Supplementary Material for
Zero-Shot Audio-Visual Editing via Cross-Modal Delta Denoising

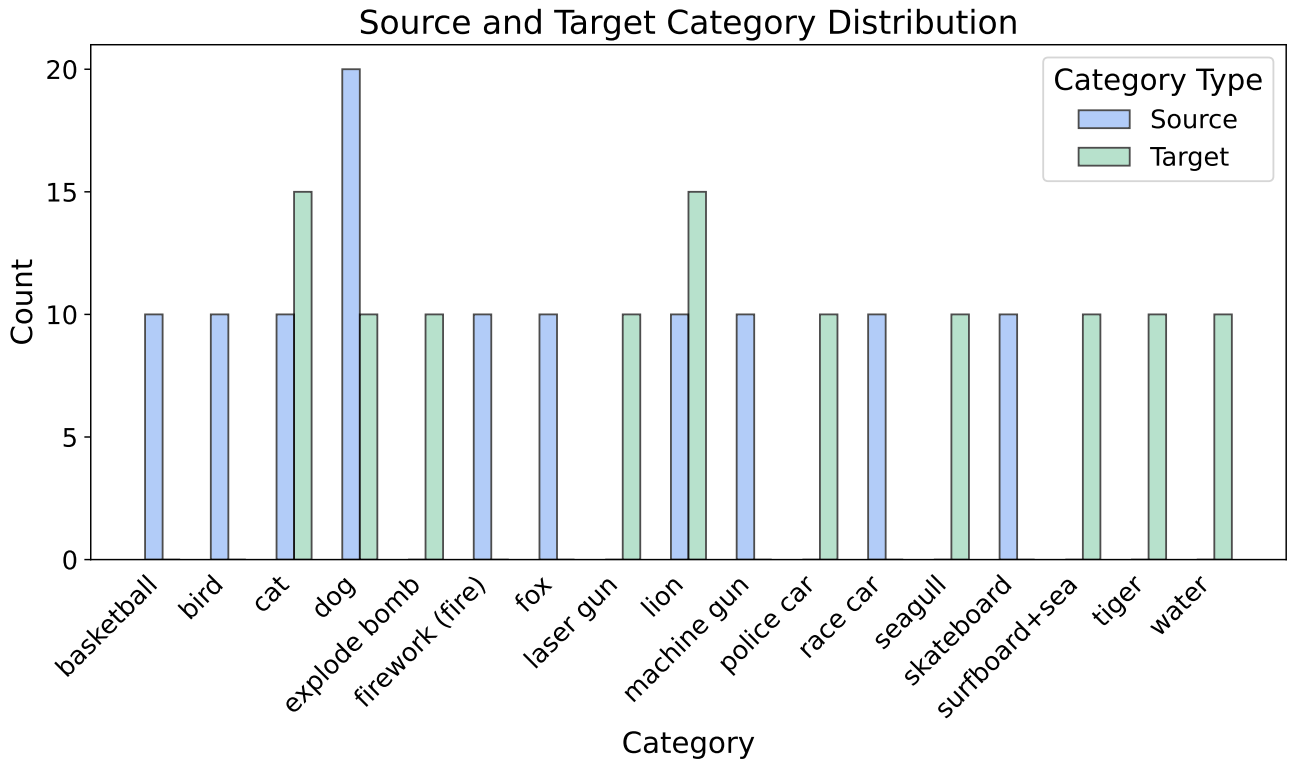


Figure 1. **Category Distribution of AVED-Bench.** We present the source and target category distribution of the AVED-Bench dataset. The source categories represent the initial categories, while the target categories indicate their edited categories. This distribution highlights AVED-Bench 's capability to effectively evaluate a variety of audio-video editing.

