Supplementary Materials: Diffusion-based Multimodal Video Captioning

Jaakko Kainulainen, Zixin Guo, and Jorma Laaksonen

Aalto University, Finland
kainulainen.jaakko@gmail.com
{zixin.guo,jorma.laaksonen}@aalto.fi

1 Generated Descriptions

The generated descriptions consist of a few sentences that provide a general overview of the video content. An example of is shown in Figure 1, where a comparison is made between five ground truth captions and a generated description.

Ground truth captions:
1. a basket ball game between the lakers and the heat
2. a basketball game with a replay of a slam dunk
3. a basketball match and a player is putting a ball into the basket
4. a basketball player is putting the ball into the basket from distance
5. a couple of slam dunk highlights from nba games

Generated description:
This video captures a thrilling moments of a basketball game in progress. The camera focuses on the players as they dribble the ball, make successful shots, and engage in intense competition on the court. The crowd is filled with excitement, cheering for their favorite team. The atmosphere is electric, and the action is unforgettable.

Fig. 1: An example of a description generated by NExT-GPT [11] from a video in the MSR-VTT dataset.

2 Datasets

YouCook2 [14] dataset contains 2,000 instructional cooking videos, each showing a step-by-step instructions for preparing 89 distinct recipes from various culinary styles and cultural backgrounds around the world. Each video is divided into segments that demonstrate a specific step in a recipe. The segments are annotated with short captions that describe the cooking step shown in the segment. While the average segment duration is 19.6 seconds, the durations vary significantly from short 1 second segments to more elaborate segments lasting up to 264 seconds. The dataset is divided into 1,333 training and 457 test videos with 9,776 training and 3,369 test video-text pairs.

MSR-VTT [12] dataset contains 10,000 short video clips from 20 different categories, including music, sports, politics, movies and more. Each video is annotated by 20 human generated short annotations, providing diverse descriptions of the actions and context of the videos. The duration of the video clips range from 10 to 30 seconds. The dataset is divided into 6,513 training, 497 validation and 2,990 testing videos with 130,260 training, 9,940 validation and 59,800 testing captions.

VATEX [10] dataset contains 41,250 video clips covering a wide range of human activities. Each clip lasts around 10 seconds and is annotated with 10 captions. The training set contains 25,591 videos and the public test set contains 6,000 videos. However, due to some videos no longer being available, we use 24,977 videos for training and 5,770 for evaluation.

VALOR-32K [2] dataset contains 32,000 short video clips across various categories. Each clip is annotated with a caption describing both the visual and audio content of the video. The dataset is split into 25,000 training videos, 3,500 validation videos, and 3,500 testing videos.

Following UniVL [5], video features pre-extracted by S3D [13] are used for both the YouCook2 and MSR-VTT datasets. The video features for VATEX and VALOR-32K were extracted using CLIP-ViT-L/14 model [8]. The automatic speech recognition (ASR) transcript for the YouCook2 dataset was provided by UniVL [5], and the ASR transcripts for the MSR-VTT, VATEX and VALOR-32K datasets were generated specifically for these experiments using the Azure AI speech-to-text service [6]. Audio features for all datasets were extracted using the pretrained ImageBind model [3], which uniformly samples three segments of 2-second audio clips, converts the clips into 2D spectrograms, and encodes the spectrograms into audio features with 1024 dimensions. The generated descriptions were created using the pretrained NExT-GPT model [11] with the prompt "Describe this video". The model uniformly samples five segments of 2-second clips from the videos and generates descriptions based on them. Video features longer than 96 frames for YouCook2 and 50 frames for MSR-VTT are truncated, while video features for VATEX and VALOR-32K are sufficiently short, so truncation is not necessary. ASR transcripts and generated descriptions longer than 128 words are also truncated.

3 Evaluation Measures

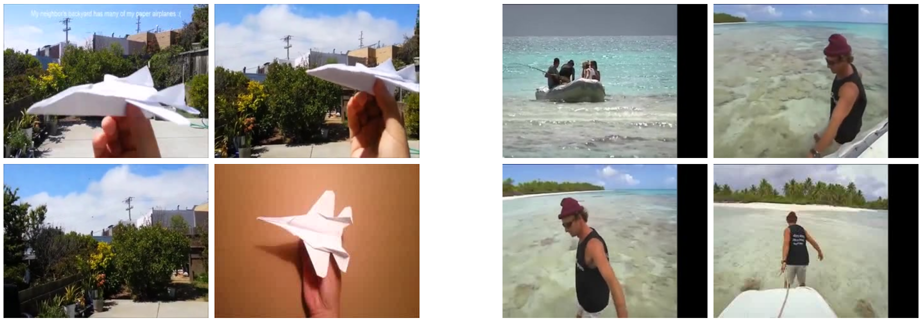
BLEU (Bilingual Evaluation Understudy, "B" in the result tables) [7] is a metric widely used to evaluate machine generated text. It measures the similarity between the generated sentence and one or more reference sentences by counting the number of matching n-grams in both.

