AUDIO-JOURNEY: OPEN DOMAIN LATENT DIFFUSION BASED TEXT-TO-AUDIO GENERATION (APPENDIX)

Jackson Michaels[†△], *Juncheng B Li*[⋆△], *Laura Yao*[⋆], *Lijun Yu*[⋆], *Zach Wood-Doughty*[†], *Florian Metze*[⋆]

Appendix Outline:

1. Section A will cover more details about prompt engineering, and the usage of entailment scoring.
2. Section B will cover more details about our model training and hyperparameters
3. Section C offers a further in-depth understanding of our **cross-attention mechanism** as described in Equation (4) in the main paper, and why it produced the results in Table 2 in the main paper.
4. Section D includes more generation results from our diffusion model.
5. Section E describe the broader impact and limitation of our work.

A. MORE ON PROMPT ENGINEERING:

A.1. Prompting given audio-only weak labels:

Table A1 shows the few-shot examples we feed Alpaca model with to generate audio captions based on the labels given.

Prompt	**Response**
alarm, burp, inside, small room.	burping while an alarm plays inside a small room.
dog, bark, howl, speech.	a dog barking and howling with a person speaking as well.
Music, jazz, piano, singing, speaking.	a person plays jazz piano with a singer while people talk.
engine, vehicle, wind, music, speech.	people talking inside a car while driving and listening to music.
water, gargle, inside, small room.	air is passing through the water in their mouth in a small room with water.
scratch, hammer, metal.	hammer striking a metal surface and scratching sounds can be heard.
thunder, wind, bark, small room.	a dog is barking in a small room during a thunderstorm with audible wind.
gunshot, vehicle engine, siren, crash.	a car chase with gunfire and sirens where a vehicle crashes.
waterfall, wind, sizzle, crackle.	a fire is cracking with something sizzling near a waterfall with wind.
stream, cough, cat, Purr.	a cat purrs near a coughing person while a stream can be heard.

Table A1. Few-shot examples of Audio-only labels to captions. The list of audio labels is preceded by the prompt: *"For each of these, summarize the sounds into a single sentence: \n describe a situation with all of these sounds together:"*

† Northwestern University, Chicago, IL
⋆ Carnegie Mellon University, Pittsburgh, PA
△ These authors contributed equally to this work.

A.2. Filter Hallucination and Obtain Visual Captions:

We noticed that many of the single-label audio captions often had hallucinations where there would be extra details added (usually from one of the examples given). One example of this would be *"A person is sprinting while a dog is barking and howling."* when only the *"run"* label was given.

To address this, we created visual captions to help enrich our audio data. Table A2 shows the specific examples we used to join together the three visual captions that were generated using BLIP2 [10].

Prompt	**Response**
a video game with a dragon in the water final fantasy xv - the end of an era final fantasy x-2 - the end of an era.	The end of an era awaits in Final Fantasy XV and X-2.
a person is using a printer to print out a document a white cash register with a keyboard and a keypad a person is holding a small piece of plastic on a plane.	a person using a piece of plastic to make a purchase at a cash register.
a man with a beard sitting on a bed a man in a cowboy hat is playing an acoustic guitar a man standing in the desert holding an acoustic guitar.	A man in a cowboy hat is playing a guitar and strumming away in the desert.

Table A2. Few-shot merging visual caption examples. Each set of captions is preceded by this prompt: *"Create one sentence that summarizes these three simply:"*

A.3. Merging Audio and Visual Captions:

Table A3 shows the specific examples that were used in the prompt to join our single-label audio class and summarized video captions. These examples help us filter the hallucination generated when creating captions from audio-only weak labels by utilizing visual information.

A.4. Audio-visual Entailment Score:

These scores can be used to address the audio-visual false positive issue in which the audio of the video does not match the visual frames presented. These scores also increase human understanding of image-audio-text correlations. We can further utilize these scores to help perform caption filtering, gating based on the entailment score between the audio and visual concepts demonstrated within each video.