Our supplementary material consists of:

1. Details of AVED-Bench.
2. Implementation Details
3. Human Evaluation Details
4. Additional Quantitative Results.

1. Details of AVED-Bench

Category Distribution. In Figure 1, we demonstrate the source and target category distributions in the AVED-Bench dataset to provide a comprehensive overview of its diverse and balanced composition. The source categories represent the initial events or objects, while the target categories indicate their corresponding editing events or objects. AVED-Bench includes a wide variety of events from animal sounds (e.g., *dog*, *cat*, *bird*) to mechanical noises (e.g., *machine gun*, *race car*) and environmental effects (e.g., *firework*, *water*). All categories are well-balanced to ensure that no single category dominates the dataset, which is essential for effective zero-shot audio-video evaluation.

Mapping of Source and Target Categories. In Figure 2, We present a heatmap visualizing the count of mappings between source and target categories in the AVED-Bench dataset. This provides an intuitive understanding of the re-

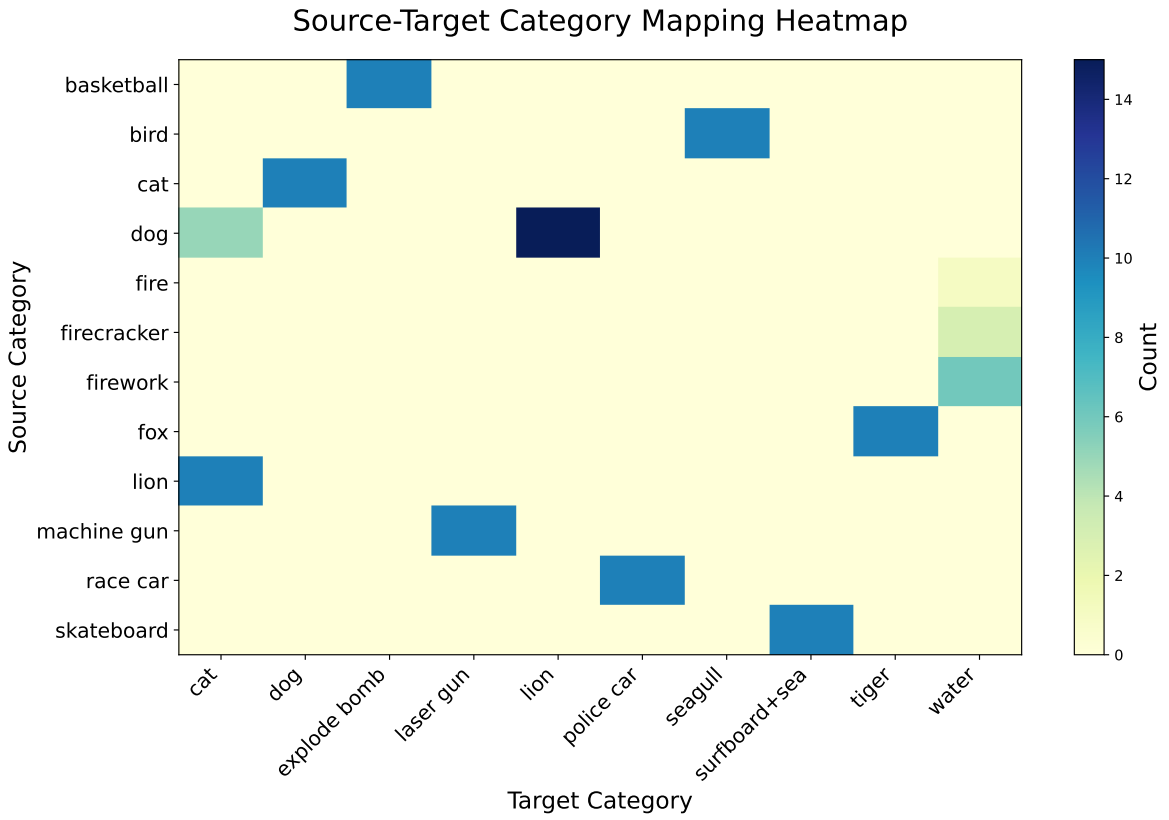


Figure 2. **Mapping of Source and Target Categories.** This figure summarizes the count of mappings between source and target categories in the dataset. Each cell represents the frequency of a specific source-to-target mapping, providing an intuitive overview of the relationships and transitions present in AVED-Bench.

lationships and transitions from source to target prompts. Each cell in the heatmap represents the frequency of a specific source-to-target mapping, with darker shades indicating higher counts. We note that the mappings include transformations such as *dog* to *lion* and *firework* to *water*, reflecting both logical relationships and imaginative diversity in these pairings. These logical and imaginative pairs can fairly and robustly evaluate the effectiveness of audio-video editing tasks.

