Divot: Diffusion Powers Video Tokenizer for Comprehension and Generation

Supplementary Material

6. Implementation Details

6.1. Divot Tokenization.

Model Architecture. The Divot tokenizer is composed of a pre-trained ViT-H/14, a Spatial-Temporal Transformer and a Perceiver Resampler. Specifically, given a video clip with a duration of two seconds, we sample 5 frames at 2 fps, which are fed into the ViT to extract frame-level features. Subsequently, the extracted frame-level features are fed into the Spatial-Temporal Transformer, which consists of a 6-layer temporal transformer for temporal fusion, average pooling with a pool size of 5, and a 4-layer transformer for spatial and temporal fusion. To reduce the number of video tokens, these features after the Spatial-Temporal Transformer are further fed into the Perceiver Resampler, which contains 6-layer Perceiver Attention [1], for obtaining the final 64 video tokens. We adopt the de-noising U-Net in DynamiCrafter [78] as the de-tokenizer, but reduce the input channel of the 3D convolution from 8 to 4 since we remove the original concatenation of a conditional image with noisy latents.

Training Pipeline. Since the original DynamiCrafter concatenates the conditional image with per-frame initial noise and feeds them to the denoising U-Net as a form of guidance, it cannot be directly applied to video representation learning due to its extra dependence on low-level image inputs. To address this, we first fine-tune the pre-trained DynamiCrafter by removing the concatenation of the conditional image. This modification makes the model utilize only the image and caption features, along with temporal embeddings, as the sole conditions for denoising the noisy video clips. Then we replace the image and caption features with spatiotemporal representations produced by Divot tokenizer as the conditions, and train the Divot tokenizer and the denoising U-Net in an end-to-end manner with v prediction for denoising. After this stage, to further enhance the generation quality of our de-tokenizer, we freeze the Divot tokenizer and only fine-tune the denoising U-Net. During this fine-tuning process, we introduce a probability of 5% to drop the conditions, enabling us to leverage classifier-free guidance during inference. Note that in previous stage for optimizing the Divot tokenizer, we do not drop conditions to ensure that the denoising process fully relies on the spatiotemporal representations to optimize representations.

Training Data. The Divot tokenizer is trained on pure videos of a subset of WebVid-10M [2] and Panda-70M [9], totaling 10M videos. For WebVid-10M dataset, we employ LLaMA-3 to filter out videos with captions that do not contain dynamic content, resulting in a refined dataset of 4.8 million videos. For Panda-70M dataset, we download a total of 5.3 million videos, all of which are utilized for training purposes.

6.2. Pre-training and Instruction Tuning.

Pre-training. Divot-LLM adopts next-word prediction and GMM modeling on video-text data for video comprehension and generation during pre-training. Specifically, the video features from the Divot tokenizer, the special tokens indicating the start and end of video features, along with the text tokens of the caption are fed into the pre-trained Mistral-7B [24] for next token prediction trained with cross-entropy loss. Two fully-connected layers are trained to align the dimensions of the Divot features with those of the LLM. For GMM Modeling, text tokens of the caption and N learnable queries are input into the LLM, where the output of the learnable queries are fed into two fully-connected layers to predict $2kd + k$ parameters per video token (kd mean and kd variance parameters for the mixture components, and k mixture probabilities). We adopt $k = 16$ in our experiment. We utilize bidirectional attention for N learnable queries within the LLM and optimize the model using NLL loss.

Instruction Tuning. We perform multimodal instruction tuning on Divot-LLM to align it with human instructions through supervised fine-tuning on public datasets as listed in Tab. 2. We fine-tune a LoRA module on the pre-trained Divot-LLM with the template as below,

$$\text{[INST]} \quad \langle\text{Instruction}\rangle \quad \text{[/INST]} \quad \langle\text{Answer}\rangle \tag{2}$$

We further fine-tune the pretrained Divot-LLM on an animated series called "Curious George" to achieve video storytelling, which generates storyline and corresponding video clips in an interleaved manner. Specifically, after downloading the videos of "Curious George" series, we adopt the video splitting algorithm in Panda-70M to cut a long video into several semantically coherent clips including splitting based on shot boundary detection, and stitching based on semantics similarity. Subsequently, we employ GPT-4V to generate captions for each video clip by uniformly sampling eight frames from each clip. Finally, we use GPT-4 to summarize the instructions and corresponding storylines based on the captions of three consecutive video clips.