A.4.1. Entailment scoring with Alpaca, T5, and BERT:

We conducted entailment scoring using Alpaca[21], T5[17], and BERT[3] on the balanced set of AudioSet[4]. This allowed us to compare the results of the three methods and how they would consider entailment differently. Since Alpaca[21] is a decoder-only model, we noticed that the entailment scoring was not always reliable, scoring differently on additional runs. This is due to its auto-regressive nature, which is highly dependent on its context. To stabilize the scoring, we employ few-shot examples (in Table A4, Table A5) to better utilize the in-context learning ability of Alpaca. When conducting entailment scoring for audios that only had one label, we utilized the AudioSet ontology description [4] and label due to the hallucination in the default audio captions.

Despite using these examples, we still noticed that Alpaca would sometimes output "incorrect" scores, e.g. given the following pairing: (*"A background of traditional Indian music with lyrics from a bhangra song playing in the foreground."*, *"A bright star is twinkling in the night sky, shining amidst the dark velvety backdrop."*) a score of 0.9 or (*"A person dribbling a*

Prompt	Response
man in a cowboy hat is throwing a rope while standing on a green field. Whip	An audio scene emphasizing the sharp sound of a whip, set against a visual backdrop of a man in a cowboy hat throwing a whip on a green field.
a man eating a cupcake at a table Tick	An audio scene centered around the subtle ticking sound, with a backdrop of a man enjoying a cupcake at a table.
a person using a piece of plastic to make a purchase. Cash register	An audio scene capturing the distinctive sound of a cash register, accompanying the moment when a person uses a piece of plastic to make a purchase.
a man brushing his teeth Toothbrush	a man cleans his teeth with a toothbrush's audible scrubbing.
man standing in the desert holding an acoustic guitar Country	An audio scene focused on the Country genre, featuring a man standing in the desert holding an acoustic guitar.

Table A3. Few-shot audio and visual caption merge examples. Caption-label pair is preceded by the following prompt: *"Summarize these two captions conditioned on the second caption, the second caption describes an audio class and is the main concept:"*

basketball, slamming it on the ground, and speaking.", *"A man in a purple shirt is playing basketball, a man in a red shirt is playing soccer."*) a score of 0.1. These cases are very counter-intuitive for human to explain.

On the other hand, we noticed that encoder-decoder models (we only use the encoder) such as T5 [17] and BERT[3] scoring tended to score some single-label (ontology-based) captions lower than an alpaca-generated caption that had less which audio-visual correspondence. For example, when the caption pairing was (*"An ice cream truck outside a small room, playing music."*, *"An old digital audio box sits proudly on a table, displaying its unique blue and white design."*), T5 gave it a score of 0.81 whereas when the pairing was (*"Fire : Sounds resulting from the rapid oxidation of a material in the exothermic chemical process of combustion, releasing heat, light, and various reaction products."*, *"Firefighters are putting out a blazing fire in a building."*), T5 only gave it a score of 0.80 despite the two captions being much closer in meaning.

Although there were some discrepancies when analyzing these scores, we still found that overall, T5 scoring seemed to be more consistent with the actual content of the captions. The different score distributions for the three metrics utilized on the balanced set can be found here (Figure 1(a), Figure 1(b), Figure 1(c)).

These distributions reinforce how T5 had a generally higher score range whereas BERT had scores in a lower range (even below 0) as compared to Alpaca. It also shows how T5 and BERT had similarly shaped distributions whereas Alpaca tended to spread a bit more evenly across the spectrum (with peaks at different points in the score distribution). The result of the 3 different types of scoring on the balanced set can be found on our open-source page[1]. We have also utilized T5 scoring on the unbalanced set for AudioSet which will also be made available.

