

Fine-grained Audible Video Description

*Xuyang Shen², *Dong Li¹, *Jinxing Zhou³, Zhen Qin², Bowen He², Xiaodong Han², Aixuan Li⁴,
Yuchao Dai⁴, Lingpeng Kong⁵, Meng Wang³, Yu Qiao¹, ✉ Yiran Zhong¹

¹Shanghai Artificial Intelligence Laboratory, ²OpenNLPLab, ³Hefei University of Technology,

⁴Northwestern Polytechnical University, ⁵The University of Hong Kong

Abstract

We explore a new task for audio-visual-language modeling called *fine-grained audible video description (FAVD)*. It aims to provide detailed textual descriptions for the given audible videos, including the appearance and spatial locations of each object, the actions of moving objects, and the sounds in videos. Existing visual-language modeling tasks often concentrate on visual cues in videos while undervaluing the language and audio modalities. On the other hand, FAVD requires not only audio-visual-language modeling skills but also paragraph-level language generation abilities. We construct the first fine-grained audible video description benchmark (FAVDBench) to facilitate this research. For each video clip, we first provide a one-sentence summary of the video, i.e., the caption, followed by 4-6 sentences describing the visual details and 1-2 audio-related descriptions at the end. The descriptions are provided in both English and Chinese. We create two new metrics for this task: an EntityScore to gauge the completeness of entities in the visual descriptions, and an AudioScore to assess the audio descriptions. As a preliminary approach to this task, we propose an audio-visual-language transformer that extends existing video captioning model with an additional audio branch. We combine the masked language modeling and auto-regressive language modeling losses to optimize our model so that it can produce paragraph-level descriptions. We illustrate the efficiency of our model in audio-visual-language modeling by evaluating it against the proposed benchmark using both conventional captioning metrics and our proposed metrics. We further put our benchmark to the test in video generation models, demonstrating that employing fine-grained video descriptions can create more intricate videos than using captions. Code and dataset are available at <https://github.com/OpenNLPLab/FAVDBench>. Our online benchmark is available at www.avlbench.opennlpplab.cn.

1. Introduction

Language serves as the primary form of human communication, providing not only complementary information to other modalities such as vision and audio, but also an efficient means of exchanging information [17, 27, 46]. For example, we can use voice navigation to guide us to our destination. Visually impaired people can watch a movie by listening to its narration. The former shows that the language can provide complementary information to other modalities, while the latter indicates that the language can carry the most information in other modalities.

Recent multi-modal modeling tasks attempt to connect the language to other modalities, including image/video captioning [16, 32, 45, 48, 51, 64, 65, 75, 76, 85, 92], text-to-image/video synthesis [15, 19–21, 58, 59, 62, 63, 74, 80], text-driven image/video manipulation [3, 30, 44, 88, 89], and *etc.* However, in these tasks, since the language is frequently used to provide complementary information to other modalities, they often fail to give a fine-grained description of the exchange of information between modalities, and only consider the simplified language, i.e., one-sentence captions. Due to the conciseness of captions, only salient objects and activities have got descriptions. As a result, the information carried by the captions will be much less than that carried by other modalities, causing significant information loss when exchanging information from other modalities to language.

In this paper, we consider language as a tool for conveying information from other modalities in multi-modal modeling and propose a new task called Fine-grained Audible Video Description (FAVD). We compare it with existing captioning tasks in Fig. 1. Video captioning aims to generate one concise sentence to summarize the whole video. Dense video captioning first detects the multiple temporal events and then provides one caption for each event. The generated sentences of these two tasks merely cover the main objects or movements while losing many details when translating the video information into language. Unlike this, the FAVD task requires a model to describe videos in a similar way that humans do, i.e., starting with an overview of the video and then focusing on each fine-grained detail to con-

*These authors have equal contributions. ✉Yiran Zhong is the corresponding author (e-mail: zhongyiran@gmail.com).

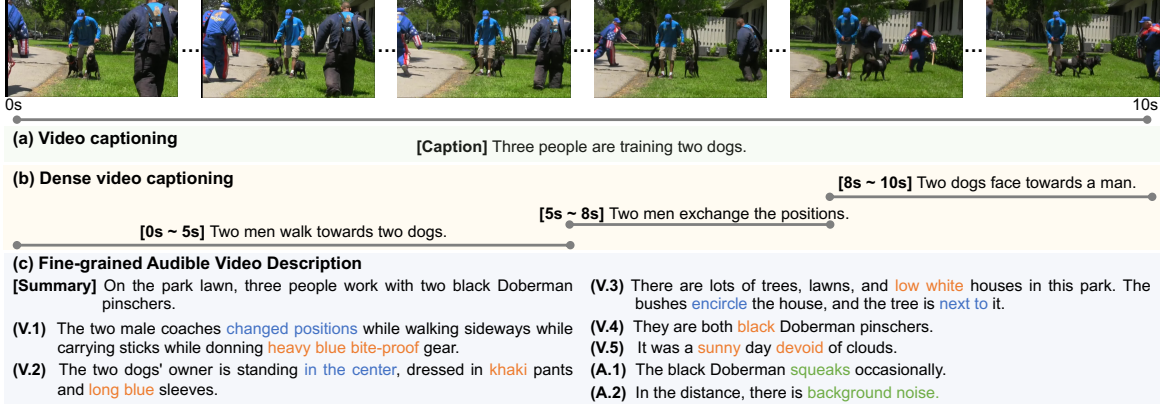


Figure 1. **Comparison of the proposed FAVD task with existing captioning tasks.** (a) Video captioning (VC) uses one sentence to describe the main content of the video. (b) Dense video captioning aims to localize the multiple temporal events and generate corresponding descriptions. Both VC and DVC describe the salient events in videos while losing many details, such as the appearance of objects, spatial relations, and sounds. (c) The proposed FAVD tries to generate a paragraph-level description that contains the caption, named as **Summary**, and the audio-visual descriptions, abbreviated as **A.** and **V.**

create the description. In this case, most video information can be preserved in the language modality. Since a video contains both visual and audio signals, we also include audio descriptions in our task.

FAVD is a non-trivial task. In addition to long-text modeling abilities, it calls for a finer-grained visual understanding than earlier video captioning tasks, *i.e.*, it needs to recognize all objects and actions in videos, as well as describe the appearance and spatial locations of each object and the sounds in videos. To facilitate this task, we construct the first fine-grained audible video description benchmark (FAVDBench), which allows the model to be trained in a supervised manner.

FAVDBench is made up of over 11,000 video clips culled from millions of YouTube videos and each clip is sourced by querying more than 70 life-related categories to fulfill the diversity of the benchmark. We annotate the video descriptions based on human behaviors. The annotations of each video clip begin with a one-sentence caption that describes the salient objects and activities, followed by 4-6 sentences that describe visual details including the appearance and spatial locations of each object, and the actions or movements of moving objects. To make the task consider audio information, we include 1-2 sentences of audio-related descriptions at the end of the whole descriptions. The descriptions are provided in both English and Chinese with human translation. An annotation example can be found in Fig. 1 and Fig. 3. We design two new metrics for the FAVD task. One called *EntityScore*, which assesses the completeness of the information that is transferred from the videos to the descriptions by gauging the completeness of entities in the visual descriptions. The other is *AudioScore*, which measures the audio description in the feature space of a pre-trained audio-visual-language model.

We provide a baseline model for this new task. The

model is based on an existing end-to-end video captioning model [35] with an additional audio branch. We also extend the visual-language transformer to the audio-visual-language transformer in order to mix multi-modal tokens. Existing video captioning models often adopt Masked Language Modelling (MLM) loss to optimize the caption generation. However, because our task requires the model to describe fine-grained details, the model should also be capable of general language modeling. To this end, we include the auto-regressive language modeling (ALM) loss in our baseline model.

We extensively evaluate the baseline model against the proposed FAVDBench using both the conventional captioning metrics and our proposed metrics and demonstrate the effectiveness of our model in audio-visual-language modeling and fine-grained description generation. To qualitatively illustrate the information preservation in modality information exchange, we further put our benchmark to the test in video generation models, showing more intricate videos than using captions.

2. Related Work

Video captioning. Video captioning aims to describe the main content in a video with one concise natural language sentence. The pioneer works mainly follow the sequence to sequence pipeline, where the Convolutional Neural Network is used to encode the video frame features and the Recurrent Neural Network is used to decode the predicted sentence [51, 73, 76, 82]. Some following works adopt the Transformer architecture for efficient captioning [7, 13, 35, 38, 39, 53, 84, 86]. Recently, Tang *et al.* [69] use the pretrained CLIP [55] to extract advanced text and visual features that improve the captioning performance significantly. Besides, there are several works that try to explore the spatio-temporal clues [1] or object rela-

Table 1. **Statistics of FAVDBench and other existing video captioning datasets.** FAVDBench is the first dataset to provide audible descriptions of videos. “#Sentence” and “#Word” denote the average number of sentences and words contained in the annotation per video, respectively. We also display the part-of-speech tagging (POS tag) in the percentage of each dataset. FAVDBench has the highest percentage of adjective words (Adj.) which indicates the rich annotation of fine-grained audio-visual details.