After instruction tuning, to further enhance the quality of video generation, we adopt a de-tokenizer adaptation technique, which fine-tunes the de-tokenizer based on the features sampled from the predicted GMM distribution derived from the LLM output.

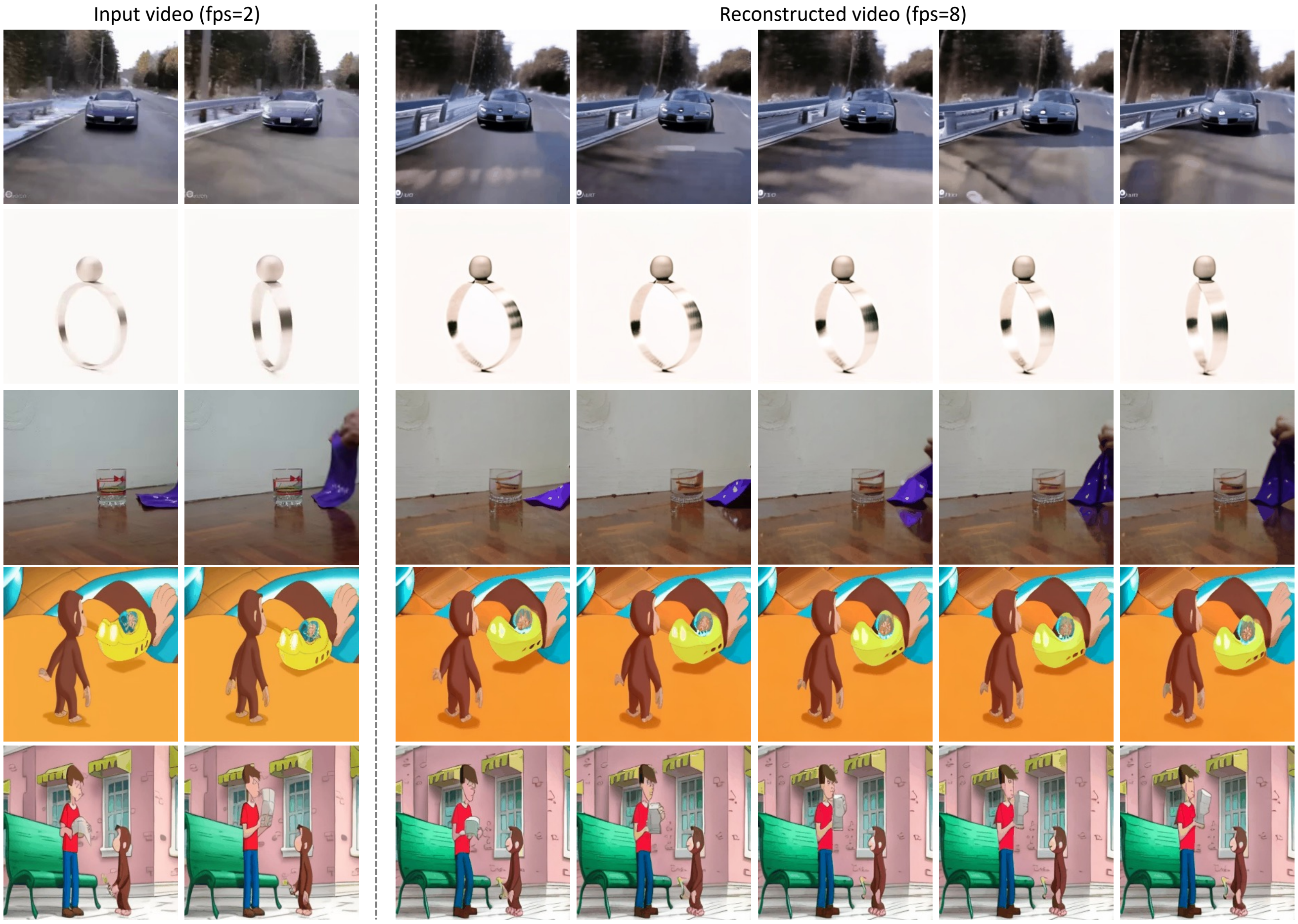


Figure 8. More qualitative examples of reconstructed videos, where the Divot tokenizer obtains spatiotemporal representations of sparsely sampled video frames and the de-tokenizer decodes these representations into semantically aligned and temporally coherent video clips.