A.4.2. Effect of Audio-visual Entailment Scores

To test the effectiveness of our entailment score, we employ a similar audio-visual classification pipeline as described in [9], which is illustrated in Figure 2. Our audio encdoer is the same AST/DeiT model[5] used in the main paper. The video is encoded by a pre-trained R2+1D[22] model. The naive fusion is pure concatenation. We plug in our entailment scores in the attention mechanism by discounting the attention score for the video portion. Specifically, $\mathbf{c}_t = \sum_{i=1}^{T} \alpha_{t,i}\mathbf{h}_i$ the corresponding embedding indexes' attention score gets discounted by the entailment score. The result is shown in Table A6. Although our

[1] https://audiojourney.github.io/

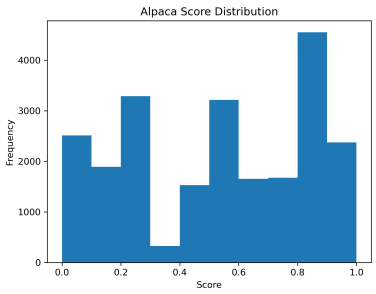
Prompt	Response
a man in a cowboy hat is throwing a rope while standing on a green field. Whip : The sound of whipping, i.e., the greatly accelerated motion of the tip of a flexible structure, as the result of concentrated angular momentum.	0.85
a man eating a cupcake at a table Tick : A metallic tapping sound.: A metallic tapping sound.	0.00
a person using a piece of plastic to make a purchase. Cash register : Sounds of a mechanical or electronic device for registering and calculating transactions, usually attached to a drawer for storing cash.	0.45
a man brushing his teeth Toothbrush : Sound of an instrument used to clean the teeth and gums consisting of a head of tightly clustered bristles mounted on a handle.	1.00
man standing in the desert holding an acoustic guitar Country : A genre of United States popular music with origins in folk, Blues and Western music, often consisting of ballads and dance tunes with generally simple forms and harmonies accompanied by mostly string instruments such as banjos, electric and acoustic guitars, dobros, and fiddles as well as harmonicas.	0.90

Table A4. Alpaca audio-visual entailment score examples for single-label. The caption-label pair is preceded by the following prompt: *"on a scale from 0 to 1, output the probability that the first caption describes a scenario with the second caption's sound description:"*

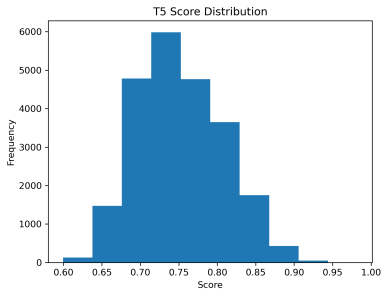
best score still lags the best SoTA model on AudioSet, the point is our improvement over naive fusion shows all three types of entailment scores outperform the naive version of fusion.

A.5. More Details about Caption Generation:

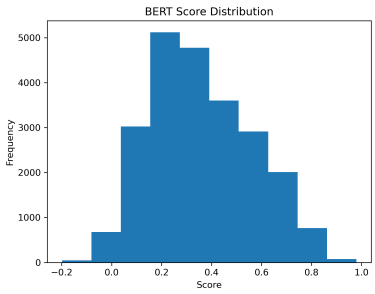
Similarity Score (Table 2 of main paper) We recruited 5 human subjects to evaluate 500 samples and report their average ratings for zero-shot, one-shot, and few-shot and A-V Merge captions and WavCaps [13] generated results in Table 2 of the main paper. When evaluating our captions, we ask the human subject to provide a similarity score between the generated captions and the AudioCaps[7] ground truth, ranging from 0-1. In the cases of ambiguous samples or when AudioCaps[7] labels are not reliable, we also provide the original youtube links to the evaluators and ask them to use the *audio content* as the ground truth.
Key Statistics Comparison: To further compare the different caption sets, we analyzed the vocabularies of AudioCaps [7] versus the WavCaps [13] and captions our system generated from the same clips. By observing the top 20 words (Figure 5, Figure 3, Figure 4), we can see that AudioCaps and WavCaps have many similarities with 14 of the top 20 words being identical.



