

2 IMPLEMENTATION DETAILS

Dataset. We use two downstream datasets, SVAD as primary dataset and Youku-mplug caption [14] as auxiliary dataset. Pretraining is conducted on an advertisement short video dataset, in which videos are sampled from the same short video platform with the SVAD dataset. This pretraining dataset is significantly larger than the SVAD dataset, encompassing around 1 million advertisement entries. We generate pseudo-labels using existing image caption models [5] and get temporal information by concatenating queries from several frames.

Experimental Settings. The Q-Former parameter from BLIP2 [5] is employed to initialize our Q-Former. GMHRA [7] and learnable queries are initialized with VideoChat [6] parameters, while the dual alignment layer undergoes random initialization. Only the aforementioned parameters are trained, with the remainder being frozen. For the vision encoder, ViT-g/14 from EVA-CLIP [12] is utilized alongside Baichuan-13B [16] for the LLM. Each video is processed to sample eight frames at equal intervals for visual feature extraction. The batch size is set at 32 for pretraining and reduced to 16 for fine-tuning. Model optimization is achieved using AdamW [10], coupled with a linear warm-up and cosine learning rate adjustment. The initial learning rate for the Q-Former is set to 3e-6, while that for other parameters is set to 3e-5. We train SVAD-VLM 1 epoch for pretraining and 3 epoch for fine-tuning. Training commences with the pre-training dataset, followed by employing a mixed-dataset approach as detailed in our methodology in main text.

Text-to-video Retrieval Details. We obtain video representations using a Vision Transformer (ViT) [4] with temporal attention, which models the sequential relationships of the frames. Drawing on the method of CLIP-ViP [15], we aggregate the video representations onto a single video proxy token. We then use the video frames and the video's ASR (Automatic Speech Recognition) to train this model. We adopt video-text contrastive learning loss as our training objective. For a reference video, we generate general caption and fine-grained description are used by BLIP2 [5] and SVAD-VLM. Text-to-video retrieval experiments are then conducted using these two types of text message.

3 QUALITATIVE RESULTS

Video description. Figure 8 illustrates more qualitative examples, where we compare SVAD-VLM with the BLIP2 [5], Qwen-vl [1], LLaVA(v1.5) [9], VideoChat [6], GVT [2] and Video-LLaVA [8]. SVAD-VLM outperforms all competitors by providing accurate visual details like "cheongsam" and "ponytail" in Figure 8a and accurately describing movements of the woman in Figure 8b. This demonstrates that SVAD-VLM extracts key information due to prompt-guided keyword generation task and shows strong generalization capability benefited from mixed-datasets training.

Text-to-video Retrieval. In Figure 9, we provide qualitative results of general caption and our fine-grained description on text-to-video retrieval. Fine-grained descriptions outperform general captions, demonstrating that spatiotemporal fine-grained video description task is more suitable for short videos by providing a more detailed description both spatially and temporally.