METEOR (Metric for Evaluation of Translation with Explicit Ordering, "M") [1] evaluates the alignment between the words in the generated and reference texts. Unlike BLEU, it uses both precision and recall, and also matches synonyms.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation, "R") [4] is a set of metrics that count the quality of the generated text by considering overlapping units. ROUGE-L is a variant of ROUGE that calculates the quality based on the longest common subsequence. It selects only one subsequence, ignoring alternative longest subsequences.

CIDEr (Consensus-based Image Description Evaluation, "C") [9] evaluates the quality of the generated sentence by considering the set of n-grams present in the sentence, containing one to four words. Lower weight is given to n-grams that are more common in the reference sentences. CIDEr-D is a variant of CIDEr that uses a Gaussian to prevents candidate sentences with repeated words to receive high score.

4 Qualitative Results on MSR-VTT



Ground truth captions:
1. paper airplane in flight example
2. a person shows you how to make a paper airplane
3. someone throwing a paper airplane outside
4. someone is doing rocket on a paper
5. a person shows you how to make a paper airplane

Visual: a person shows a paper airplane
Audio: a man is explaining a paper
Speech transcript: a man is talking
Generated description: a plane is flying
All modalities: a man is folding a paper airplane

Ground truth captions:
1. a group of people are fishing
2. a group of people are riding on a raft in a body of water
3. a group of people are swimming in a pontoon
4. a man is dragging a boat in the ocean
5. a man pulls a boat in the water

Visual: a man is surfing the water in the water
Audio: a man is talking
Speech transcript: a man is boiling water in a water
Generated description: a man is playing something on the beach
All modalities: people are surfing in the ocean

Fig. 2: Qualitative results generated by using the different modalities for two videos from the MSR-VTT dataset.

Figure 2 show sample captions generated for MSR-VTT employing various modalities. The first example highlights how the visual modality effectively captures essential information in the video. Both audio and generated descriptions partially capture relevant details, but they fail to capture all aspects of the video. The caption generated from the speech transcript lack specificity. Many videos in the dataset lack speech and the model might struggle to extract meaningful information from the transcripts, resulting to caption that lack detail. The second example shows the model struggling with repeated words and its occasional inability to recognize specific actions depicted in the videos.

References

1. Banerjee, S., Lavie, A.: METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In: Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. pp. 65–72 (2005)
2. Chen, S., He, X., Guo, L., Zhu, X., Wang, W., Tang, J., Liu, J.: VALOR: Vision-audio-language omni-perception pretraining model and dataset. arXiv preprint arXiv:2304.08345 (2023)
3. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15180–15190 (2023)
4. Lin, C.Y., Och, F.J.: Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics. pp. 605–612 (2004)
5. Luo, H., Ji, L., Shi, B., Huang, H., Duan, N., Li, T., Li, J., Bharti, T., Zhou, M.: UniVL: A unified video and language pre-training model for multimodal understanding and generation. arXiv preprint arXiv:2002.06353 (2020)
6. Microsoft Corporation: Azure AI Speech to Text, `https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text/`
7. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: BLEU: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318 (2002)
8. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
9. Vedantam, R., Zitnick, C., Parikh, D.: CIDEr: Consensus-based image description evaluation. In: Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4566–4575 (2015)
10. Wang, X., Wu, J., Chen, J., Li, L., Wang, Y.F., Wang, W.Y.: Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4581–4591 (2019)
11. Wu, S., Fei, H., Qu, L., Ji, W., Chua, T.S.: NExT-GPT: Any-to-any multimodal llm. arXiv preprint arXiv:2309.05519 (2023)
12. Xu, J., Mei, T., Yao, T., Rui, Y.: MSR-VTT: A large video description dataset for bridging video and language. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5288–5296 (2016)

13. Zhang, D., Dai, X., Wang, X., Wang, Y.F.: S3D: single shot multi-span detector via fully 3D convolutional networks. arXiv preprint arXiv:1807.08069 (2018)
14. Zhou, L., Xu, C., Corso, J.J.: Towards automatic learning of procedures from web instructional videos. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 7590–7598 (2018)