(a) Alpaca-generated entailment score distribution for balanced set



(b) T5-generated entailment score distribution for balanced set



(c) BERT-generated entailment score distribution for balanced set

Fig. 1. Comparison of entailment score distributions for different models on AudioSet[4] balanced set.

Prompt	Response
two young men wearing red and white shirts A person is speaking while there is a loud gush of air.	0.10
goat grazing on grass in the mountains A goat making music and someone speaking aswell.	0.75
a young boy is riding a skateboard down a sidewalk A male is singing and a child is singing along to the same music.	0.05
a person is holding a gun with a glove on it Gunfire and cap gun.	0.80
a screenshot of a game with three people in red and blue outfits Gunfire and cap gun.	0.00
a person playing a ukulele with their hands There is background music with a mandolin being played.	0.70
a man in a green shirt is talking to the camera A cat meowing and a person speaking.	0.50
a person holding a snake in front of a door A snake hissing.	0.60

Table A5. Alpaca audio-visual entailment score examples for multi-label. The caption pair is preceded by the following prompt: *"on a scale from 0 to 1, output the probability that the first caption describes a scenario with the second caption's sound description:"*

Model	Backbone	PT	AS-20k (mAP)			AS-2M (mAP)		
			A	V	A+V	A	V	A+V
Naive Fusion	DeiT-B/R2+1D	IN+KI-SL	$34.6_{\pm.20}$	$18.1_{\pm.09}$	$37.4_{\pm.18}$	$45.4_{\pm.70}$	$23.9_{\pm.12}$	$46.5_{\pm.29}$
Alpaca score Fusion	DeiT-B/R2+1D	IN+KI-SL	$34.6_{\pm.20}$	$18.1_{\pm.09}$	$38.4_{\pm.14}$	$45.4_{\pm.70}$	$23.9_{\pm.12}$	$47.6_{\pm.33}$
T5 score Fusion	DeiT-B/R2+1D	IN+KI-SL	$34.6_{\pm.20}$	$18.1_{\pm.09}$	$\mathbf{39.1}_{\pm.10}$	$45.4_{\pm.70}$	$23.9_{\pm.12}$	$\mathbf{49.5}_{\pm.42}$
BERT score Fusion	DeiT-B/R2+1D	IN+KI-SL	$34.6_{\pm.20}$	$18.1_{\pm.09}$	$38.7_{\pm.15}$	$45.4_{\pm.70}$	$23.9_{\pm.12}$	$49.1_{\pm.37}$
AST [5]	DeiT-B	IN	34.6	-	-	45.4	-	-
MBT [14]	ViT-B	IN-SL	31.3	27.7	43.9	41.5	31.3	49.6
CAV-MAE [6]	ViT-B	SSL	37.7	19.8	42.0	46.6	26.2	51.2

Table A6. **Comparison with other state-of-the-art models** on audio-visual classification evaluated on AudioSet[4] test set, using both audio and visual features. Metrics are mAP for AS. For pre-training (PT) dataset, AS:AudioSet, KI:Kinetics (for R2+1D[22]), and IN:ImageNet. SSL: self-supervised learning, SL: supervised learning; We gray-out baselines. Best single models in AS-2M are compared (no ensembles).

These similarities help support the closeness of vocabulary between the two sets (resulting in high automatic metrics), meanwhile reducing the diversity and generalizability of the captions. We also noticed that in the top 20 words of our audio-video merged captions contain the term 'audio scene' which reflects the examples we used when prompting Alpaca. This shows a clear path that we could inject our own inductive bias into any future LLM-based dataset augmentations. In addition to the vocabulary size difference, we also noticed a key difference in the length of captions generated through our system versus those of AudioCaps or WavCaps. The minimum caption length for all three main types of captions were all 3. The maximum caption differed a lot, with the AudioCaps [7] samples having a maximum of 31, WavCaps [13] samples having a maximum of 25, and our audio-video merged captions having a maximum 45. Additionally, the average length of AudioCaps is 10.49, for WavCaps is 6.89, and for our audio-video merge was 17.07. This demonstrates a drastic increase in length and thus richness of the captions generated by our LLM-based audio-video merging methods. We see that the WavCaps [13] captions are generally the shortest which suggests