REFERENCES

- [1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A Frontier Large Vision-Language Model with Versatile Abilities. *CoRR abs/2308.12966* (2023). <https://doi.org/10.48550/ARXIV.2308.12966> arXiv:2308.12966
- [2] Aozhu Chen, Ziyuan Wang, Chengbo Dong, Kaibin Tian, Ruixiang Zhao, Xun Liang, Zhanhui Kang, and Xirong Li. 2023. ChinaOpen: A Dataset for Open-world Multimodal Learning. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, Abdullmotaheb El-Saddik, Tao Mei, Rita Cucchiara, Marco Bertini, Diana Patricia Tobon Vallejo, Pradeep K. Atrey, and M. Shamim Hossain (Eds.). ACM, 6432–6440. <https://doi.org/10.1145/3581783.3612156>
- [3] David L. Chen and William B. Dolan. 2011. Collecting Highly Parallel Data for Paraphrase Evaluation. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, Dekang Lin, Yuji Matsumoto, and Rada Mihalcea (Eds.). The Association for Computer Linguistics, 190–200. <https://aclanthology.org/P11-1020/>
- [4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xi-aohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
- [5] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597* (2023).
- [6] Kunchang Li, Yanan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2023. VideoChat: Chat-Centric Video Understanding. *CoRR abs/2305.06355* (2023). <https://doi.org/10.48550/ARXIV.2305.06355> arXiv:2305.06355
- [7] Kunchang Li, Yali Wang, Yanan He, Yizhuo Li, Yi Wang, Limin Wang, and Yu Qiao. 2023. UniFormerV2: Unlocking the Potential of Image ViTs for Video Understanding. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*. IEEE, 1632–1643. <https://doi.org/10.1109/ICCV51070.2023.00157>
- [8] Bin Lin, Yang Ye, Bin Zhu, Jiayi Cui, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. *CoRR abs/2311.10122* (2023). <https://doi.org/10.48550/ARXIV.2311.10122> arXiv:2311.10122
- [9] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved Baselines with Visual Instruction Tuning. *CoRR abs/2310.03744* (2023). <https://doi.org/10.48550/ARXIV.2310.03744> arXiv:2310.03744
- [10] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).
- [11] Xuyang Shen, Dong Li, Jinxing Zhou, Zhen Qin, Bowen He, Xiaodong Han, Aixuan Li, Yuchao Dai, Lingpeng Kong, Meng Wang, Yu Qiao, and Yiran Zhong. 2023. Fine-grained Audible Video Description. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 10585–10596. <https://doi.org/10.1109/CVPR52729.2023.01020>
- [12] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389* (2023).
- [13] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. VaTeX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 4580–4590. <https://doi.org/10.1109/ICCV.2019.00468>
- [14] Haiyang Xu, Qinghao Ye, Xuan Wu, Ming Yan, Yuan Miao, Jiabo Ye, Guohai Xu, Anwen Hu, Yaya Shi, Guangwei Xu, Chenliang Li, Qi Qian, Maofei Que, Ji Zhang, Xiao Zeng, and Fei Huang. 2023. Youku-mPLUG: A 10 Million Large-scale Chinese Video-Language Dataset for Pre-training and Benchmarks. *CoRR abs/2306.04362* (2023). <https://doi.org/10.48550/ARXIV.2306.04362> arXiv:2306.04362
- [15] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. 2023. CLIP-ViP: Adapting Pre-trained Image-Text Model to Video-Language Alignment. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net. <https://openreview.net/pdf?id=GNjzMAgawq>
- [16] Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, Fan Yang, Fei Deng, Feng Wang, Feng Liu, Guangwei Ai, Guosheng Dong, Haizhou Zhao, Hang Xu, Haoze Sun, Hongda Zhang, Hui Liu, Jiaming Ji, Jian Xie, Juntao Dai, Kun Fang, Lei Su, Liang Song, Lifeng Liu, Liyun Ru, Luyao Ma, Mang Wang, Mickel Liu, MingAn Lin, Nuolan Nie, Peidong Guo, Ruiyang Sun, Tao Zhang, Tianpeng Li, Tianyu Li, Wei Cheng, Weipeng Chen, Xiangrong Zeng, Xiaochuan Wang, Xiaoxi Chen, Xin Men, Xin Yu, Xuehai Pan, Yanjun Shen, Yiding Wang, Yiyi Li, Youxin Jiang, Yuchen Gao, Yupeng Zhang, Zenan Zhou, and Zhiying Wu. 2023. Baichuan 2: Open Large-scale Language Models. *CoRR abs/2309.10305* (2023). <https://doi.org/10.48550/ARXIV.2309.10305> arXiv:2309.10305

117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232

INSTRUCTION:

1. Describe all significant individuals and movements presented in the video.
2. Describe the setting of the event if it is clear; otherwise, omit.
3. If there are no human figures in the video, describe the main subject.
4. If there is no clear movement, only describe the characters or main subject.
5. If multiple characters or subjects appear sequentially, label them accordingly; separate them with semicolons.
6. Describe details related to the main subject, omit unimportant details.
7. Each sentence should contain at least fifteen words.
8. Avoid typing or grammatical errors.
9. Maintain objectivity, refrain from including personal feelings, and do not use "I" or "my."
10. Do not use phrases like "here is" or "there is."
11. Do not describe events that occurred in the past or are expected to happen in the future.
12. Do not assign names to people on your own; if specific names are mentioned in the video, do not label them.