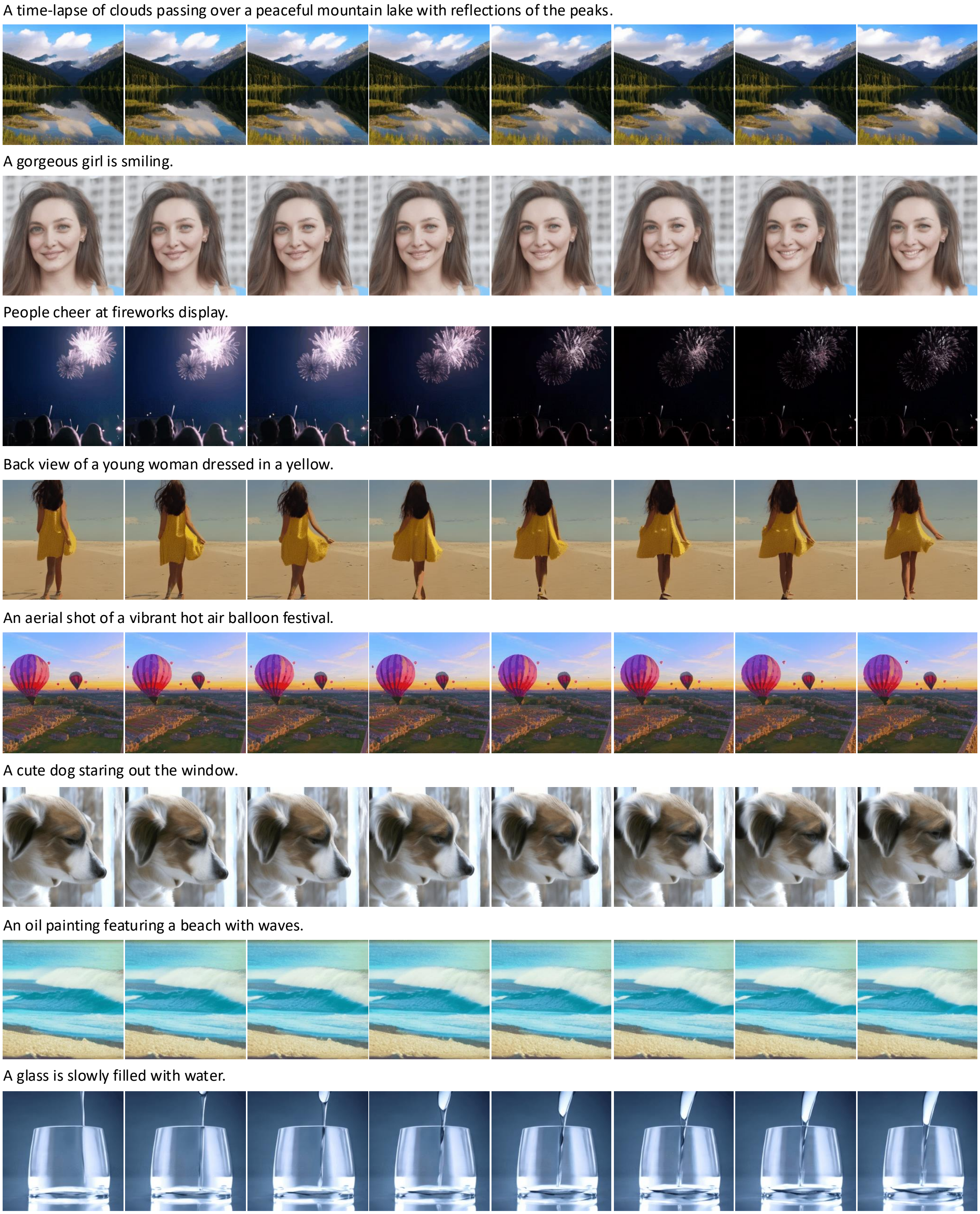
7. Qualitative Examples

Video Reconstruction. We provide additional qualitative examples of video reconstruction in Fig. 8, where the spatiotemporal representations are obtained from the Divot tokenizer and subsequently fed into the denoising U-Net to denoise realistic video clips from noise. The decoded video clips, generated from the learned spatiotemporal representations, exhibit semantic alignment with the original videos and maintain temporal coherence. For the adaptation to the animated series ”Curious George,” we fine-tune only the de-tokenizer while keeping the Divot tokenizer frozen. The satisfactory reconstruction results demonstrate the generalizability of our Divot tokenizer in obtaining robust video representations.

Video Generation. We present more qualitative examples of text-to-video generation in Fig. 9. Through modeling the distributions of Divot features with a GMM and training the LLM to predict GMM parameters, our Divot-LLM can generate videos that are both semantically aligned with text prompts and temporally coherent across frames. This is achieved through the dual-function de-tokenizer, utilizing only 4.8 million video-caption pairs for training.