Full Dataset. In Table 2, we provide full annotation of AVED-Bench for the reference. We note that the target category is used for **OBJ.** metrics.

2. Implementation Details

We use pretrained Stable Diffusion 2.1 [5] and AudioLDM2-Large [3] as the backbone for video and audio processing, respectively. Following the setup of RAVE [2], we structure a 10-second video (at 4 fps) into a 2×2 grid. At each DDS iteration, the latent frames within each grid are randomly shuffled across different grids. The optimization process consists of 200 steps in total. We optimize the first 15 steps using only the DDS loss to ensure that the target latent is initially related to the desired editing prompt. In the remaining steps, we introduce the cross-modal denoising loss $\mathcal{L}_{\text{cmd}}$ by a factor of 10 for both audio and video. We adjust the DDS scaling for different phases: for video, the scale is set to 2000 for the first 15 steps and then to 4000 for the remainder. For audio, it is set to 1000 initially and increases to 5000 after that. The target latent $\mathbf{z}(\theta)$ is updated using the SGD optimizer with a learning rate of 1, decaying by multiplying 0.99 at each iteration. We set the threshold both τ_a and τ_v to 0.8. Positive patches are sampled randomly, taking 50% of the patches where $\tilde{\mathbf{S}}^a_{\text{trg}} > \tau$ or $\tilde{\mathbf{S}}^v_{\text{trg}} > \tau$ for audio and video, respectively. For negative sampling, we randomly select 80% of patches where $\tilde{\mathbf{S}}^a_{\text{src}} < \tau$ or $\tilde{\mathbf{S}}^v_{\text{src}} < \tau$, from both source and target branches for $\mathbf{H}^-_a$ and $\mathbf{H}^-_v$. If the number of audio-video patches differs, we randomly drop selected patches to align them. This entire process takes approximately 20 minutes on a NVIDIA A6000 GPU.

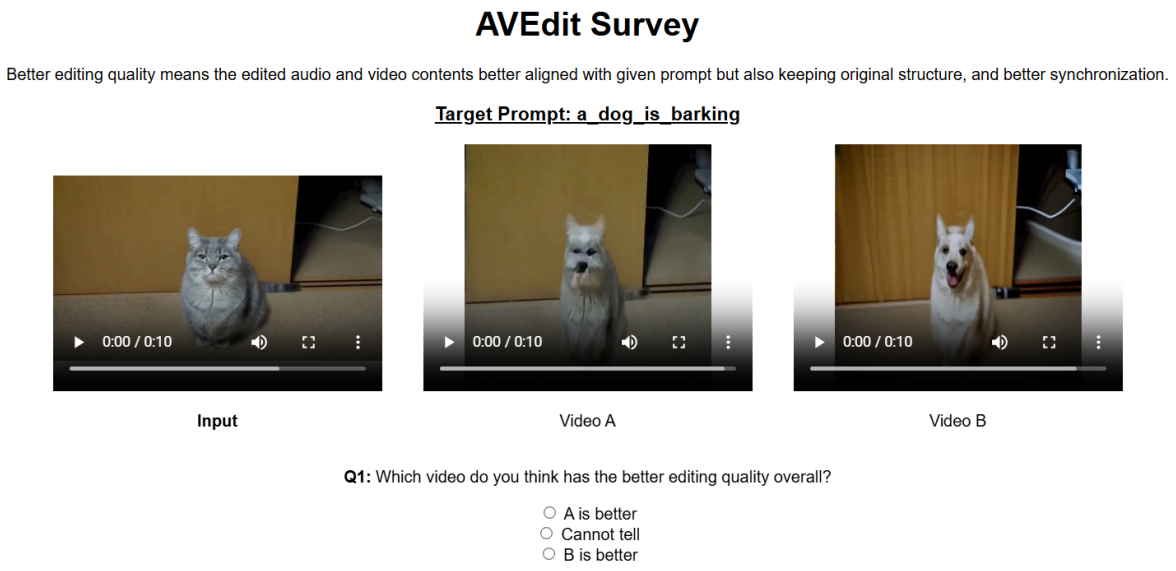


Figure 3. **Human Evaluation.** Human raters are asked to select the edited video that best aligns with the target prompt. We report the average human preference rate for each method. Note that all samples are presented in a random order.