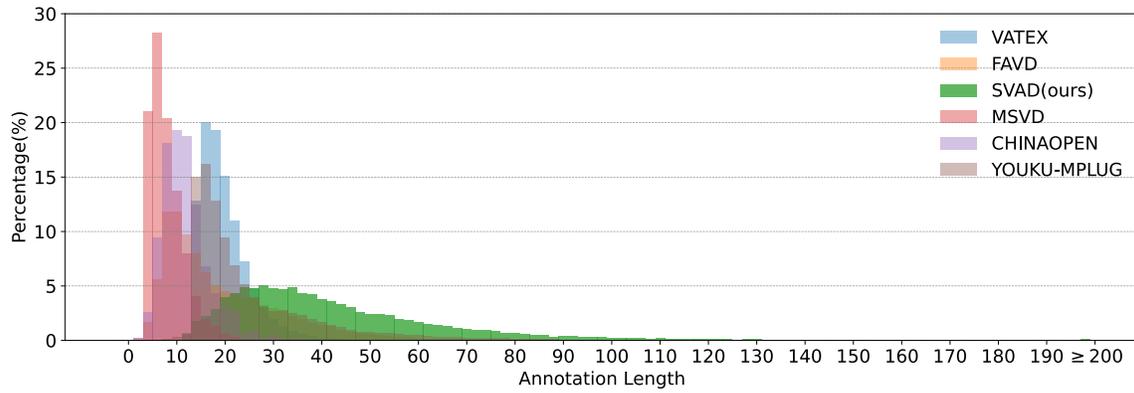
EXAMPLES:

Annotation: A man holds a glass with light green tea soup and green tea leaves inside. A man uses a white plate to scoop up raw tea leaves in a white woven bag for display.

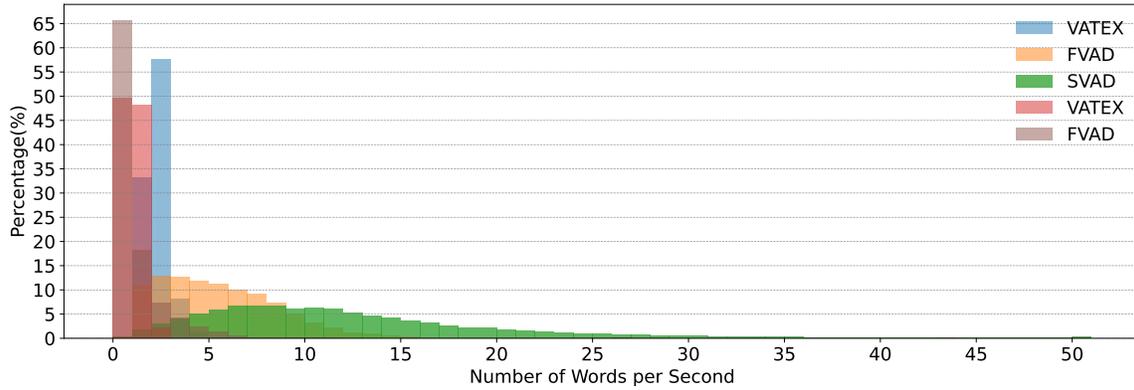


Annotation: In the shooting game, there are multiple flames in the scene, and a man wearing a red suit, carrying a black gun and a tan backpack opens a box and picks up the game equipment.

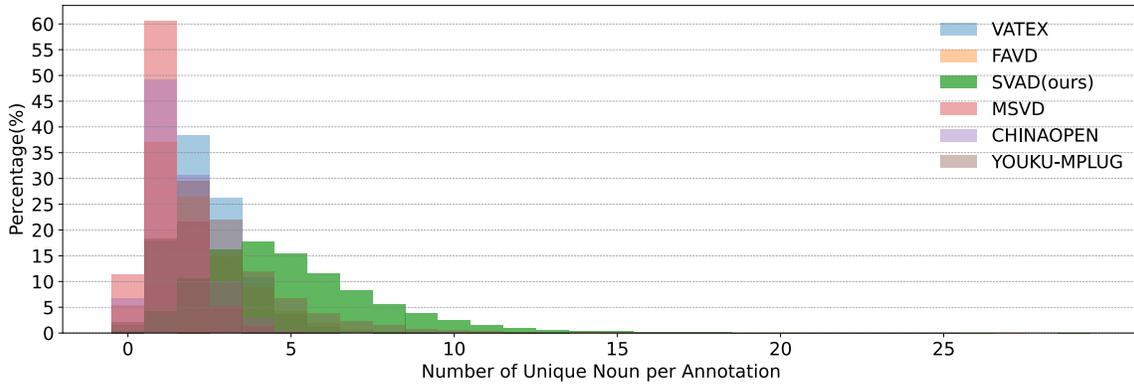
Figure 2: Annotation instructions and examples.



