

TAVGBench: Benchmarking Text to Audible-Video Generation Supplementary Materials

Anonymous Authors

ABSTRACT

In the supplementary materials, we provide additional details about the dataset annotation pipeline, the proposed AVHScore, and the baseline methodology. Moreover, we include enhanced visualizations and a demo video to further illustrate our findings.

1 MORE DETAILS

1.1 More details of dataset annotation pipeline

We show the prompts used at different stages of our dataset annotation pipeline.

Prompt of Audio Re-captioning: *This is an audio captioning, please rewrite in more details, do not add personal feeling/opinion, and avoid excessive text. If details are missing, please add details in one sentence directly without asking a question in reply.*

Prompt of Video Re-captioning: *This is a video captioning, please rewrite in more details, do not add personal feeling/opinion, and avoid excessive text. If details are missing, please add details directly without asking a question in reply.*

Prompt of Fused Re-captioning: *The following contents include one video captioning and one audio captioning, please smoothly fuse them and give one complete paragraph. Do not add personal feeling/opinion, and avoid excessive text.*

Finally, we analyze the descriptions across the entire dataset and generate a wordcloud, as illustrated in Fig. 1.



Figure 1: Wordcloud of the proposed TAVGBench.

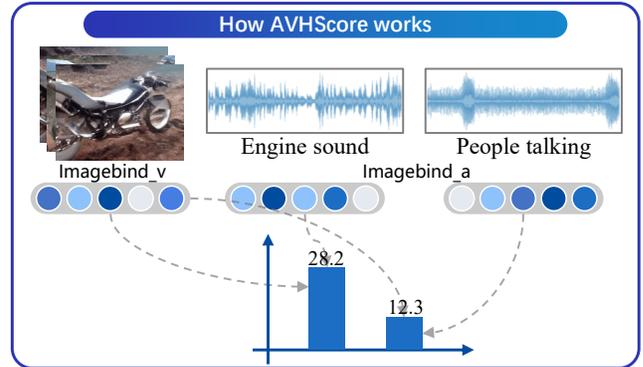


Figure 2: We detail the AVHScore calculation process with two specific examples. The first video features a motorcycle paired with its corresponding audio, “Engine sound”. After extracting their features via ImageBind, we calculate the score using cosine similarity. For contrast, we provided a negative example, “People talking”. Following the same process, this pairing yielded a significantly lower score.

1.2 More details of the proposed AVHScore

Our proposed Audio-Visual Harmony Score (AVHScore) is designed to assess the alignment between generated audio and video. For well-aligned pairs, this metric yields higher values. Conversely, poorly aligned pairs result in lower values. In Fig. 2, we show how AVHScore works through a qualitative example. We further examine the effectiveness of AVHScore in assessing audio-visual alignment. We input unpaired audio and video, calculate their AVHScore, and present the results in Table 1.

Table 1: Performance of AVHScore in evaluating audio-visual alignment using unpaired audio and video inputs.

	paired	unpaired
AVHScore	23.35	3.18

2 LIMITATIONS OF THE TAVGBENCH

The TAVGBench dataset we proposed comprises 1.7 million entries sourced from AudioSet [2], each containing 10 seconds of synchronized audio and video. However, the videos are not precisely split according to specific events, resulting in numerous transitions within the data. Additionally, the dataset includes videos with subtle movements. These characteristics may impact the performance of the TAVG model. In the future, we aim to develop more precise video split methods [1] and create higher-quality datasets that maintain consistent scene composition.

3 MORE EXAMPLES

We provide additional visualizations to illustrate our results. We show the results of more categories of video generation in the demo video. Including music scenes, landscapes, daily scenes, animation scenes, game scenes, *etc.* The details are shown in `demo.mp4`.

REFERENCES

- [1] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. 2024. Panda-70M: Captioning 70M Videos with Multiple Cross-Modality Teachers. *arXiv preprint arXiv:2402.19479* (2024).
- [2] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 776–780.

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
174

175
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