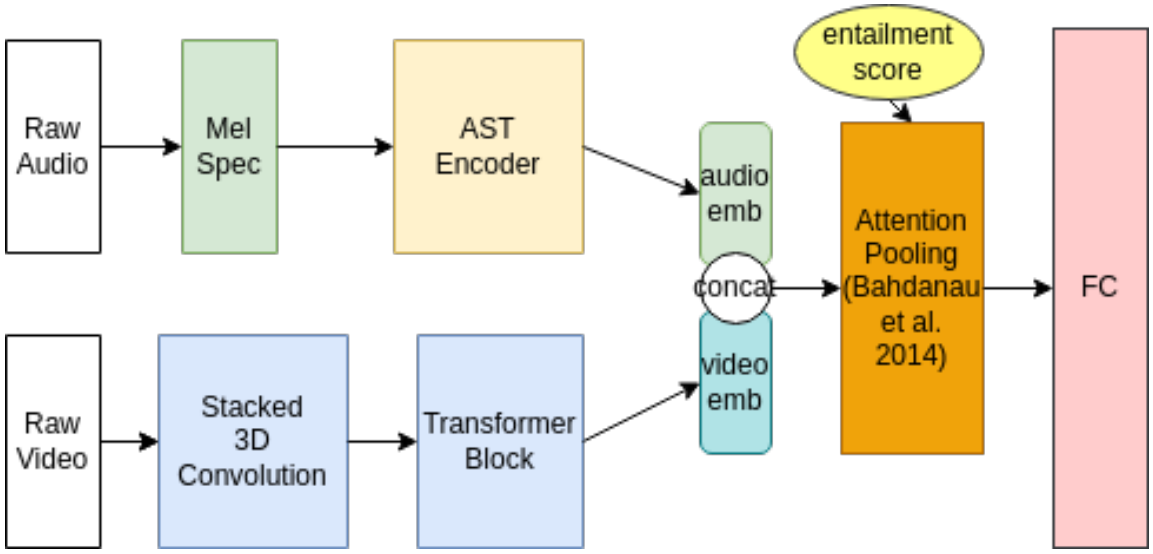


Fig. 2. Entailment score penalizing attention score at the attention pooling layer

a lack of variation in structure when generating captions.

Some key examples of where utilizing audio-video merged captions benefited our model over WavCaps[13] are included in Table A7. We see that in these examples, our captions utilized details such as the flag or the caption on the video to determine what exactly the sound was, adding key details that weren't apparent in the WavCaps captions. These also demonstrate how simple WavCaps captions are overall.

Overall, we believe this is a very promising paradigm to augment the existing audio dataset. In the future, if we could leverage high-performance classifiers to auto-label audio from the wild or to fine-tune the LLM to plug into Automatic Audio Captioning system, it would be a clear pathway to scale up audio training datasets.

Youtube ID	**Audio Label**	**Video Caption**	**Merged Caption**	**WavCaps[13]**
BjWf0keANT8	Sine Wave	10,000 hertz sine wave audio frequency.	10,000 hertz audio frequency, represented by a sine wave.	a continuous sine wave sound
83IJft_3Z4E	Throbbing	An ultrasound image of a baby in the womb, proof of new life and potential new beginnings.	An audio scene depicting the throbbing sound of a heartbeat, accompanying the visual of an ultrasound image of a baby in the womb, proof of new life and potential new beginnings.	a heartbeat is being recorded
6hj2F5xvGYE	Male singing	A man is singing into a microphone in front of an American flag.	A male singing into a microphone in front of an American flag, capturing the emotion of the patriotic song.	someone is singing a song

Table A7. Examples of our Audio-Video merged captions versus WavCaps[13]