3. Human Evaluation Details

As depicted in Figure 3, we conduct a human evaluation to assess the quality of edited audio-video samples based on their alignment with the target prompt. Participants are presented with a source (unedited) video and two edited versions generated by different methods (one must come from AVED). They are asked to select their preferred sample based on *Which video do you think has the better editing quality overall?* For each question, participants can choose one of the two samples or a third option, "Cannot tell." Each subject evaluates five randomly selected video pairs from a pool of 110 comparisons, ensuring a diverse sample set. One sample in each pair is always from AVED, while the other is from a competing method [1, 2, 6]. To prevent bias, all methods remain anonymized during evaluation. Our study involves 300 participants recruited via Amazon Mechanical Turk. Results are reported as the average human preference rate for each method, providing insights into the perceived quality of audio-video edits.

4. Additional Quantitative Results

In Figure 4, we present detailed quantitative results evaluating the performance of AVED across different thresholds (i.e., τ_v and τ_a in the main draft). We report the metrics DINO, LPAPS, and AV-Align, which are highly related to how synchronized edited audio and video are. For simplicity, we set τ_v and τ_a **equally** in these experiments. These results highlight the impact of different threshold settings on each metric.

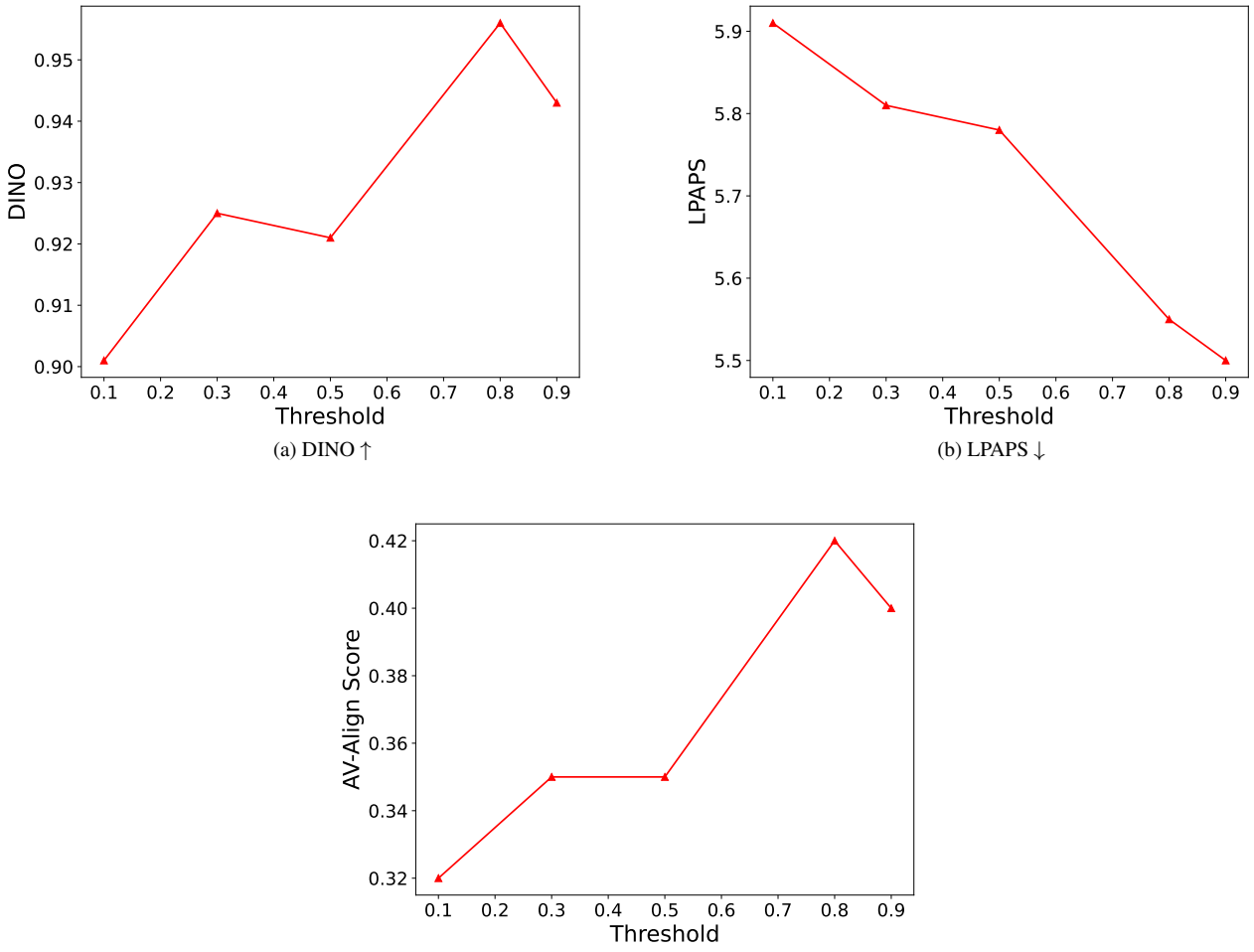
DINO and LPAPS. In Figure 4a andFigure 4b, these metrics evaluate structural similarity and coherence in visual outputs and perceptual similarity in audio, respectively. The results demonstrate that the score achieves peaks (close to peak) around **0.8** to suggest the optimal hyper-parameters contributing to aligned audio-video editing.