Video StoryTelling. We provide more qualitative examples of video storytelling in Fig. 10. Given a brief story instruction, after instruction tuning, our Divot-LLM can generate a sequence of multimodal stories that feature rich narrative text alongside contextually relevant videos, all while maintaining temporal coherence.

Video Comprehension. As illustrated in Fig. 11, we provide qualitative examples to demonstrate the video comprehension capability of Divot-LLM. It can effectively understand sequences of events depicted in a video, reason using common sense, track and summarize the outcomes of specific actions or events, and deliver comprehensive and detailed descriptions of the videos. By utilizing diffusion procedure for video representation learning, our Divot tokenizer effectively captures robust spatiotemporal representations, enhancing the comprehension capabilities of Divot-LLM.



A time-lapse of clouds passing over a peaceful mountain lake with reflections of the peaks.

A gorgeous girl is smiling.

People cheer at fireworks display.

Back view of a young woman dressed in a yellow.

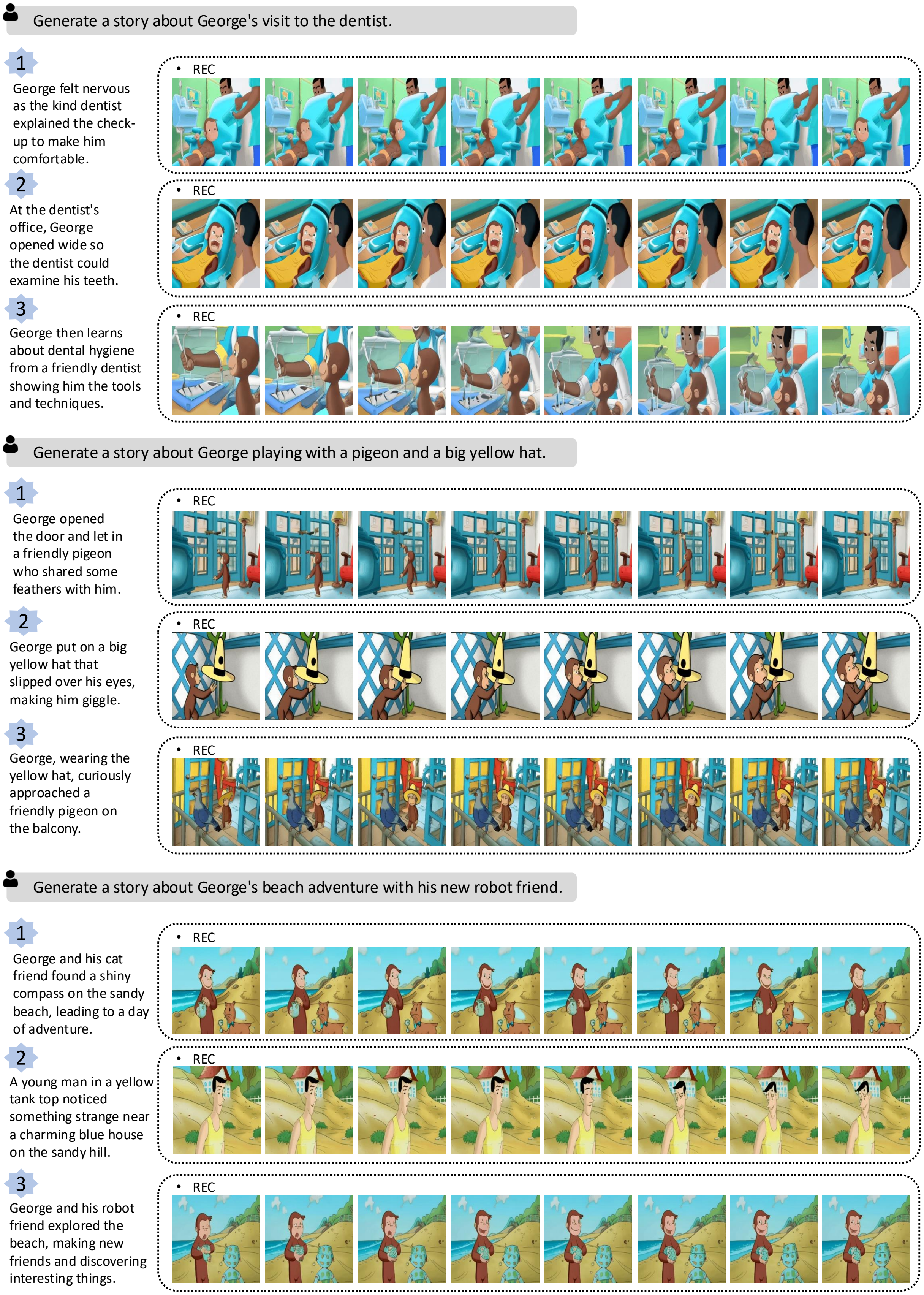
An aerial shot of a vibrant hot air balloon festival.