Dataset	Video		Annotation				POS tag		
	#Clip	DUR. (h)	Audio	#Sentence	#Word	Vocabulary	%Adj.	%Noun	%Prep.
MSVD [5]	1,970	5.3	✗	35.5	308.3	13,010	2.6	31.8	7.7
MSR-VTT [81]	10,000	41.2	✗	20.0	185.7	29,316	4.8	33.9	11.5
VATEX [79]	41,250	114.6	✗	20.0	291.8	82,654	4.4	31.8	12.4
TVC [29]	21,793	461.3	✗	5.0	67.0	57,100	2.2	36.4	12.7
YouCookII [90]	15,433	176.0	✗	1.0	7.9	2,583	4.1	40.3	11.6
FAVDBench	11,424	24.4	✓	12.6	218.9	73,245	13.0	30.5	12.4

tions [47, 85] to give more accurate descriptions. However, it is hard for these captioning models to detail the rich semantics in a video with only one sentence. Our proposed FAVD task carries the FAVDBench dataset which provides fine-grained annotations allowing us to train an intelligent robot that can tell rich video details.

Dense video captioning. Since there are usually complex content and various events contained in a long and untrimmed video, it is insufficient to describe the video with only one sentence. For this reason, dense video captioning is proposed [26] which aims to automatically localize the multiple temporal events and generate corresponding captions. Most of the related works follow the “localize then describe” scheme [33, 43, 66, 68, 77, 91] which first predicts many event proposals and then generates its captions. Unlike this, Wang *et al.* [78] propose an end-to-end captioning model that decodes the event proposals and captions parallelly. Some methods utilize the multi-modal information such as audio [22, 23, 57, 87] and motion [6] in videos for better captioning. Although these methods achieve satisfactory performance on dense video captioning, they all manipulate the frame-level features that prohibit the network from learning fine-grained details, which is indeed the essential expectation of the proposed FAVD task. Unlike them, the proposed baseline framework encodes the patch-level video representations with the aggregation of multi-modal tokens that enables to give more detailed descriptions.

Audio captioning. Unlike video captioning and dense video captioning which mainly focus on the visual domain, audio captioning aims to directly generate text description for audio in the wild [12, 24, 36, 40, 83]. In the initial work, Kim *et al.* [24] propose an encoder-decoder method, where the Bi-LSTM is used to generate the captions word by word. Recently, Mei *et al.* [41] use the Transformer backbone to encode the long-range dependencies in the audio signal and output the whole predicted sentence. Though these methods enable us to describe the audio to some extent, we argue that it is hard for accurate audio captioning without giving any visual information. FAVD takes both audio and visual sig-

nals as inputs, which eases the description generation.

Audio-visual-language dataset. Existing captioning-related multi-modal datasets mainly focus on two modalities: the *vision-language* datasets for (dense) video captioning, such as MSVD [5], MSR-VTT [81], VATEX [79], TVC [29], and YouCookII [90]; or the *audio-language* datasets for audio captioning, such as AudioCaps [24] and Clotho [11]. A few video captioning works try to introduce the audio track contained in videos of vision-language datasets, but the audio actually limits to speech domain [29, 42, 60, 90]. While videos in the proposed FAVDBench are collected from various life-related scenes that extend the diversity of the audio. Besides, the audio and visual signals are usually semantic corresponding which makes it a truly audio-visual-language dataset.

3. The FAVDBench

3.1. Task description

The FAVD task involves generating fine-grained textual descriptions for the given audible videos that include information about appearance and spatial location of each object, the movements of moving objects, and the sounds in the videos. Specifically, given an audible video clip, this task aims to generate a paragraph with 6-9 sentences. The first sentence needs to be an overview of the video clip that emphasizes the key visual elements as well as the main event. The next 4-6 sentences should then be used to describe the minute visual details of each object, *i.e.*, the rich semantics that each one possesses and the function that it serves in the video. The final 1-2 sentences characterize the audio information contained in the video clip, describing what is producing the sound or what it sounds like. We compare the annotation differences between the FAVD and previous video captioning tasks in Fig. 1.

3.2. Dataset statistics

We construct a benchmark called FAVDBench to facilitate the FAVD task. We adopt the technique described in VGGSound [2] to make sure the audio and visual clips match the intended semantics. All videos are collected from

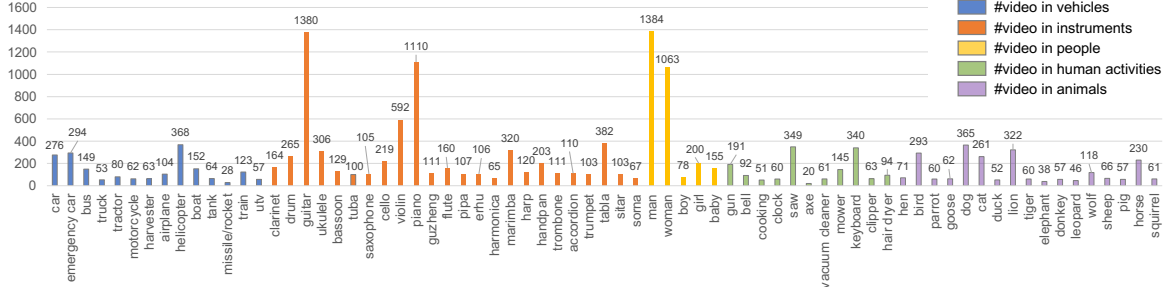


Figure 2. **Video distribution of 5 major categories and 71 sub-classes in FAVDBench.** The value represents the total occurrence in visual of each class. The vehicles class is colored blue, contains 14 sub-classes. There are 24 sub-classes in instruments, colored orange. The people category includes man, woman, boy, girl, and baby, which are colored yellow. The category of human-related activities is colored green and contains 11 sub-classes. There are 17 sub-classes in animals, colored purple. Best view in color.

YouTube under *Creative Commons* license. We take into account the number of sounding objects when collecting video clips as we find the difficulty of audio description is positively correlated with the number of sounding objects. As a result, 60.2% of the video clips have one sounding object and the rest have 2-3 sounding objects.

The FAVDBench consists of 11,424 audible video clips and 143,548 sentences of annotations. Each video clip has a description of 12.6 sentences, 218.9 word tokens on average. Among the clips, 5,289 videos are trimmed to 5 seconds and 6,135 videos to 10 seconds. The total video duration is 24.4 hours and 2.6 million frames. To avoid data leakage and maintain diversity, we enforce a minimum interval of 5 seconds between adjacent videos. We also manually filter clips with similar content. The benchmark is divided into 4 partitions: 7,500 for training, 1,500 for validation, 1,000 for testing, and 1,424 withhold for online benchmarking. The withhold partition will only be visible for contestants in the future FAVD Benchmark competition.

We compare FAVDBench with other video captioning datasets in Table 1. In terms of context differences, we include 5 major categories for video content, *i.e.*, vehicles, instruments, animals, people, and common human activities. We further divide the 5 major categories into 71 sub-classes and the distribution of each category/subclass is plotted in Fig. 2. By contrast, the contexts of existing captioning datasets TVC [29] and YouCookII [90] are restricted to movie and cooking scenarios while the MSVD [5], MSR-VTT [81], and VATEX [79] merely include human activities. Besides, we also provide audio descriptions for each video clip in FAVDBench, which distinguishes us from other datasets. Furthermore, since our benchmark provides fine-grained descriptions for video objects, the percentage of adjectives in our dataset is more than 3 times higher than in other datasets.

3.3. Annotation details

FAVDBench provides both English and Chinese annotations. All video clips are primarily annotated in Chinese using crowdsourcing and then translated to English with



Figure 3. **An example of FAVDBench annotation.** For each video clip, we provide both Chinese and English annotations, each contains 1 summary [S.], 4-6 detailed visual descriptions [V.] and 1-2 audio-related descriptions [A.]. Details of spatial locations, appearance, and sounds are highlighted in blue, orange, and green.

human translation. To collect large-scale fine-grained descriptions of videos, we design several rules of annotation. Specifically, each video is required to be watched completely, and then described by 1 summary sentence, at least 4 visual-related points, and at least 1 audio-related point. The summary only covers high-level information about the most salient event that the video tries to convey. The order to describe the visual and audio details follows from the most salient objects to the background. The statuses of objects are required to be thoroughly captured, including the actions, characteristics (adjective), and relative spatial relations (preposition). Furthermore, each sentence is required to be larger than 5 Chinese characters. Any subjective thoughts and speculative words (e.g., “maybe”) are forbidden to appear in labels, especially in the audio description. In addition, information beyond the video and audio is not allowed to include in the annotations, such as the wiki of objects, subtitles, and speech contents. A video sample with bilingual annotations is shown in Fig. 3.