AV-Align Results. In Figure 4c, the suggested threshold, **0.8**, also presents the best results in the AV-Align metric to lead the synchronization and coherence between audio and video editing results.

Different Settings for τ_v and τ_a. The best performance is achieved with $\tau_v = 0.8$ and $\tau_a = 0.7$, yielding the following results: CLIP-F (**0.905**↑), CLIP-T (**0.260**↑), Obj. (**0.180**↑), DINO (**0.961**↑), CLAP (**0.229**↑), LPAPS (**5.41**↓), IB (**0.24**↑), and AV-Align (**0.48**↑). These results demonstrate the benefits of separately tuning audio and video thresholds to improve overall performance.

Grid Design Ablation. In Table 1, we study different sizes of the grids. We note that larger grids slightly enhance video temporal consistency (CLIP-F) and visual structure preservation (DINO), while a smaller grid (e.g., 2×2) yields better visual and audio fidelity (CLIP-T, Obj, CLAP, LPAPS) and synchronization (IB, AV-Align).

Additional Alignment Metric. We use the ACC metric [4], which predicts the probability of synchronization, for additional reference. In AVED-Bench, the ACC



(a) DINO ↑ (b) LPAPS ↓

(c) AV-Align ↑

Figure 4. **Impact of the Threshold.** The sub-figures illustrate the performance of DINO, LPAPS, and AV-Align metrics on AVED-Bench across varying threshold settings, where the threshold decides whether a patch is a prompt-relevant patch (i.e., τ_v and τ_a in the main draft).

Grid	CLIP-F↑	CLIP-T↑	Obj.↑	DINO↑	CLAP↑	LPAPS↓	IB.↑	Align.↑
2×2	0.903	**0.260**	**0.180**	0.956	**0.226**	**5.55**	**0.23**	**0.42**
3×3	0.910	0.229	0.157	0.960	0.214	5.71	0.21	0.40
4×4	**0.915**	0.221	0.150	**0.961**	0.211	5.65	0.21	0.40

Table 1. **Grid Design.** Performance comparison across different grid sizes.

results↑ show that ControlVideo achieves 52.7%, Token-Flow reaches 45.4%, and RAVE obtains 55.4%. In comparison, AVED significantly outperforms these methods with an ACC of **72.7%**, highlighting AVED's effectiveness. This substantial improvement demonstrates a similar trend of AV-align in the main draft, which ensures better synchronization and alignment of edited content.

