Fine-grained Audible Video Description
– Supplementary Material –

In this supplemental material, 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 ("FAVDBench_Demo.mp4") containing some samples in FAVDBench is attached to this supplementary material.

A. The FAVDBench

Word Frequency. As FAVDBench provides both English and Chinese descriptions, we plot the frequency of English and Chinese vocabularies in Fig. 1. The most common words in descriptions, such "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.

B. Experimental Results

Impact of different AVLFormer attention masks. We ablate the impact of different AVLFormer attention masks in Fig. 2. 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).

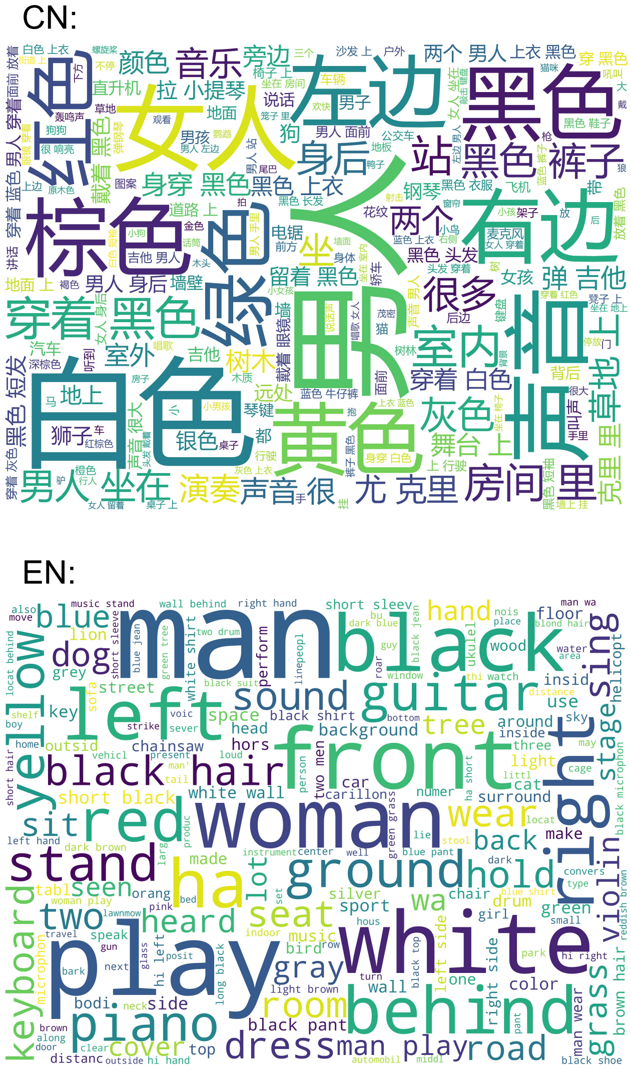


Figure 1. **Word Frequency.** We count the word frequency of Chinese annotation (upper) and English annotation (bottom) on FAVDBench.

Extended quantitative results. We report 6 additional metrics together with the 8 metrics from the main paper in Table 2. We extend the Clipscore [2] 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 vari-

Table 1. **Video generation performance on FAVDBench.** FVD [7] 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 (↓)	1,540,010	**1,457,802**

ous pretrained weights and fine-tuning on FAVDBench in Table 3. Pretrained weights from other captioning dataset contribute to the performance improvement, especially the VATEX. As a comparison, AVLFormer beats all settings of SwinBERT over all evaluation metrics.

C. 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) [7] metric in Table 1. We employ Cogvideo [3] 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. 3 and Fig. 4, we provide 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 [3]. FAVDBench can guide video generation more accurately than captions, generating videos that are closer to the reference videos.