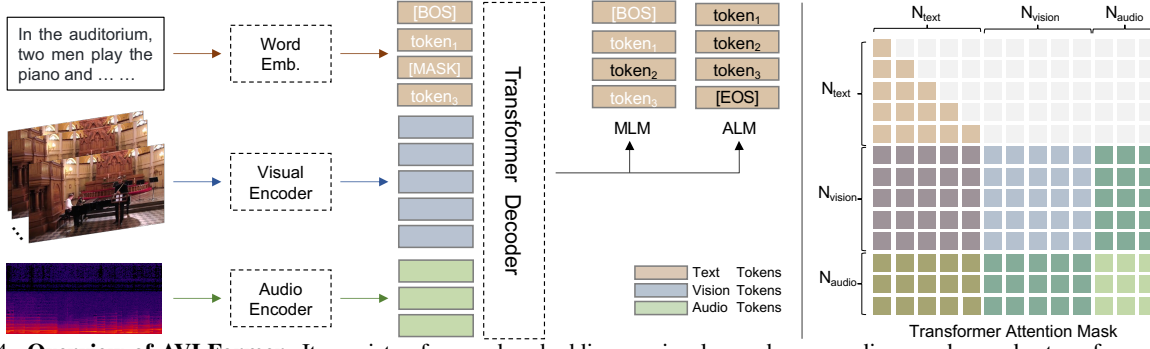


Figure 4. **Overview of AVLFormer.** It consists of a word embedding, a visual encoder, an audio encoder, and a transformer decoder. We adopt the video swin transformer and patchout audio transformer as the visual encoder and audio encoder, respectively. They extract visual and audio features from video frames and audio. Masked language modeling and auto-regressive language modeling are configured in training. The attention mask strategy of AVLFormer is illustrated on the right, where the masked attention is colored in gray. The tokens and attention masks of text, vision, and audio are colored brown, blue, and green, respectively.

To reduce the annotation bias and ensure the quality of annotation, the annotators are restricted to around 60 people, and all of them are native speakers of Chinese. Each video is annotated by one worker, checked by 2 peer workers, and examined by grammar-checking tools. The English and the Chinese version of descriptions are congruent to prevent any changes to the semantics of sentences.

3.4. Evaluation metrics

The generated descriptions should be evaluated in terms of semantics rather than word-for-word precision. For example, we would like to know whether the entities in the videos are addressed and the sounding objects are properly described, rather than whether they are using the same way to describe these things. Existing video caption metrics, such as BLEU [49], ROUGE [34], Meteor [9], and CIDEr [72] often concentrate on word-for-word precision by measuring the token similarity between the generated and ground truth texts, which does not meet our needs. Therefore, to properly evaluate the generated descriptions, we design two new metrics called EntityScore \mathbb{ES} and AudioScore \mathbb{AS} .

EntityScore measures the extent to which consistently referred words or series of words, known as entities and often manifested as nouns, in the predicted text match those in the annotated text. We use an off-the-shelf Natural Language Toolkit library to extract nouns as entities. The mathematical expression of the EntityScore \mathbb{ES} is as follows:

$$R(\mathbf{p}, \mathbf{r}) = \frac{\#\{\mathbf{p} \cap \mathbf{r}\}}{\#\{\mathbf{r}\}}, C(\mathbf{p}, \mathbf{r}) = \frac{\cos(\mathbf{T5}(\mathbf{p}), \mathbf{T5}(\mathbf{r})) + 1}{2}, \quad (1)$$

$$\mathbb{ES}(\mathbf{p}, \mathbf{r}) = \frac{2R(\mathbf{p}, \mathbf{r})C(\mathbf{p}, \mathbf{r})}{R(\mathbf{p}, \mathbf{r}) + C(\mathbf{p}, \mathbf{r})},$$

where $\mathbf{p} \in \mathbb{R}^n$ is the predicted entities from the model predictions, and $\mathbf{r} \in \mathbb{R}^m$ is the ground truth entities, n and m are the entity numbers.

$\#\{\mathbf{r}\}$ means the number of entities in the reference, and $\#\{\mathbf{p} \cap \mathbf{r}\}$ counts the number of correctly described entities in the prediction. \cos denotes the cosine similarity between

two vectors and $\mathbf{T5}$ represents the pretrained model $\mathbf{T5}$ [56] used to extract entity features. Besides, it is normalized into the range $[0, 1]$ to guarantee its positive. Therefore, $R(\mathbf{p}, \mathbf{r})$ indicates the recall of the predicted entity that is described using the same word as the reference and $C(\mathbf{p}, \mathbf{r})$ measures the comprehensiveness of semantics between the predicted entities and the ground truth entities. In other words, R focuses on syntactic level accuracy on entities while C concentrates on semantic level entity similarities. The \mathbb{ES} considers both syntactic and semantic level accuracy. However, \mathbb{ES} has a limitation in that it places emphasis solely on the presence of entities while disregarding their placement and frequency.

AudioScore assesses the accuracy of audio descriptions by computing the product of the extracted audio-visual-text unit features, where CLIP [55] is used to extract features for video frames and the corresponding descriptions and PaSST [25] is used for audio waves. Moreover, we fine-tune these models using contrastive learning on FAVDBench to ensure that the latent features are mapped to a shared space. Specifically, the AudioScore \mathbb{AS} is defined as:

$$\mathbf{e}_a = \text{PaSST}(\mathbf{A}), \mathbf{e}_v = \text{CLIP}(\mathbf{V}), \mathbf{e}_t = \text{CLIP}(\mathbf{T}),$$

$$s = \left(\frac{1}{2} \cos(\mathbf{e}_a, \mathbf{e}_t) + \frac{1}{2} \cos(\mathbf{e}_a, \mathbf{e}_v) + 1 \right) \times 0.5, \quad (2)$$

$$\mathbb{AS}(\mathbf{A}, \mathbf{V}, \mathbf{T}) = \mathbf{f}(s), \mathbf{f}(x) = a \exp(-b \exp(-cx)),$$

where \mathbf{A} and \mathbf{V} denote the audio and visual frames of one video, respectively, while \mathbf{T} is the predicted audio description. \cos and \mathbf{f} denote cosine similarity and a specific form of Gompertz function, respectively. We set $c=10$ empirically, and choose values for a and b ($a = \frac{1}{e^{-0.69e-10}}$, $b=0.693$) to force specific values of x and $\mathbf{f}(x)$ ($x=1$, $\mathbf{f}(x)=1$ for $a = \frac{1}{e^{-0.69e-10}}$, and $x=0$, $\mathbf{f}(x)=0.5$ for $b=0.693$). We select the last one or two sentences from the whole generated paragraph description as \mathbf{T} , as we find that the model prefers describing audio information at the end of the description, mimicking the ground truth. Therefore, we report the Top-1 and Top-2 accuracy as the \mathbb{AS} score for the last

Table 2. **Comparison with different methods on FAVDBench.** We report different backbone settings for PDVC, SwinBERT, BMT and AVLFormer. In the backbone freeze (FRZ.) column, - represents that input is not available; \rightarrow represents that is trainable; * represents frozen. The visual and audio backbone of BMT is offline by default. For all metrics, higher values are better. The maximum human evaluation score of ten (10) is regarded as representing the ground truth.

Method	Backbone FRZ.		Conventional Metric					Proposed Metric		Human Evaluation
	Visual	Audio	B@1	B@4	Meteor	CIDEr	Clipscore	EntityScore	AudioScore	
PDVC [78]	\rightarrow	-	34.91	6.33	13.80	6.27	65.77	33.45	56.88	3.7
PDVC w. Audio	\rightarrow	\rightarrow	35.59	6.47	14.49	13.13	66.15	34.12	60.72	4.1
SwinBERT [35]	\rightarrow	-	39.68	9.05	16.78	21.69	68.13	39.60	62.60	6.0
SwinBERT [35]	*	-	41.41	9.08	17.17	23.03	67.98	41.00	58.72	5.9
BMT [22]	*	*	41.28	9.23	16.57	16.46	65.00	35.99	61.03	6.4
AVLFormer	\rightarrow	\rightarrow	44.10	10.29	18.36	29.85	69.74	44.46	63.88	8.2
AVLFormer	*	\rightarrow	44.10	10.23	18.24	26.31	68.98	43.47	61.53	7.9
AVLFormer	\rightarrow	*	42.82	10.16	17.96	25.45	70.07	43.05	62.84	7.8
AVLFormer	*	*	42.35	9.84	17.73	25.65	69.33	43.13	61.90	7.6

one or two sentences. We use the Top-1 score by default.