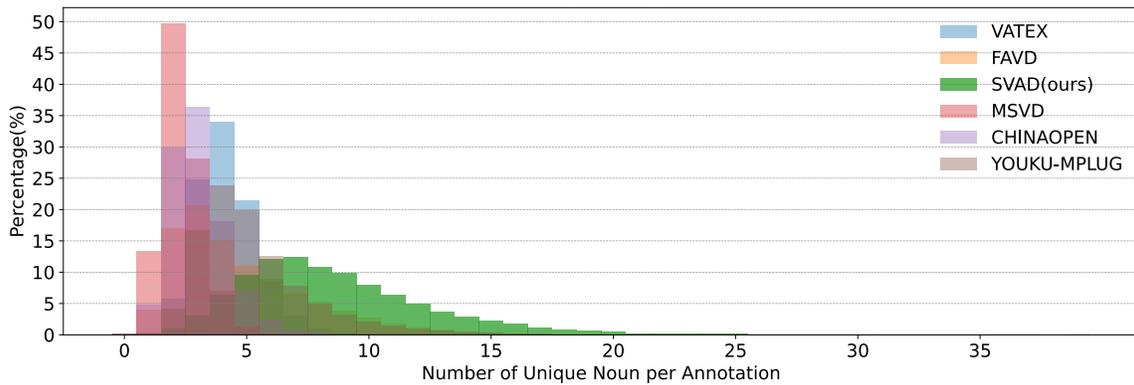
(a) Distributions of annotation lengths.



(b) Distributions of number of word percent.



(c) Distributions of unique verb per annotation.



(d) Distributions of unique noun per annotation

Figure 3: Statistical histogram distributions on VATEX [13], FAVD [11], MSVD [3], Chinaopen [2], Youku-mplug Caption (YOUKU) [14] and SVAD. Compared to other datasets, distributions of SVAD shift to the right, which means its annotations are longer and richer in verbs and nouns, with a higher text information density.