References

[1] 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. 4

[2] 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. 1

[3] 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. 2

[4] 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

[5] 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. 4

[6] 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. 3, 4

[7] 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. 2

[8] 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

[9] 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. 4

[10] 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. 4

[11] 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. 4

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

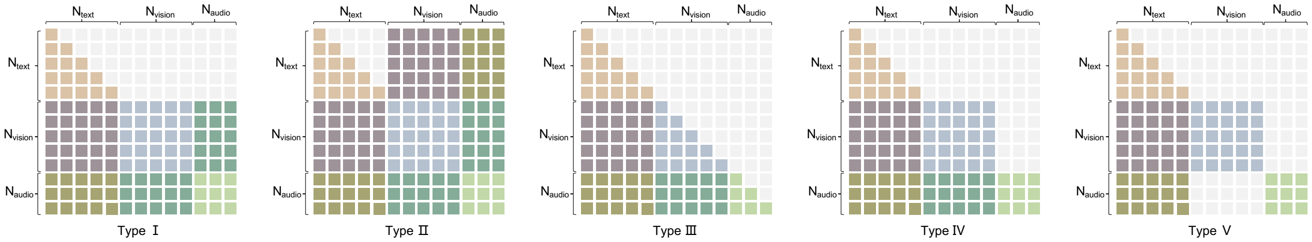


Figure 2. **Impact of transformer attention masks.** 5 different types of attention masks in the table are visualized in bottom figure. The masked attention is colored gray. The attention masks of text, vision, and audio are colored brown, blue, and green, respectively. The type I is set as the default 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.