4. The AVLFormer

We propose a transformer-based model called Audio-Visual-Language Transformer (AVLFormer) as a baseline method for the FAVD task. AVLFormer is in a form of encoder-decoder structures as shown in Fig. 4. A visual encoder and an audio encoder are adapted to process the video clips and audio, respectively. Similar to previous video captioning models [35], AVLFormer also takes textual descriptions as input and uses a word tokenizer and a linear embedding to embed the text. The output of AVLFormer is the fine-grained descriptions of the input videos.

The encoders. We employ the video swin transformer [37] as our visual encoder as it achieves state-of-the-art performance in video understanding tasks [35, 37]. The weights of the visual encoder are pretrained on the ImageNet-22k [8] and the Kinetics-600 [4] datasets. The visual features $\mathbf{F}_v \in \mathbb{R}^{N_v \times d_v}$, where $d_v = 512$ is the feature dimension and $N_v = 784$ is the patch sequence length. We adopt the patchout audio transformer (PaSST) [25] as our audio encoder. We pretrain it on the ImageNet-1K [8] and the AudioSet [14] datasets. The audios are firstly processed by a Fast Fourier Transform and a Mel feature extractor and then sent to PaSST. The audio features $\mathbf{F}_a \in \mathbb{R}^{N_a \times d_a}$, where $d_a = 768$ is the dimension and $N_a = 473$ is the sequence length. For the text input, we apply a word tokenizer and a linear embedding to embed the text as $\mathbf{F}_t \in \mathbb{R}^{N_t \times d_t}$, where $d_t = 768$ and $N_t = 300$.

Audio-visual-language transformer. Before passing the audio features \mathbf{F}_a , visual features \mathbf{F}_v , and text embedding \mathbf{F}_t to the audio-visual-language transformer (AVLFormer), a linear projection layer is employed to unify their feature dimensions to $d = 768$. AVLFormer takes the concatenated feature $\mathbf{F}_{avt} \in \mathbb{R}^{(N_t+N_v+N_a) \times d}$ of \mathbf{F}_a , \mathbf{F}_v , and \mathbf{F}_t as input and processes it with a stack of multi-head self-attention

modules. The number of layers of the transformer is set to 12. We illustrate the masking strategy of AVLFormer in Fig. 4. Specifically, we use the auto-regressive language model masking strategy to mask the text attention matrix, which ensures that the predicted tokens are only dependent on the previous tokens. For the audio and visual tokens, since there is no ordering restriction for these modalities, we use full attention matrices instead. We also make audio and visual tokens visible to text tokens, ensuring that other modality information can contribute to text generation.

Loss functions. Previous video captioning methods [28, 35] mainly adopt the masked language modeling loss to make the model concentrate on salient objects and actions, whereas previous vision-language generation methods [31] often use auto-regressive language modeling loss for longer text generation. For the FAVD task, since it requires not only fine-grained descriptions for each object but also paragraph-level text generation capability, we employ both the masked language modeling loss \mathcal{L}_{MLM} and the auto-regressive language modeling loss \mathcal{L}_{ALM} to train our AVLFormer. In masked language modeling, AVLFormer is required to recover the words that are replaced by [MASK] symbol, as shown in Fig. 4. In this way, the model is forced to recuperate the information based on audio-visual information. In auto-regressive language modeling, AVLFormer is asked to understand the semantic context for the entire sentence. Note that AVLFormer works in an auto-regressive manner in inference, the \mathcal{L}_{ALM} can also mitigate the gap between the training and the inference. The total object function \mathcal{L} can be computed as follows:

$$\mathcal{L} = \lambda \mathcal{L}_{MLM} + (1 - \lambda) \mathcal{L}_{ALM}, \quad (3)$$

where λ is a scaling factor between 0 and 1.

Inference. AVLFormer takes video frames and audios as the input for visual and audio encoders and the [BOS] token for the word embedding in inference. It then generates

descriptions in an auto-regressive manner. Specifically, it predicts the next word iteratively until the [EOS] token or max sequence length is reached.

5. Experimental Results

5.1. Implementation details

We conduct experiments on the proposed FAVDBench. All video clips are sampled uniformly to 32 frames, and all frames are resized to the shape of 224×224 . All audios are sampled by 32kHz in a mono channel. We set the max sequence length and mask probability to 300 and 0.25, respectively for natural language processing. We use the Adam optimizer with a linear warm-up learning rate. The batch size is set to 8 and the maximum number of training epochs is set to 150.

Evaluation metrics. We include 5 conventional captioning metrics as well as the two newly proposed metrics as our evaluation metrics, namely, BLEU [50], Meteor [10], Rouge-L [61], CIDEr [71], Clipscore [18], EntityScore, and AudioScore. We also perform the human evaluation on the generated descriptions. Among the entire test-set, 10% of samples are randomly selected to be measured, scoring from 0 to 10, and average values of 20 volunteers are reported. The ground truth descriptions are set to 10. Evaluators are required to focus on the smoothness of descriptions and the correlation between the descriptions and the inputs.

5.2. Main results

We compare our AVLFormer with the state-of-the-art video captioning method SwinBERT [35], video captioning method with original audio input BMT [22], dense video captioning method PDVC [78] on our FAVDBench for the new FAVD task. Both quantitative and qualitative results are provided in the following sections.

Quantitative comparison. We report the quantitative results over the 8 evaluation metrics in Table 2. We compare AVLFormer with PDVC, SwinBERT and BMT under different backbone freezing settings. AVLFormer beats competitors with a clear margin and achieves the best performance when both the visual and audio backbones are unfrozen. Among the 7 automatic evaluation metrics, the EntityScore is trending closer to human evaluation, demonstrating the validity of the evaluation metric.

Qualitative comparison. We illustrate the qualitative results of AVLFormer and other competitors in Fig. 5. AVLFormer provides finer-grained details of entities and their status than other methods, especially for the spatial location descriptions. For example, AVLFormer attempts to capture the relative positions of major objects, such as the spatial relations among the harbor, the reefs and the yacht, and generates descriptions that are close to the reference. For the audio descriptions, AVLFormer describes the sound and the sounding objects whereas PDVC and SwinBERT often neglect the sounding sources.

Table 3. **Impact of video and audio.** Four different inference modes from top to bottom: positive pairs of video and audio, negative pairs of video and audio, random audio, random video.

Inference	CIDEr	EntityScore	AudioScore
Pos. V.&A.	29.85	44.46	63.88
Neg. V.&A.	18.63	8.25	31.25
Video w Random	20.12	40.77	61.22
Random w Audio	1.58	11.86	44.70

Table 4. **Audio impact on AudioScore.** Four different audio-visual-text pairs are typed from top to bottom: original, audio-shuffled, audio random initialized and audio fixed initialized pairs.

Pair Type	Top-1 AudioScore	Top-2 AudioScore
Type I	63.88	64.60
Type II	52.59	56.31
Type III	41.93	38.06
Type IV	39.31	40.12

5.3. Ablations

We first ablate the impact of audio and visual modalities on AVLFormer by replacing the positive audio-visual pairs with negative pairs or random noises in inference. The results are shown in Table 3. We then analyze the contribution of the masked language modeling loss and the auto-regressive language modeling loss in Fig. 7. Finally, we evaluate the validity of our proposed metrics.

Impact of audio. We empirically find that the trained model can infer partial sound-related descriptions without an audio branch as shown in Fig. 5 and Table 3. It is probably because audio is often related to actions, and the network can predict sound descriptions by understanding the actions. However, in this situation, the sound descriptions are often incorrect or unrelated to the actual audio. For example, the model generates one-third of samples with a pattern like: *somebody/something makes noise*.

Impact of visual. As shown in Table 3, replacing visual information with random noises will cause a significant performance drop in all metrics. This indicates that in FAVD, the visual modality is most important. Even with the aid of correct audio information, the model cannot produce related descriptions without visual guidance.

Analysis of losses.