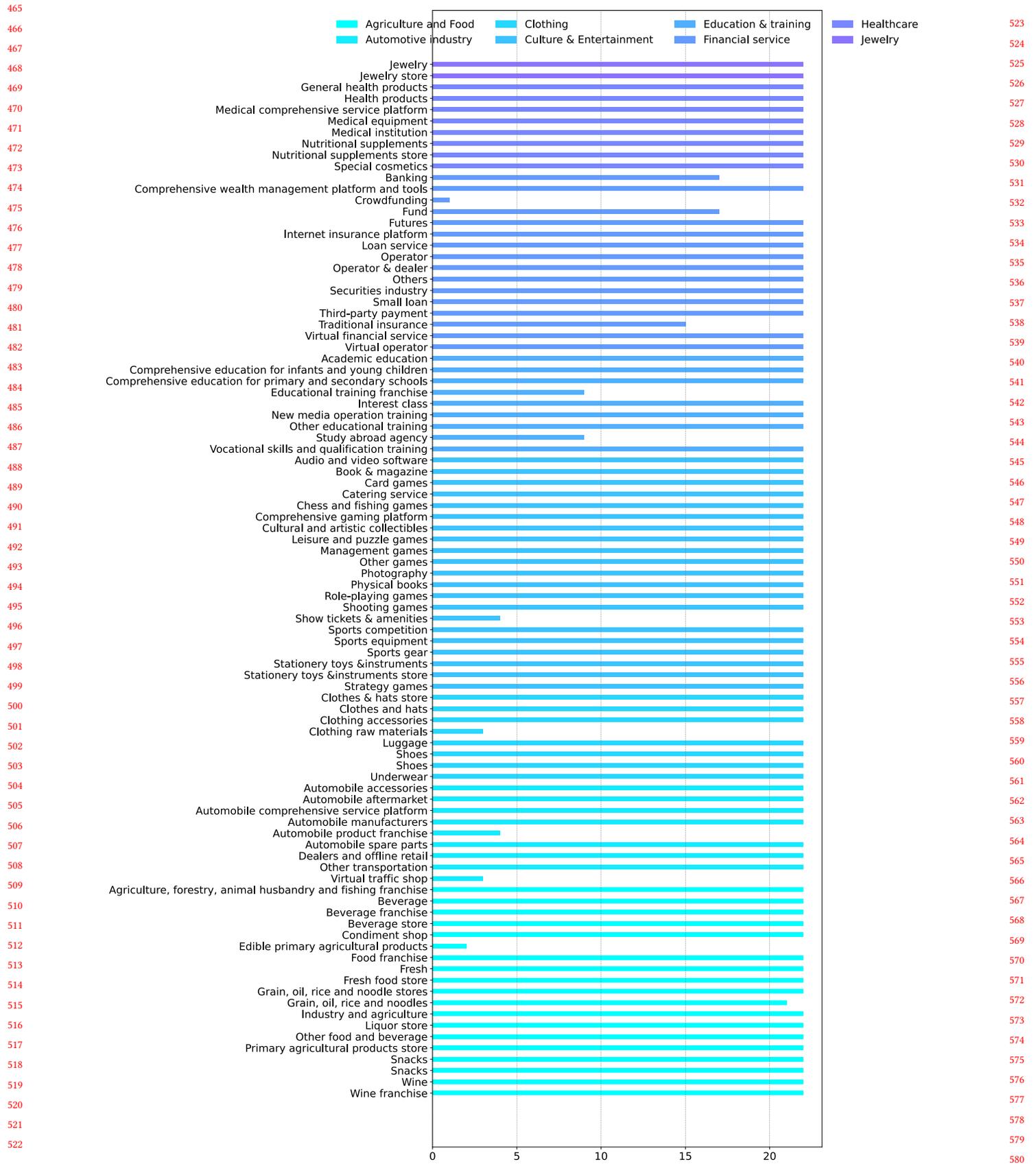


Figure 4: Statistical histograms of videos from 191 sub-industries across 15 industry categories (first 8 categories).