References

[1] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. In *ICLR*, 2023. 3

[2] Ozgur Kara, Bariscan Kurtkaya, Hidir Yesiltepe, James M Rehg, and Pinar Yanardag. Rave: Randomized noise shuffling for fast and consistent video editing with diffusion models. In *CVPR*, 2024. 2, 3

[3] Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *TASLP*, 2024. 2

[4] Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. In *NeurIPS*, 2023. 3

[5] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2

[6] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. In *ICLR*, 2024. 3

Table 2. Full Dataset Overview

YouTube ID and Time	Source (top) and Target (bottom) Prompt	Source Category	Target Category
MkJuGDph-1k_000070	a dog is barking a lion is roaring	dog	lion
2jIv5qBTS88_000009	a dog is barking a lion is roaring	dog	lion
3LO8iqmX_LM_000030	a dog is barking a lion is roaring	dog	lion
5nAvo9b2_SY_000020	a dog is barking a lion is roaring	dog	lion
21vfIdV2B60_000030	a dog is barking a lion is roaring	dog	lion
92vVMZ4Zkcc_000000	a dog is barking a lion is roaring	dog	lion
AyRHwZyYxEs_000003	a dog is barking a lion is roaring	dog	lion
BCcdatk468s_000070	a dog is barking a lion is roaring	dog	lion
CM5W31FRa7o_000030	a dog is barking a lion is roaring	dog	lion
cRt6_axWZqY_000020	a dog is barking a lion is roaring	dog	lion
3d4Q9iU5Je8_000006	a cat is meowing a dog is barking	cat	dog
3gjF-cVb_m4_000030	a cat is meowing a dog is barking	cat	dog
3S8bTR5VAp0_000030	a cat is meowing a dog is barking	cat	dog
04_1N_55VFw_000030	a cat is meowing a dog is barking	cat	dog
4le_42J5qkU_000012	a cat is meowing a dog is barking	cat	dog
5b04Vp0loTM_000030	a cat is meowing a dog is barking	cat	dog
5pncSb7fxEQ_000030	a cat is meowing a dog is barking	cat	dog
6kdiESHOO3w_000030	a cat is meowing a dog is barking	cat	dog
6VpcBh1KPSM_000001	a cat is meowing a dog is barking	cat	dog
7cIVRdCLnW4_000064	a cat is meowing a dog is barking	cat	dog
JQ39VPSbQ2o_000000	a race car is driving on the road a police car is driving on the road	race car	police car
M-bmsXbJBdw_000034	a race car is driving on the street a police car is driving on the street	race car	police car
kJe373Z1qsc_000050	a race car is driving down the street a police car is driving down the street	race car	police car
KW_OFhShBfI_000040	a race car is driving down the road a police car is driving down the road	race car	police car

