefly answer: If there is a backpack in the image, what is its color?
InternVL: Red.
Prompt: Are there any orange objects in the image other than a backpack. Include Yes or No in your answer and a small explanation.
InternVL: No, there are no orange objects in the image other than a backpack. The image shows a group of hikers on a snowy mountain, and the only orange object visible is a backpack worn by one of the hikers.

Prompt: Please briefly answer: What is the color of the cardinal in the image?
CogVLM: The cardinal in the image is red.
Prompt: Please briefly answer: Find the color of the cardinal in the image.
CogVLM: The color of the cardinal in the image is red.
Prompt: Please briefly answer: If there is a cardinal in the image, what is its color?
CogVLM: Let's start by observing the image closely. We can see a bright red bird perched on a branch. Given the context and the color of the bird, it is likely a cardinal.
Prompt: Are there any red objects in the image other than a cardinal. Include Yes or No in your answer and a small explanation.
CogVLM: Yes, there are red objects in the image other than the cardinal. The cardinal's beak and eyes are also red. Additionally, there are red berries on some of the branches.

Prompt: Please briefly answer: What is the color of the backpack in the image?
CogVLM: The backpacks in the image are primarily red
Prompt: Please briefly answer: Find the color of the backpack in the image.
CogVLM: The backpacks in the image are primarily red.
Prompt: Please briefly answer: If there is a backpack in the image, what is its color?
CogVLM: Starting from the leftmost person: The first backpack is red. The second person has a blue backpack. The third person has a red backpack. The fourth person also has a red backpack.
Prompt: Are there any orange objects in the image other than a backpack. Include Yes or No in your answer and a small explanation.
CogVLM: Yes, there are orange objects in the image other than a backpack. Starting from the top, there's an orange jacket worn by the person in the middle. Moving downwards, there are orange trekking poles held by the individuals. Lastly, at the bottom, there's an orange piece of equipment or bag.

Figure 14. Example case where the CogVLM fails in the task of color attribution to identify only the color of the foreground objects

Figure 15. Example case where a T2V model fails to create the correct color in the backpacks which was correctly identified by the LVLMs.