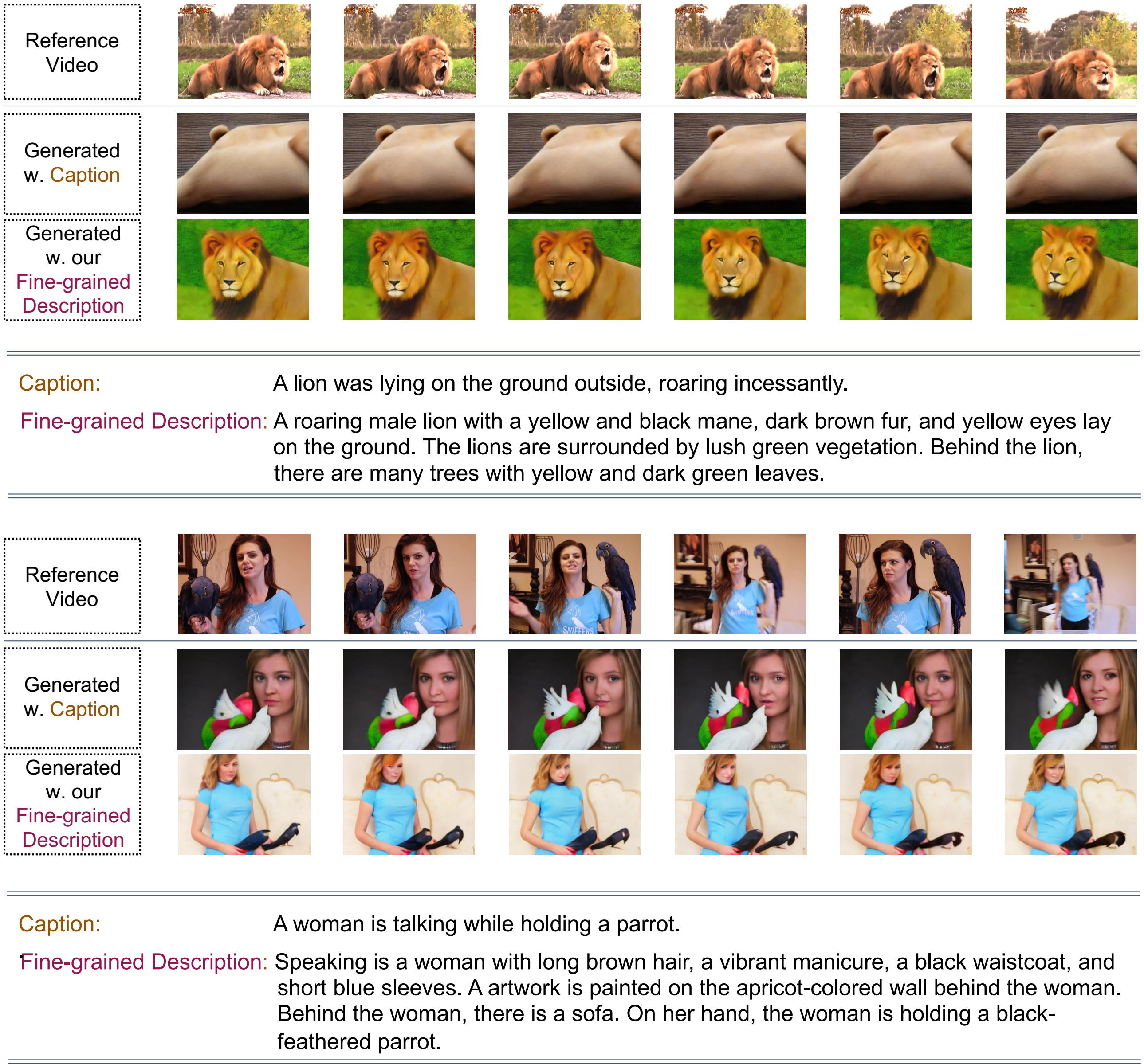
Table 2. **More quantitative comparisons with different methods on FAVDBench.** We report different backbone settings for PDVC, SwinBERT and AVLFormer. In the backbone freeze (FRZ.) column, **-** represents that input is not available; ➜ represents that is trainable; ✻ represents frozen. As an extension 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 [8]	➜	-	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 [6]	➜	-	39.68	23.54	14.26	9.05	16.78	29.60	21.69
SwinBERT [6]	✻	-	41.41	24.43	14.57	9.08	17.17	29.49	23.03
BMT [4]	✻	✻	41.28	24.38	14.76	9.23	16.57	30.12	16.46
AVLFormer	➜	➜	**44.10**	**26.61**	**16.21**	**10.29**	**18.36**	**30.93**	**29.85**
AVLFormer	✻	➜	**44.10**	26.39	16.02	10.23	18.24	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 [8]	➜	-	65.85	66.33	65.77	66.27	32.40	56.88
PDVC w. Audio	➜	➜	66.35	67.16	66.15	67.04	33.09	60.72
SwinBERT [6]	➜	-	68.64	70.02	68.13	69.75	38.17	62.60
SwinBERT [6]	✻	-	68.39	70.02	67.98	69.79	39.39	58.72
BMT [4]	✻	✻	65.00	67.25	64.51	66.99	35.99	61.03
AVLFormer	➜	➜	70.24	**72.09**	69.74	**71.82**	**42.70**	**63.88**
AVLFormer	✻	➜	69.42.	71.30	68.98	71.04	41.69	61.53
AVLFormer	➜	✻	**70.58**	71.97	**70.07**	71.69	41.36	62.84
AVLFormer	✻	✻	69.82	71.33	69.33	71.05	41.39	61.90

Table 3. **Comparison with different pretrained weights of SwinBERT [6] 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	✘	39.68	9.05	16.78	21.69	68.13	39.60	62.60
SwinBERT	MSR-VTT [10]	42.59	9.66	17.84	23.16	**70.40**	41.41	61.94
SwinBERT	MSVD [1]	42.16	**10.15**	17.89	26.85	69.69	40.91	**62.90**
SwinBERT	TVC [5]	42.24	9.67	17.71	24.18	70.33	41.55	61.62
SwinBERT	VATEX [9]	**43.03**	10.02	**18.01**	25.59	69.74	**41.77**	62.28
SwinBERT	YOUCOOK II [11]	42.67	9.05	17.96	**27.64**	68.13	41.76	62.22
AVLFormer (Ref.)	✘	**44.10**	**10.29**	**18.36**	**29.85**	**69.74**	**44.46**	**63.88**