Continued on next page

Table 2. Full Dataset Overview

YouTube ID and Time	Source (top) and Target (bottom) Prompt	Source Category	Target Category
LGdZx_Ng2OU_000070	a race car is driving a police car is driving	race car	police car
LSMgG64MW2U_000130	a race car is driving down the street a police car is driving down the street	race car	police car
lVAPg1giCr8_000010	a race car is driving down the street a police car is driving down the street	race car	police car
m6iAg3VoVFo_000050	a race car is driving down the road a police car is driving down the road	race car	police car
MmIIjAzgw30_000180	a race car is driving down the road a police car is driving down the road	race car	police car
o0CMD5nm-48_000380	a race car is driving down the road a police car is driving down the road	race car	police car
2wUsyQY0R5Q_000360	a boy is riding a skateboard on a ramp a boy is riding a surfboard on the sea	skateboard	surfboard+sea
4yJFeZd3oYg_000030	a man is riding a skateboard on a ramp a man is riding a surfboard on the sea	skateboard	surfboard+sea
22b6SLUvGHk_000020	a man is riding a skateboard on a ramp a man is riding a surfboard on the sea	skateboard	surfboard+sea
38DktuXxeqQ_000260	a man is riding a skateboard on a ramp a man is riding a surfboard on the sea	skateboard	surfboard+sea
38Vg0ciQWBc_000010	a man is riding a skateboard on a ramp a man is riding a surfboard on the sea	skateboard	surfboard+sea
81FaBaq0t-8_000225	a man is riding a skateboard on a ramp a man is riding a surfboard on the sea	skateboard	surfboard+sea
1914bVcPq0k_000030	a man is riding a skateboard on a ramp a man is riding a surfboard on the sea	skateboard	surfboard+sea
Apr3yTPMZCY_000458	a man is riding a skateboard on a ramp a man is riding a surfboard on the sea	skateboard	surfboard+sea
ctJ2410j7qo_000040	a man is riding a skateboard on a ramp a man is riding a surfboard on the sea	skateboard	surfboard+sea
FIO-QGbT1-g_000000	a man is riding a skateboard on a ramp a man is riding a surfboard on the sea	skateboard	surfboard+sea
_-R5EPaybpk_000002	a lion is roaring a cat is meowing	lion	cat
_Xha6m87OOs_000060	a lion is roaring a cat is meowing	lion	cat
1-9ryaEsFc8_000029	a lion is roaring a cat is meowing	lion	cat
2PSepowyWHE_000003	a lion is roaring a cat is meowing	lion	cat
5ZtB9FnUIZs_000016	a lion is roaring a cat is meowing	lion	cat
6YS3ewiOBkU_000007	a lion is roaring a cat is meowing	lion	cat
9IIKUunyzM8_000007	a lion is roaring a cat is meowing	lion	cat
9S4-bVB6v0Q_000005	a lion is roaring a cat is meowing	lion	cat

Continued on next page

Table 2. Full Dataset Overview

YouTube ID and Time	Source (top) and Target (bottom) Prompt	Source Category	Target Category
bBMcsO6IeDE_000021	a lion is roaring a cat is meowing	lion	cat
C0i90WmUMvE_000109	a lion is roaring a cat is meowing	lion	cat
7IVwirRF3W0_000030	machine gun laser gun	machine gun	laser gun
8LVWNGRjD0g_000252	machine gun laser gun	machine gun	laser gun
35zwOkXkDt4_000224	machine gun laser gun	machine gun	laser gun
-ahgvCmiECM_000030	machine gun laser gun	machine gun	laser gun
AwW31u6wYvE_000019	machine gun laser gun	machine gun	laser gun
bnqsQhk2yX0_000040	machine gun laser gun	machine gun	laser gun
C9n__zk2DY8_000061	machine gun laser gun	machine gun	laser gun
kqfvQWAH0C0_000130	machine gun laser gun	machine gun	laser gun
LIUowNjRNwE_000125	machine gun laser gun	machine gun	laser gun
N59msUnyy1g_000054	machine gun laser gun	machine gun	laser gun
1nzZrX5JWyA_000017	a dog is howling a lion is roaring	dog	lion
6CmqiGDOtmo_000014	a dog is howling a lion is roaring	dog	lion
6WIPUATvzL4_000001	a dog is howling a lion is roaring	dog	lion
BFBW5bEqCy0_000018	a dog is howling a lion is roaring	dog	lion
d0V8UNE3Dtg_000141	a dog is howling a lion is roaring	dog	lion
ESRiXGhb-Ww_000095	a dog is howling a cat is meowing	dog	cat
fIramHEoZvg_000017	a dog is howling a cat is meowing	dog	cat
fq2SbG-7Rtk_000063	a dog is howling a cat is meowing	dog	cat
j3NuRPLUDhg_000021	a dog is howling a cat is meowing	dog	cat
J4pwTPDqrKA_000018	a dog is howling a cat is meowing	dog	cat
_wQakQ-AC3A_000020	a bird is chirping a seagull is chirping	bird	seagull
0C-5reSgR2w_000020	a bird is chirping a seagull is chirping	bird	seagull