We employ both the masked language modeling loss \mathcal{L}_{MLM} and the auto-regressive language modeling loss \mathcal{L}_{ALM} as our objective loss functions. A ratio λ is adopted to balance the weights between them (Eq. 3). We train our model with different λ s under the same setting and plot the test set accuracy in Fig. 7. As shown, using pure \mathcal{L}_{MLM} (i.e., $\lambda = 1$) achieves significant better performance than using pure \mathcal{L}_{ALM} (i.e., $\lambda = 0$), e.g., 27.11 vs 22.22 in CIDEr and 44.37 vs 42.24 in EntityScore. The model achieves the best performance when $\lambda = 0.9$.

Analysis of proposed metrics. The EntityScore indicates



Reference: The yacht was being driven by one man while the other man stood stationary. Driving the yacht on the left side is a man with a short hair, sunglasses, and a white shirt. On the right side of the yacht, a man with short black hair, sunglasses, and red jeans is immobile. The two were seated on a blue-and-white yacht with two white engines in the back and a white interior. The beach, lush trees, people, and buildings are to the left of the two, while the sea is in front of them. It's sunny and clear outside. There is a noisy yacht. The video includes the sound of pulsing water.

PDVC: A man drives a yacht on the sea. A man with a white hat and a white hat and a white hat is driving on the sea. There are many people on the shore. The sound of waves waves.

SwinBERT: On the water a speedboat is traveling quickly. On the water a speedboat with a white body and a white bottom is moving quickly. On the speedboat there are both tall buildings trees and tall buildings. The speedboat is incredibly loud.

BMT: On the water, a speedboat is moving. On the water, a white speedboat with a white bottom and a white bottom is moving quickly. On the speedboat, there are a lot of buildings, and trees. The speedboat is moving quickly. The speedboat is moving quickly. The sound of the speedboat's engine may be heard.

AVLFormer: On the water a man pilots a yacht. A man in a blue shirt and sunglasses is operating a white speedboat. On the yacht there is a man wearing glasses and a woman with glasses. On the yacht there are a lot of rocks. The harbor beneath the yacht is covered in several reefs. The boat is moving swiftly across the water creating a buzzing sound as it does so.

Figure 5. **Qualitative result of fine-grained video description on FAVDBench.** We compare the proposed AVLFormer with the PDVC [78], SwinBERT [35] and BMT [22]. The generated video summary of each method is underlined. The spatial relation, visual adjectives, and audio-related words are highlighted in blue, orange, and green, respectively.

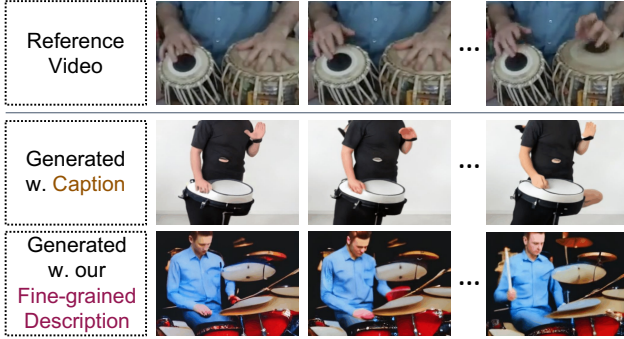


Figure 6. **Qualitative examples of video generation on FAVD-Bench.** We use Cogvideo [20] to generate videos through the caption and our fine-grained descriptions. We eliminate audio-related input descriptions as current video generation models cannot produce sound. *Caption:* “A man sings and beats the drums with his hands in the room.”; *Description:* “In the room, a man in a blue shirt sings and beats drums with both hands. The man is pounding two drums, one large and one small. The two white drums have black dots on them and are wrapped in light brown ropes.”

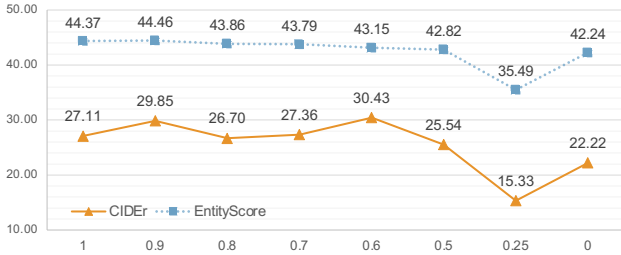


Figure 7. **Analysis of losses.** We report model performance from on and EntityScore with respect to different λ values of loss.

how comprehensively described entities are when compared to reference descriptions. As shown in Table 2 and Table 3, the EntityScore displays a scoring pattern that is fairly similar to human evaluation, which demonstrates the validity of the EntityScore. For the AudioScore, we additionally design an ablation to evaluate its validity. Specifically, we compare four settings of audio-visual-text pairs: Type I: the

original audio-visual-text pairs; Type II: shuffle the audios while keeping the visual-text pairs unchanged; Type III: replace the audio with uniformly random initialized noise; Type IV: replace the audio with a fixed 1e-3 value initialization. The results are shown in Table 4. Type I keeping the semantic-consistent audio and visual-text pairs achieves significantly higher AudioScore than others, proving the validity of the AudioScore.

5.4. Video generation

Existing video generation models can also benefit from our FAVDBench as we provide finer-grained video descriptions than captions. To prove our claim, we compare the generated videos from captions and our fine-grained description using a state-of-the-art text-video model Cogvideo [20]. As shown in Fig. 6, the video generated from the descriptions is closer to the reference video than the one generated from the captions. Furthermore, FAVD-Bench can be used to train video generation models and can substantially reduce the impact of insufficient scene description on video generation tasks as mentioned in [63].

6. Conclusion

We proposed FAVD, a new audio-visual-language modeling task that aims to generate paragraph-level fine-grained textual descriptions for every object in the given audible videos. A new benchmark called FAVDBench was constructed to facilitate the research. Two new metrics were designed to evaluate the quality of the generated descriptions, where the EntityScore is used to assess the completeness of entities in descriptions and the AudioScore evaluates the accuracy of audio descriptions. We also presented a transformer-based architecture AVLFormer for this task. Extensive experiments validated our model and metrics design. We intend to explore architectures in FAVD-Bench [52, 54, 67], and use FAVDBench to investigate audible video generation in the future.

Acknowledgement: We thank Lei Li and Hai Jie for their valuable advice on this work. This work is partially supported by the National Key R&D Program of China (NO.2022ZD0160100) and partly by the Shanghai Committee of Science and Technology (Grant No. 21DZ1100100).

7. Appendix

In the appendix, we additionally provide the word frequency of FAVDBench and more ablation study results such as the impact of different attention masking strategies in AVLFormer and extended quantitative results on FAVDBench. We also include more video generation samples as well as their quantitative results. A video (<https://youtu.be/iWJvTB-bTWk>) containing some samples in FAVDBench is attached to this supplementary material.

7.1. The FAVDBench

Word Frequency. As FAVDBench provides both English and Chinese descriptions, we plot the frequency of English and Chinese vocabularies in Fig. 8. The most common words in descriptions, such as “man” and “woman,” are nouns. Whereas, both adjectives and prepositions also appear in the most frequent vocabularies of descriptions (*e.g.*, “red”, “black”, “left”, “right”). It is worth noting that sound-related vocabularies occur in the word frequency diagram, such as “sound”. It proves that our descriptions of videos thoroughly capture the visual and audio details, including actions, characteristics, and relative spatial relations.

7.2. Experimental Results

Impact of different AVLFormer attention masks. We ablate the impact of different AVLFormer attention masks in Fig. 9. The empirical results illustrate that the full attention mask (Type II) negatively affects model coverage. It may cause by information leakage when all word tokens are visible to vision and audio tokens when we infer the descriptions in auto-regressive manner. Audio tokens are not visible to vision tokens (Type III - Type V) contributes to performance improvement, compared to the default attention mask (Type I).

Extended quantitative results. We report 6 additional metrics together with the 8 metrics from the main paper in Table 6. We extend the Clipscore [18] from the first frame to 32 frames. To obtain the later value, all frames are required to measure the cosine similarity with descriptions, and report the the averaged scores. Over the 14 evaluation metrics, AVLFormer leads a substantial performance improvement than PDVC and SwinBERT under various backbone freezing settings.

Quantitative results of SwinBERT. We evaluate SwinBERT over 7 representative metrics through loading various pretrained weights and fine-tuning on FAVDBench in Table 7. Pretrained weights from other captioning dataset



Figure 8. **Word Frequency.** We count the word frequency of Chinese annotation (upper) and English annotation (bottom) on FAVDBench.

contribute to the performance improvement, especially the VATEX. As a comparison, AVLFormer beats all settings of SwinBERT over all evaluation metrics.

7.3. Video Generation

Quantitative results. To compare the performance between the caption-generated and the description-generated videos, We report the quantitative results over Frechet Video Distance (FVD) [70] metric in Table 5. We employ Cogvideo [20] to generate the video based on captions and our fine-grained descriptions. The FVD score reflects that our fine-grained description can generate videos that are more close to the referenced videos.