A cute dog staring out the window.

An oil painting featuring a beach with waves.

A glass is slowly filled with water.

Figure 9. More qualitative examples of text-to-video generation by Divot-LLM, which effectively generates videos that are both semantically aligned with text prompts and temporally coherent across frames.



Generate a story about George's visit to the dentist.

1 George felt nervous as the kind dentist explained the check-up to make him comfortable.

2 At the dentist's office, George opened wide so the dentist could examine his teeth.

3 George then learns about dental hygiene from a friendly dentist showing him the tools and techniques.

Generate a story about George playing with a pigeon and a big yellow hat.

1 George opened the door and let in a friendly pigeon who shared some feathers with him.

2 George put on a big yellow hat that slipped over his eyes, making him giggle.

3 George, wearing the yellow hat, curiously approached a friendly pigeon on the balcony.

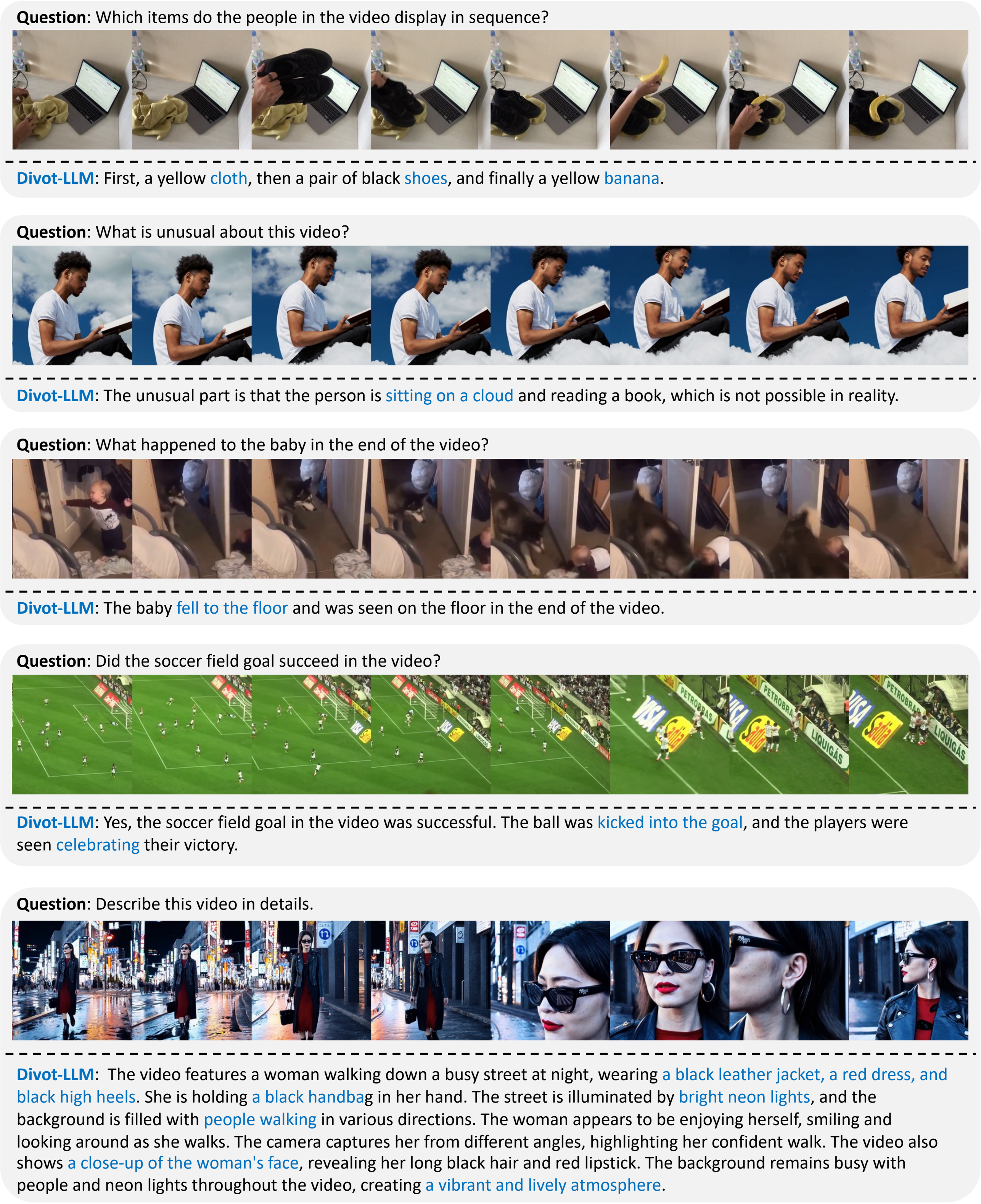
Generate a story about George's beach adventure with his new robot friend.

1 George and his cat friend found a shiny compass on the sandy beach, leading to a day of adventure.

2 A young man in a yellow tank top noticed something strange near a charming blue house on the sandy hill.

3 George and his robot friend explored the beach, making new friends and discovering interesting things.

Figure 10. More qualitative examples of video storytelling by Divot-LLM.



Question: Which items do the people in the video display in sequence?

Divot-LLM: First, a yellow cloth, then a pair of black shoes, and finally a yellow banana.

Question: What is unusual about this video?

Divot-LLM: The unusual part is that the person is sitting on a cloud and reading a book, which is not possible in reality.

Question: What happened to the baby in the end of the video?

Divot-LLM: The baby fell to the floor and was seen on the floor in the end of the video.

Question: Did the soccer field goal succeed in the video?

Divot-LLM: Yes, the soccer field goal in the video was successful. The ball was kicked into the goal, and the players were seen celebrating their victory.

Question: Describe this video in details.

Divot-LLM: The video features a woman walking down a busy street at night, wearing a black leather jacket, a red dress, and black high heels. She is holding a black handbag in her hand. The street is illuminated by bright neon lights, and the background is filled with people walking in various directions. The woman appears to be enjoying herself, smiling and looking around as she walks. The camera captures her from different angles, highlighting her confident walk. The video also shows a close-up of the woman's face, revealing her long black hair and red lipstick. The background remains busy with people and neon lights throughout the video, creating a vibrant and lively atmosphere.

Figure 11. Qualitative examples of video comprehension by Divot-LLM.