Continued on next page

Table 2. Full Dataset Overview

YouTube ID and Time	Source (top) and Target (bottom) Prompt	Source Category	Target Category
0yTG1Yrnmak_000030	a bird is chirping a seagull is chirping	bird	seagull
1QWd2fVos8s_000104	a bird is chirping a seagull is chirping	bird	seagull
3Wt_ldHN6QE_000113	a bird is chirping a seagull is chirping	bird	seagull
4Ruk56DiQj8_000030	a bird is chirping a seagull is chirping	bird	seagull
4SUPgmqKe1s_000053	a bird is chirping a seagull is chirping	bird	seagull
7WL9Orh7auA_000324	a bird is chirping a seagull is chirping	bird	seagull
9bNTnSx8eFY_000240	a bird is chirping a seagull is chirping	bird	seagull
997RTKzc39c_000146	a bird is chirping a seagull is chirping	bird	seagull
0uOCCKQ3jQk_000019	a fox is barking a tiger is roaring	fox	tiger
2Zir1UxVpxo_000005	a fox is barking a tiger is roaring	fox	tiger
7h-RkocdbiM_000016	a fox is barking a tiger is roaring	fox	tiger
dBgiqeXB7PI_000038	a fox is barking a tiger is roaring	fox	tiger
eH6SkkHQ118_000105	a fox is barking a tiger is roaring	fox	tiger
GQHfBnaGXu8_000036	a fox is barking a tiger is roaring	fox	tiger
jU3mklfml-E_000001	a fox is barking a tiger is roaring	fox	tiger
McWY8wmi1NE_000045	a fox is barking a tiger is roaring	fox	tiger
qfSpVBLpmg8_000140	a fox is barking a tiger is roaring	fox	tiger
qiWvQK-siSA_000087	a fox is barking a tiger is roaring	fox	tiger
0Cvljw2Sssk_000000	a man is playing basketball a man is playing an exploding bomb	basketball	explode bomb
6Jo_tvf0qdI_000009	a man is playing basketball a man is playing an exploding bomb	basketball	explode bomb
CWFnfL-5X-M_000030	a young girl is playing basketball a young girl is playing an exploding bomb	basketball	explode bomb
Do__idgHWJM_000490	a man is playing basketball a man is playing an exploding bomb	basketball	explode bomb
do8fVVvSvTQ_000130	a young man is playing basketball a young man is playing an exploding bomb	basketball	explode bomb
GQJ3-6ZBsfg_000200	two men are playing basketball two men are playing an exploding bomb	basketball	explode bomb

Continued on next page

Table 2. Full Dataset Overview

YouTube ID and Time	Source (top) and Target (bottom) Prompt	Source Category	Target Category
I5ECnXmzfCo_000135	a young boy is playing basketball a young boy is playing an exploding bomb	basketball	explode bomb
KEeyw8lcPKs_000000	a man is playing basketball a man is playing an exploding bomb	basketball	explode bomb
myBRFJP7Z2U_000001	a man is playing basketball a man is playing an exploding bomb	basketball	explode bomb
SCskHem1qzo_000060	a man is playing basketball a man is playing an exploding bomb	basketball	explode bomb
0aKYFdbqjHc_000032	a firework lit up in the night sky a splash of water lit up in the night sky.	firework	water
1W-i57MQ-zQ_000590	a firework lit up in the night sky a splash of water lit up in the night sky.	firework	water
2pxTI_FzoqI_000289	a firework lit up in the night sky a splash of water lit up in the night sky.	firework	water
8LfRQTJAemw_000322	a firework lit up in the night sky a splash of water lit up in the night sky.	firework	water
41SLncI_B04_000066	a firework lit up in the night sky a splash of water lit up in the night sky.	firework	water
a8fa79w2aIQ_000023	a bursting firecracker is on the road a splash of water is on the road	firecracker	water
dyi5OukAmmU_000016	a bursting firecracker is on the street a splash of water is on the street	firecracker	water
e6at9oqeIGw_000086	a bursting firecracker is on the snow a splash of water is on the snow	firecracker	water
ehnGYrg3Vaw_000252	a fire is exploding in the field water is splashing in the field	fire	water
XD1_PR3n7xY_000323	a firework lit up in the night sky a splash of water lit up in the night sky.	firework	water