Qualitative results. In Fig. 10 and Fig. 11, we provide

Table 5. **Video generation performance on FAVDBench.** FVD [70] is used to measure the generated video from captions and our fine-grained descriptions. As the text-video model is used to directly infer without fine-tuning, the score is higher than normal ranges. Besides, the smaller FVD is the better.

	Captions	Fine-grained Descriptions
FVD (\downarrow)	1,540,010	1,457,802

more video generation examples in our FAVDBench. The videos are generated from captions and our fine-grained descriptions through a state-of-the-art text-video model Cogvideo [20]. FAVDBench can guide video generation more accurately than captions, generating videos that are closer to the reference videos.

Attention Types	B@1	B@4	METEOR	ROUGE.L	CIDEr	Clipscore	EntityScore
Type I	44.10	10.29	18.36	30.93	29.85	69.74	42.71
Type II	19.72	1.97	8.94	23.03	0.89	60.25	20.06
Type III	44.44	10.24	18.39	30.65	29.21	69.91	44.30
Type IV	44.45	10.41	18.47	30.75	29.16	70.19	44.58
Type V	44.32	10.50	18.51	30.86	29.07	70.06	44.41

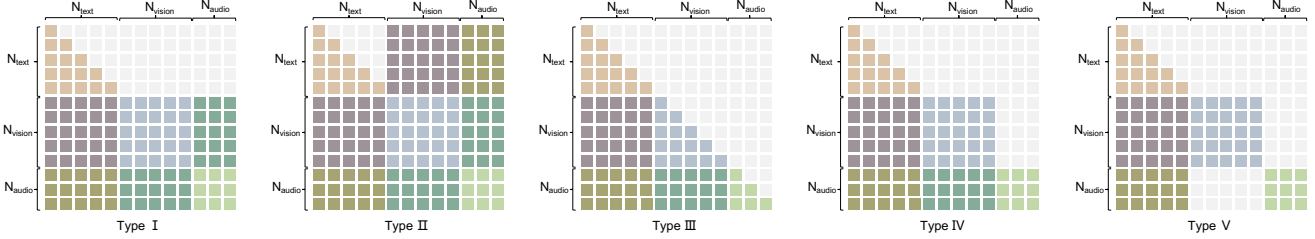


Figure 9. **Impact of transformer attention masks.** 5 different types of attention masks in the table are visualized in bottom figure. The masked attention is colored gray. The attention masks of text, vision, and audio are colored brown, blue, and green, respectively. The type I is set as the default of AVLFormer, where all vision and audio tokens are visible to text, and vision tokens and audio tokens are visible to each other. The type II is full attention among three modalities. Apart from the type I, all text tokens are visible to vision and audio tokens. The type III is single-direction attention, where the tokens in preceding order are visible to preceding tokens, and vice versa. In type IV, vision tokens are visible to audio, and vice versa. In type V, vision tokens and audio tokens are independent.

Table 6. **More quantitative comparisons with different methods on FAVDBench.** We report different backbone settings for PDVC, SwinBERT and AVLFormer. In the backbone freeze (FRZ.) column, - represents that input is not available; → represents that is trainable; * represents frozen. As an extension of Table 2 in the main paper, we report Bleu-1, Bleu-2, Bleu-3, Bleu-4, METEOR, ROUGE.L, CIDEr, Clipscore 1 frame, Reference Clipscore 1 frame, Clipscore 32 frames (in average), Reference Clipscore 32 frames (in average), EntityScore, and AudioScore. Scores of human evaluation are not repeated in this table. For all metrics, higher values are better.

Method	Backbone FRZ.		Conventional Metric						
	Visual	Audio	B@1	B@2	B@3	B@4	METEOR	ROUGE.L	CIDEr
PDVC [78]	→	-	34.91	19.90	11.15	6.33	13.80	24.67	6.27
PDVC w. Audio	→	→	35.59	20.47	11.53	6.47	14.49	26.98	13.13
SwinBERT [35]	→	-	39.68	23.54	14.26	9.05	16.78	29.60	21.69
SwinBERT [35]	*	-	41.41	24.43	14.57	9.08	17.19	29.49	23.03
AVLFormer	→	→	44.10	26.61	16.21	10.29	18.36	30.93	29.85
AVLFormer	*	→	44.10	26.39	16.02	10.23	18.25	30.54	26.31
AVLFormer	→	*	42.82	25.79	15.80	10.16	17.96	30.68	25.45
AVLFormer	*	*	42.35	25.37	15.40	9.84	17.73	30.25	25.65

Method	Backbone FRZ.		Clipscore				Proposed Metric	
	Visual	Audio	Fr@1	Ref. Fr@1	Fr@32	Ref. Fr@32	EntityScore	AudioScore
PDVC [78]	→	-	65.85	66.33	65.77	66.27	32.40	29.37
PDVC w. Audio	→	→	66.35	67.16	66.15	67.04	33.09	40.33
SwinBERT [35]	→	-	68.64	70.02	68.13	69.75	38.17	34.57
SwinBERT [35]	*	-	68.39	70.02	67.98	69.79	39.39	35.06
AVLFormer	→	→	70.24	72.09	69.74	71.82	42.70	52.43
AVLFormer	*	→	69.42	71.30	68.98	71.04	41.69	51.18
AVLFormer	→	*	70.58	71.97	70.07	71.69	41.36	51.72
AVLFormer	*	*	69.82	71.33	69.33	71.05	41.39	51.01

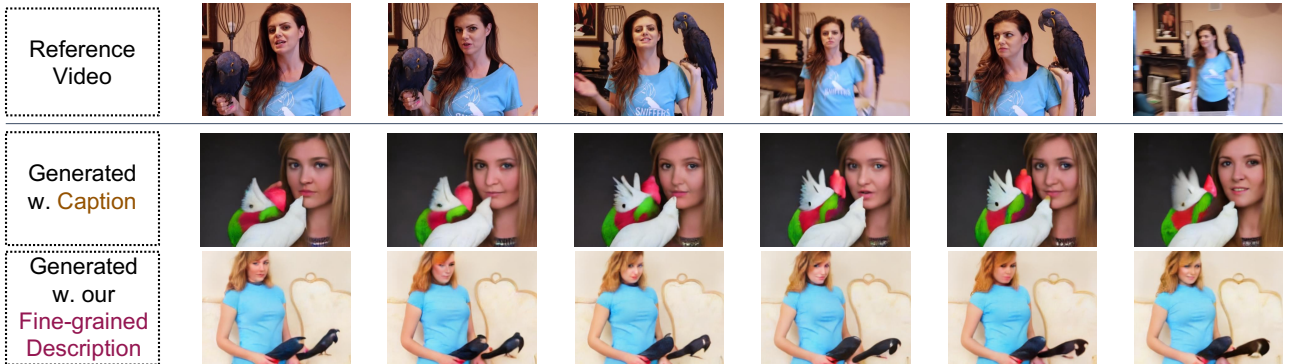
Table 7. **Comparison with different pretrained weights of SwinBERT [35] on FAVDBench.** Pretrained weights from 5 captioning datasets are evaluated: MSR-VTT, MSVD, TVC, VATEX, YOUCOOK II. As a comparison, results of SwinBERT and AVLFormer training with default settings are reported. All pretrained weights are provided by the SwinBERT official repository.

Method	pretrained Weights	Conventional Metric					Proposed Metric	
		B@1	B@4	Meteor	CIDEr	Clipscore	EntityScore	AudioScore
SwinBERT	X	39.68	9.05	16.78	21.69	68.13	39.60	34.57
SwinBERT	MSR-VTT [81]	42.59	9.66	17.84	23.16	70.40	41.41	35.11
SwinBERT	MSVD [5]	42.16	10.15	17.89	26.85	69.69	40.91	34.48
SwinBERT	TVC [29]	42.24	9.67	17.71	24.18	70.33	41.55	35.27
SwinBERT	VATEX [79]	43.03	10.02	18.01	25.59	69.74	41.77	35.44
SwinBERT	YOUCOOK II [90]	42.67	9.05	17.96	27.64	68.13	41.76	35.45
AVLFormer (Ref.)	X	44.10	10.29	18.36	29.85	69.74	44.46	52.43



Caption: A lion was lying on the ground outside, roaring incessantly.