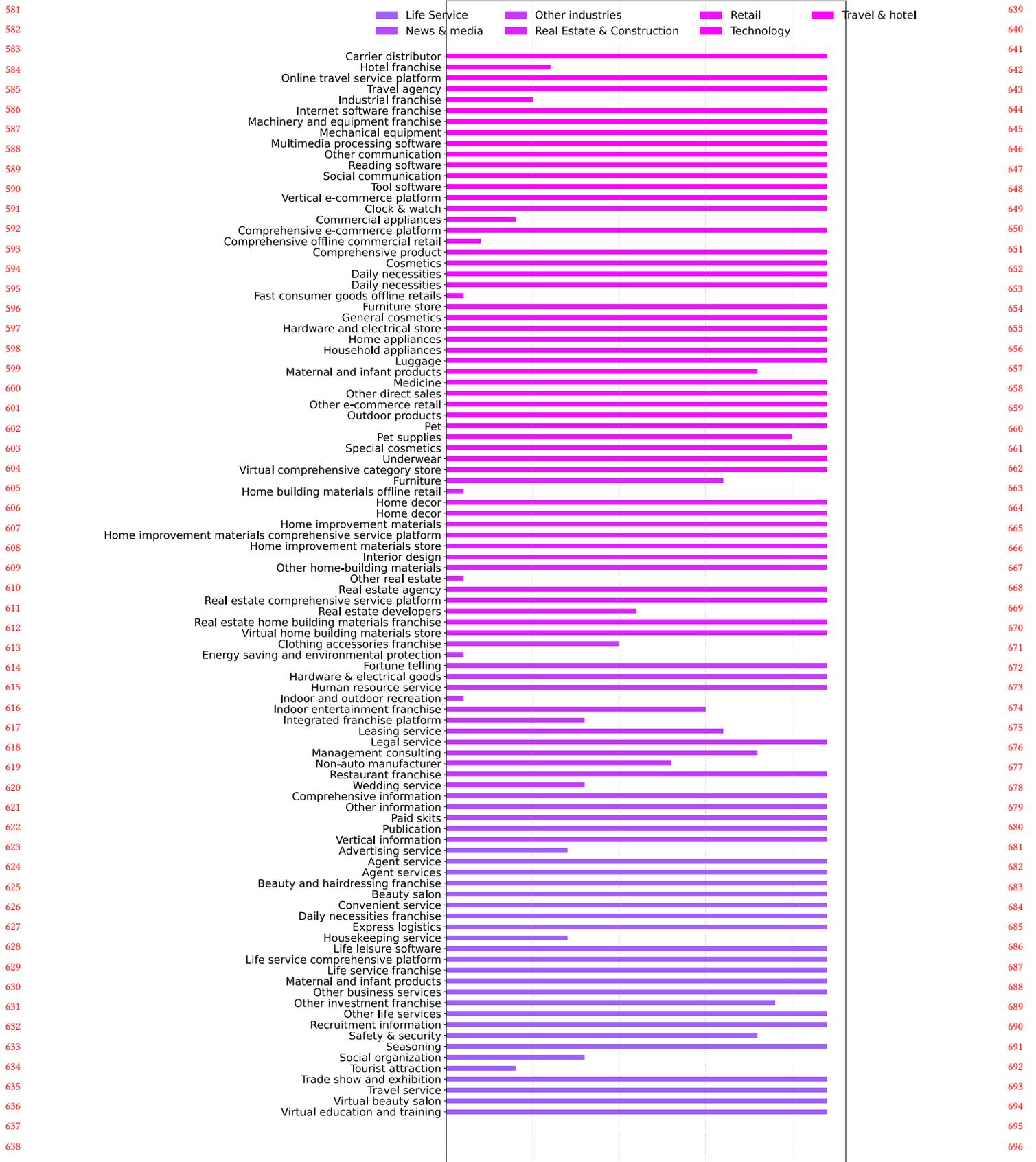


Figure 5: Statistical histograms of videos from 191 sub-industries across 15 industry categories (last 7 categories).

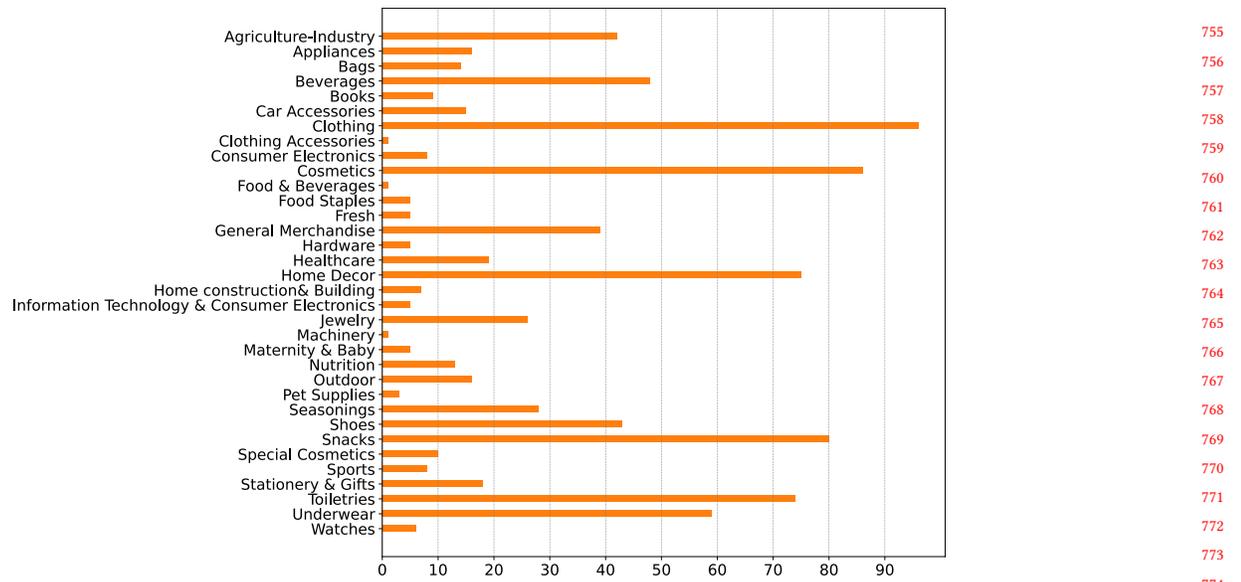


Figure 6: Statistical histograms of videos across different product categories.