Reference Video
Generated w. Caption
Generated w. our Fine-grained Description
Reference Video
Generated w. Caption
Generated w. our Fine-grained Description

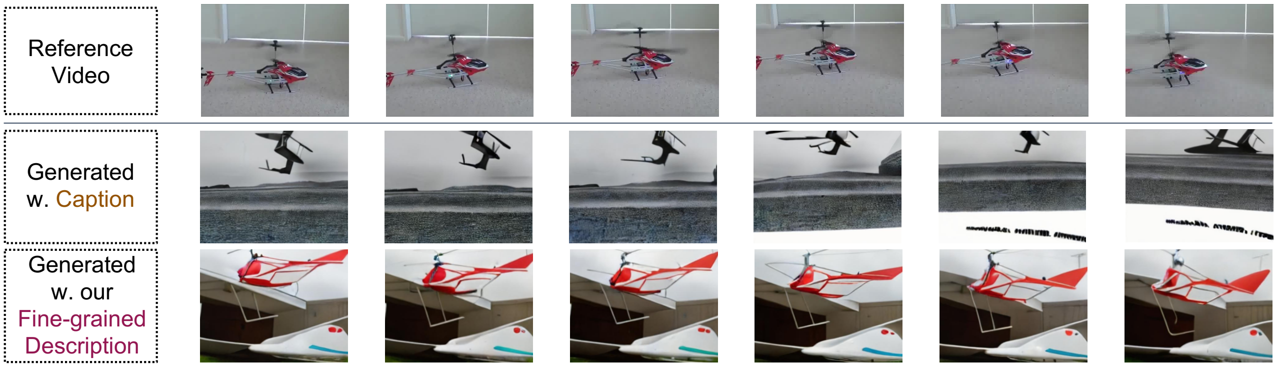
Caption: A lion was lying on the ground outside, roaring incessantly.

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.

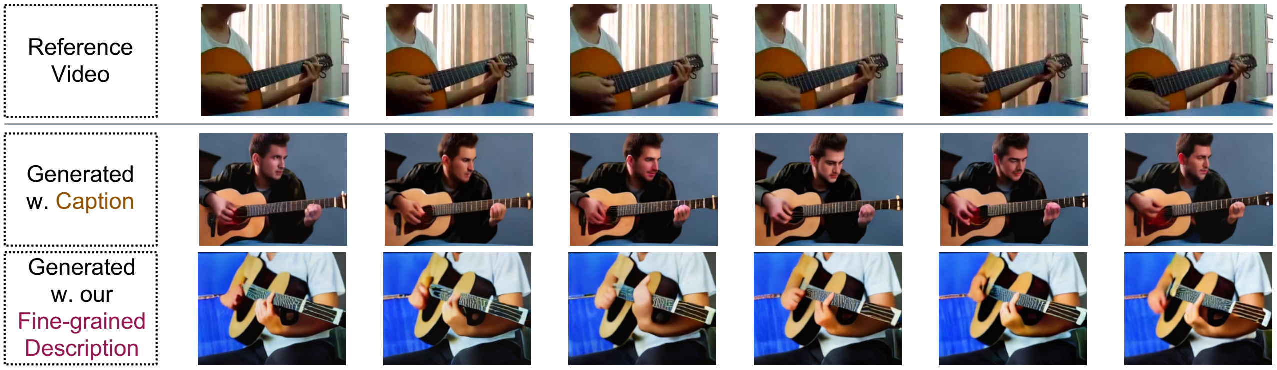
Fine-grained Description: Speaking is a woman with long brown hair, a vibrant manicure, a black waistcoat, and short blue sleeves. A artwork is painted on the apricot-colored wall behind the woman. Behind the woman, there is a sofa. On her hand, the woman is holding a black-feathered parrot.

Figure 3. **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 4. **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.