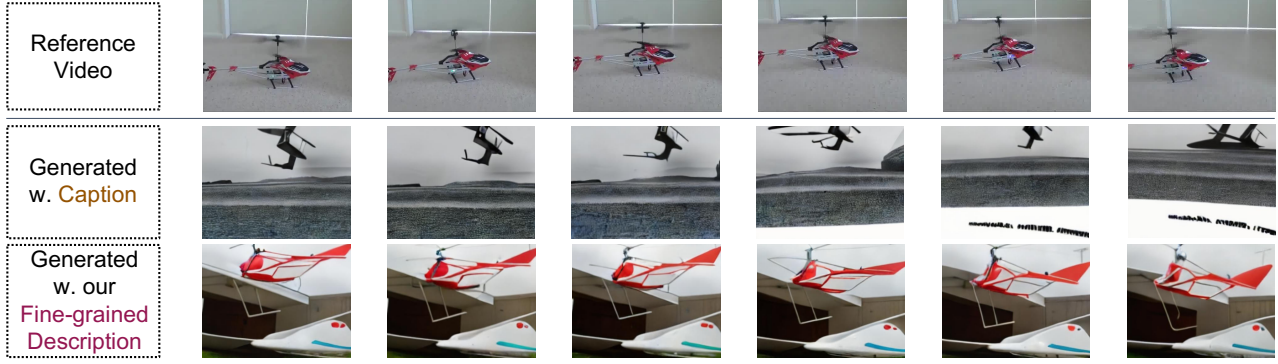
Fine-grained Description: A roaring male lion with a yellow and black mane, dark brown fur, and yellow eyes lay on the ground. The lions are surrounded by lush green vegetation. Behind the lion, there are many trees with yellow and dark green leaves.



Caption: A woman is talking while holding a parrot.

Fine-grained Description: Speaking is a woman with long brown hair, a vibrant manicure, a black waistcoat, and short blue sleeves. A artwork is painted on the apricot-colored wall behind the woman. Behind the woman, there is a sofa. On her hand, the woman is holding a black-feathered parrot.

Figure 10. **More qualitative examples of video generation on FAVDBench (1).** The images in each group are sampled from the ground-truth videos and videos produced by Cogvideo through the caption and our fine-grained descriptions, respectively.

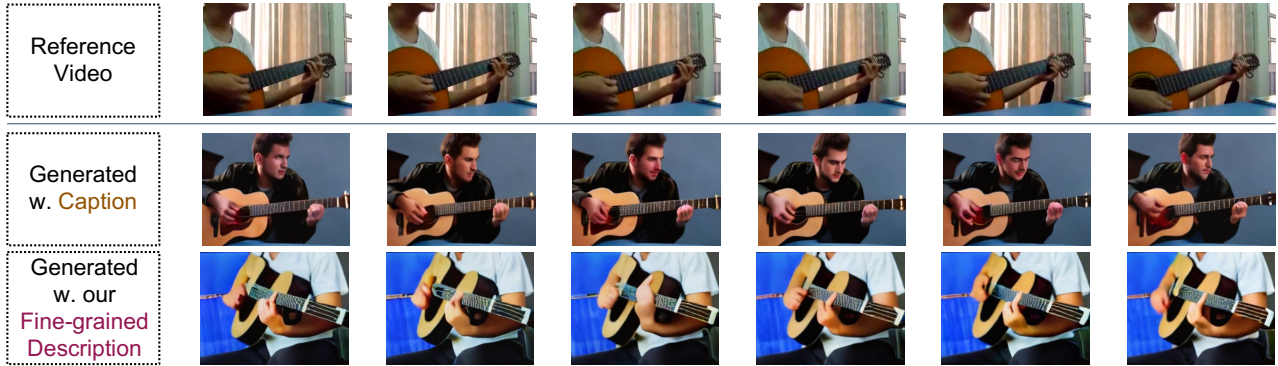


Caption:

From the ground, a remote-controlled helicopter is attempting to take off.

Fine-grained Description:

A remote-controlled helicopter with flashing colored lights inside the fuselage is attempting to take off from the ground. The cockpit is red and black, the middle is silver, the tail is red with white lattice, and there are two sizable propellers on the top. Off-white is the color of the ground where the toy helicopter is situated. A white door and a white wall are located to the left of the model helicopter.



Caption:

A man is singing and playing his guitar while seated at a blue table.

Fine-grained Description:

The guitarist is a man with yellow complexion, wearing white short sleeves, and holding the guitar in both hands. The man is holding a brown top guitar with a black bridge that is being played. Behind the left-facing man, beige drapes are hanging. The man is seated in front of a blue table.

Figure 11. **More qualitative examples of video generation on FAVDBench (2).** The images in each group are sampled from the ground-truth videos and videos produced by Cogvideo through the caption and our fine-grained descriptions, respectively.

References

- [1] Nayyer Aafaq, Naveed Akhtar, Wei Liu, Syed Zulqarnain Gilani, and Ajmal Mian. Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12487–12496, 2019. 2
- [2] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European conference on computer vision (ECCV)*, pages 435–451, 2018. 3
- [3] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *European Conference on Computer Vision*, pages 707–723. Springer, 2022. 1
- [4] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018. 6
- [5] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011. 3, 4, 12
- [6] Shaoxiang Chen, Wenhao Jiang, Wei Liu, and Yu-Gang Jiang. Learning modality interaction for temporal sentence localization and event captioning in videos. In *European Conference on Computer Vision*, pages 333–351. Springer, 2020. 3
- [7] Xuelian Cheng, Huan Xiong, Deng-Ping Fan, Yiran Zhong, Mehrtash Harandi, Tom Drummond, and Zongyuan Ge. Implicit motion handling for video camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13864–13873, 2022. 2
- [8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 6
- [9] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380, Baltimore, Maryland, USA, June 2014. Association for Computational Linguistics. 5
- [10] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014. 7
- [11] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 736–740. IEEE, 2020. 3
- [12] Ayşegül Özkaya Eren and Mustafa Sert. Audio captioning using sound event detection. *arXiv preprint arXiv:2110.01210*, 2021. 3
- [13] Zhiyuan Fang, Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang. Video2commonsense: Generating commonsense descriptions to enrich video captioning. *arXiv preprint arXiv:2003.05162*, 2020. 2
- [14] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017. 6
- [15] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022. 1
- [16] Danna Gurari, Yanan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning images taken by people who are blind. In *European Conference on Computer Vision*, pages 417–434. Springer, 2020. 1
- [17] Sayuri Hayakawa and Viorica Marian. How language shapes the brain. *Scientific Am.*, Apr, 30, 2019. 1
- [18] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021. 7, 9
- [19] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 1
- [20] Wenyi Hong, Ming Ding, Wendi Zheng, Xinghan Liu, and Jie Tang. Cogvideo: Large-scale pretraining for text-to-video generation via transformers. *arXiv preprint arXiv:2205.15868*, 2022. 1, 8, 9, 10
- [21] Mengqi Huang, Zhendong Mao, Penghui Wang, Quan Wang, and Yongdong Zhang. Dse-gan: Dynamic semantic evolution generative adversarial network for text-to-image generation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4345–4354, 2022. 1
- [22] Vladimir Iashin and Esa Rahtu. A better use of audio-visual cues: Dense video captioning with bi-modal transformer. *arXiv preprint arXiv:2005.08271*, 2020. 3, 6, 7, 8
- [23] Vladimir Iashin and Esa Rahtu. Multi-modal dense video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 958–959, 2020. 3
- [24] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132, 2019. 3
- [25] Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, and Gerhard Widmer. Efficient training of audio transformers with patchout. *arXiv preprint arXiv:2110.05069*, 2021. 5, 6
- [26] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017. 3