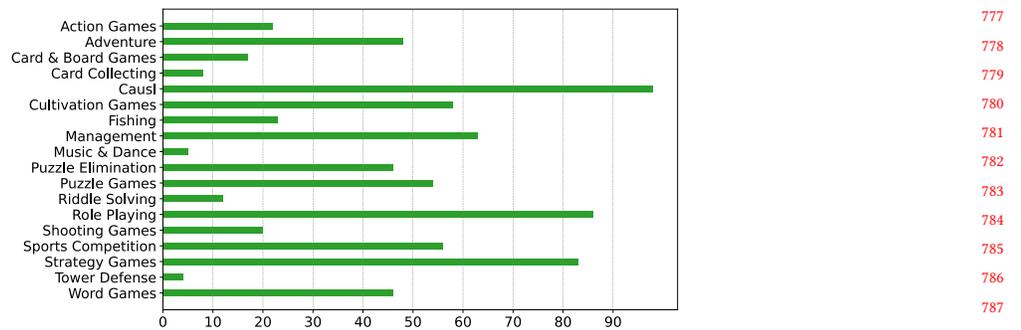


Figure 7: Statistical histograms of videos across different game categories.



Reference: A woman in a pink cheongsam with a ponytail is sitting on a chair playing the erhu.

BLIP2: A Chinese woman is playing a **musical instrument**.

Qwen-v1: In the picture, a woman in a pink cheongsam is playing the erhu. A **huqin stool** is placed under the erhu, along with an erhu and some **music books**.

LLaVA (v1.5): The image depicts a woman dressed in traditional Chinese clothing, sitting on a chair and **holding a long, thin musical instrument**. She appears to be playing this instrument, probably a **violin or similar stringed instrument**.

VideoChat: A woman in fancy clothes is playing a **violin** on a pink wall. On the back are the words “**how to adjust the violin frequency**.”

GVT: A woman in a cheongsam is sitting on a chair, holding a **black wind controller** in her right hand.

VideoLLaVA: A woman dressed in pink, holding a **long black stick**, **speaks** to the camera.

SVAD-VLM: In front of a pink background, a woman in a **cheongsam** with a **ponytail**, sitting on a chair, is **playing the erhu**. She is **holding** the headstock of the erhu **with the left hand** and **pulling** the bow **with the right hand**.

(a)



Reference: A woman dressed in gray opens a brown fanny bag tied around her waist, pulls her cell phone and perfume out of the bag, pulls the key, and shakes it from side to side to show a brown fanny waist.

BLIP2: A woman is **holding a cell phone**.

Qwen-v1: This picture is an advertising copy for a product. At the center is a female model in a gray dress holding a **cell phone bag** made of coffee-colored leather material with lanyards for crossbody use and a **separate space** inside the bag for the phone.

LLaVA (v1.5): The image shows a woman wearing a **black dress** and a brown **handbag**. She has a **cell phone in her hand**. The woman is standing in a room with **handbags on the floor nearby**. There are several **other handbags** in the scene.

VideoChat: A woman wearing a gray coat and a gray skirt holds a brown fanny pack in her left hand, taps the leather of the fanny pack with her fingers, and then puts the cellphone into it.

GVT: A woman dressed in gray shows off an **invisible anti-theft and degaussed** fanny pack.

VideoLLaVA: A woman in a gray coat **hangs** a brown fanny pack around her waist, then holds it in both hands to **show the inner structure** of the fanny pack.

SVAD-VLM: A woman in a **gray coat** stands and **shows off a fanny bag**, then **unzips it with both hands** and **takes cell phone and the perfume inside out**. She then **put the items back** and **fasten the buckle** of the bag.

(b)

Figure 8: Qualitative results from the SVAD dataset, comparing our proposed SVAD-VLM with the BLIP2 [5], Qwen-v1 [1], LLaVA(v1.5) [9], VideoChat [6], GVT [2] and Video-LLaVA [8].

Original Video	Text	Top1	Top2	Top3	Top4	Top5
	General Caption: A woman pours red liquid into a glass on a table.					
	Fine-grained Description (Ours): A woman in a yellow top pours red liquid from a clear glass into another glass.					

(a)

Original Video	Text	Top1	Top2	Top3	Top4	Top5
	General Caption: A video game screen showing a person playing a game.					
	Fine-grained Description (Ours): In a video game, a character dressed in red uses Spider-Man's skills to fly to a tall building.					

(b)

Figure 9: Qualitative results general caption and fine-grained description in text-to-video retrieval.