- [27] Geoffrey Leech. Semantics. England, 1974. 1
- [28] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7331–7341, 2021. 6
- [29] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *European Conference on Computer Vision*, pages 447–463. Springer, 2020. 3, 4, 12
- [30] Bowen Li, Xiaojuan Qi, Thomas Lukasiewicz, and Philip HS Torr. Manigan: Text-guided image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7880–7889, 2020. 1
- [31] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. 6
- [32] Yehao Li, Yingwei Pan, Ting Yao, and Tao Mei. Comprehending and ordering semantics for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17990–17999, 2022. 1
- [33] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. Jointly localizing and describing events for dense video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7492–7500, 2018. 3
- [34] Chin-Yew Lin and Eduard Hovy. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1, NAACL '03*, page 71–78, USA, 2003. Association for Computational Linguistics. 5
- [35] Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. Swinbert: End-to-end transformers with sparse attention for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17949–17958, 2022. 2, 6, 7, 8, 11, 12
- [36] Xubo Liu, Qiushi Huang, Xinhao Mei, Tom Ko, H Lilian Tang, Mark D Plumbley, and Wenwu Wang. Cl4ac: A contrastive loss for audio captioning. *arXiv preprint arXiv:2107.09990*, 2021. 3
- [37] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022. 6
- [38] Kaiyue Lu, Zexiang Liu, Jianyuan Wang, Weixuan Sun, Zhen Qin, Dong Li, Xuyang Shen, Hui Deng, Xiaodong Han, Yuchao Dai, et al. Linear video transformer with feature fixation. *arXiv preprint arXiv:2210.08164*, 2022. 2
- [39] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020. 2
- [40] Xinhao Mei, Qiushi Huang, Xubo Liu, Gengyun Chen, Jingqian Wu, Yusong Wu, Jinzheng Zhao, Shengchen Li, Tom Ko, H Lilian Tang, et al. An encoder-decoder based audio captioning system with transfer and reinforcement learning. *arXiv preprint arXiv:2108.02752*, 2021. 3
- [41] Xinhao Mei, Xubo Liu, Qiushi Huang, Mark D Plumbley, and Wenwu Wang. Audio captioning transformer. *arXiv preprint arXiv:2107.09817*, 2021. 3
- [42] Mathew Monfort, SouYoung Jin, Alexander Liu, David Harwath, Rogerio Feris, James Glass, and Aude Oliva. Spoken moments: Learning joint audio-visual representations from video descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14871–14881, 2021. 3
- [43] Jonghwan Mun, Linjie Yang, Zhou Ren, Ning Xu, and Bohyung Han. Streamlined dense video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6588–6597, 2019. 3
- [44] Seonghyeon Nam, Yunji Kim, and Seon Joo Kim. Text-adaptive generative adversarial networks: manipulating images with natural language. *Advances in neural information processing systems*, 31, 2018. 1
- [45] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24, 2011. 1
- [46] Gloria Origgi and Dan Sperber. Evolution, communication and the proper function of language. *Evolution and the human mind: Language, modularity and social cognition*, pages 140–169, 2000. 1
- [47] Boxiao Pan, Haoye Cai, De-An Huang, Kuan-Hui Lee, Adrien Gaidon, Ehsan Adeli, and Juan Carlos Nibbles. Spatio-temporal graph for video captioning with knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10870–10879, 2020. 3
- [48] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10971–10980, 2020. 1
- [49] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. 5
- [50] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002. 7
- [51] Wenjie Pei, Jiayuan Zhang, Xiangrong Wang, Lei Ke, Xiaoyong Shen, and Yu-Wing Tai. Memory-attended recurrent network for video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8347–8356, 2019. 1, 2
- [52] Zhen Qin, Xiaodong Han, Weixuan Sun, Bowen He, Dong Li, Dongxu Li, Yuchao Dai, Lingpeng Kong, and Yiran

- Zhong. Toeplitz neural network for sequence modeling. In *The Eleventh International Conference on Learning Representations*, 2023. 8
- [53] Zhen Qin, XiaoDong Han, Weixuan Sun, Dongxu Li, Lingpeng Kong, Nick Barnes, and Yiran Zhong. The devil in linear transformer. *arXiv preprint arXiv:2210.10340*, 2022. 2
- [54] Zhen Qin, Weixuan Sun, Hui Deng, Dongxu Li, Yunshen Wei, Baohong Lv, Junjie Yan, Lingpeng Kong, and Yiran Zhong. cosformer: Rethinking softmax in attention. *arXiv preprint arXiv:2202.08791*, 2022. 8
- [55] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 5
- [56] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. 5
- [57] Tanzila Rahman, Bicheng Xu, and Leonid Sigal. Watch, listen and tell: Multi-modal weakly supervised dense event captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8908–8917, 2019. 3
- [58] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 1
- [59] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 1
- [60] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212, 2015. 3
- [61] Lin CY ROUGE. A package for automatic evaluation of summaries. In *Proceedings of Workshop on Text Summarization of ACL, Spain*, 2004. 7
- [62] Shulan Ruan, Yong Zhang, Kun Zhang, Yanbo Fan, Fan Tang, Qi Liu, and Enhong Chen. Dae-gan: Dynamic aspect-aware gan for text-to-image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13960–13969, 2021. 1
- [63] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 1, 8
- [64] Peipei Song, Dan Guo, Jun Cheng, and Meng Wang. Contextual attention network for emotional video captioning. *IEEE Transactions on Multimedia*, 2022. 1
- [65] Peipei Song, Dan Guo, Jinxing Zhou, Mingliang Xu, and Meng Wang. Memorial gan with joint semantic optimization for unpaired image captioning. *IEEE Transactions on Cybernetics*, 2022. 1
- [66] Jiadai Sun, Yuxin Mao, Yuchao Dai, Yiran Zhong, and Jianyuan Wang. Munet: Motion uncertainty-aware semi-supervised video object segmentation. *Pattern Recognition*, 138:109399, 2023. 3
- [67] Weixuan Sun, Zhen Qin, Hui Deng, Jianyuan Wang, Yi Zhang, Kaihao Zhang, Nick Barnes, Stan Birchfield, Lingpeng Kong, and Yiran Zhong. Vicinity vision transformer. *arXiv preprint arXiv:2206.10552*, 2022. 8
- [68] Weixuan Sun, Jing Zhang, Zheyuan Liu, Yiran Zhong, and Nick Barnes. Getam: Gradient-weighted element-wise transformer attention map for weakly-supervised semantic segmentation. *arXiv preprint arXiv:2112.02841*, 2021. 3
- [69] Mingkan Tang, Zhanyu Wang, Zhenhua Liu, Fengyun Rao, Dian Li, and Xiu Li. Clip4caption: Clip for video caption. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 4858–4862, 2021. 2
- [70] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018. 9, 10
- [71] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 7
- [72] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2015. 5
- [73] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. Sequence to sequence-video to text. In *Proceedings of the IEEE international conference on computer vision*, pages 4534–4542, 2015. 2
- [74] Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022. 1
- [75] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 1
- [76] Bairui Wang, Lin Ma, Wei Zhang, and Wei Liu. Reconstruction network for video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7622–7631, 2018. 1, 2
- [77] Jingwen Wang, Wenhao Jiang, Lin Ma, Wei Liu, and Yong Xu. Bidirectional attentive fusion with context gating for dense video captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7190–7198, 2018. 3
- [78] Teng Wang, Ruimao Zhang, Zhichao Lu, Feng Zheng, Ran Cheng, and Ping Luo. End-to-end dense video captioning with parallel decoding. In *Proceedings of the IEEE/CVF*

- International Conference on Computer Vision*, pages 6847–6857, 2021. 3, 6, 7, 8, 11
- [79] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4581–4591, 2019. 3, 4, 12
- [80] Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. Nüwa: Visual synthesis pre-training for neural visual world creation. In *European Conference on Computer Vision*, pages 720–736. Springer, 2022. 1
- [81] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 3, 4, 12
- [82] Jun Xu, Ting Yao, Yongdong Zhang, and Tao Mei. Learning multimodal attention lstm networks for video captioning. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 537–545, 2017. 2
- [83] Zhongjie Ye, Helin Wang, Dongchao Yang, and Yuexian Zou. Improving the performance of automated audio captioning via integrating the acoustic and textual information. *DCASE2021 Challenge, Tech. Rep.*, 2021. 3
- [84] Jing Zhang, Yunqiu Lv, Mochu Xiang, Aixuan Li, Yuchao Dai, and Yiran Zhong. Depth confidence-aware camouflaged object detection. *arXiv e-prints*, pages arXiv–2106, 2021. 2
- [85] Ziqi Zhang, Yaya Shi, Chunfeng Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zheng-Jun Zha. Object relational graph with teacher-recommended learning for video captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13278–13288, 2020. 1, 3
- [86] Yiran Zhong, Yuchao Dai, and Hongdong Li. 3d geometry-aware semantic labeling of outdoor street scenes. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2343–2349. IEEE, 2018. 2
- [87] Jinxing Zhou, Dan Guo, Yiran Zhong, and Meng Wang. Improving audio-visual video parsing with pseudo visual labels. *arXiv preprint arXiv:2303.02344*, 2023. 3
- [88] Jinxing Zhou, Xuyang Shen, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, et al. Audio-visual segmentation with semantics. *arXiv preprint arXiv:2301.13190*, 2023. 1
- [89] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio-visual segmentation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, pages 386–403. Springer, 2022. 1
- [90] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI Conference on Artificial Intelligence*, pages 7590–7598, 2018. 3, 4, 12
- [91] Luowei Zhou, Yingbo Zhou, Jason J Corso, Richard Socher, and Caiming Xiong. End-to-end dense video captioning with masked transformer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8739–8748, 2018. 3
- [92] Yuanen Zhou, Yong Zhang, Zhenzhen Hu, and Meng Wang. Semi-autoregressive transformer for image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3139–3143, 2021. 1