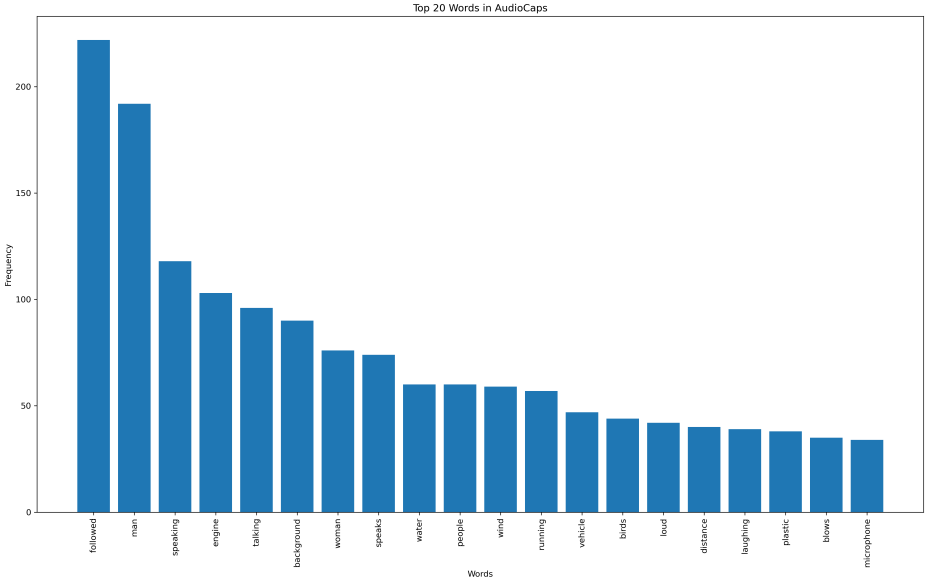


Fig. 3. Top 20 words from the vocabulary of the AudioCaps samples (not including stop words)

B. TRAINING DETAILS

B.1. Implementation

We fine-tuned our models from models available on HuggingFace Diffuser Library, specifically their v1.4 model [18]. While we initialized our model from the Stable Diffusion checkpoint, the model, training code, and larger pipeline have been heavily modified to fit our purposes. Most notably, we had to make changes to work with non-HuggingFace models, such as the Encodec model[2], along with changes to the U-Net to adapt to wide-channel inputs. All data for models other than ours in Table B.1 was copied from their respective papers training details as they do not provide *training* code as of the time of writing, only AudioLDM[11] has public code for loading checkpoints.

For our training, we exclusively trained on individual machines with eight A100 GPUs. We chose most hyperparameters following [18], with variations to fit our hardware and model. We could use significantly larger batch sizes due to the extremely high compression from the Encodec [2] model's discrete codebook; we lowered our memory overhead by 56.5% per sample. Pre-computing the codebook codes made these gains possible by allowing larger batches and faster training. As Table 2 of the main paper described, we trained multiple models, varying the text encoder with all other parameters and stages remaining the same. However, as will be described in Section C, the T5 models required an additional linear projection layer from the length 1028 T5 encoding to the expected 786 input dimension for cross-attention. This extra layer adds parameters to the model and would affect training, but the added parameters are negligible compared to the size of the U-Net. Due to our A100 GPUs are the 40GB version, without model parallel, we could not afford to unfreeze and fine-tune the text encoder together with training the UNet.

B.2. Encodec Latent Embedding Space

Figure 6 displays the Encodec [2] pipeline through the quantization stage. Perceptually the reconstruction is excellent with a subjective evaluation (MUSHRA score) of 88.0[2]. This figure does not show the entire end-to-end process raw audio to generated raw-audio, as we do not use them in training our model. We initially trained our model on the intermediate discrete stage (RVQ codebook). However, this proved difficult due to the discrete nature of this latent space. While we could easily have trained the model on the pre-quantization latent, this would not have allowed us to leverage the compressed quantized representations as this would have required reading wav files from disk. We trained our final model on the post-quantization latent representation displayed at the bottom.

[2]https://en.wikipedia.org/wiki/MUSHRA

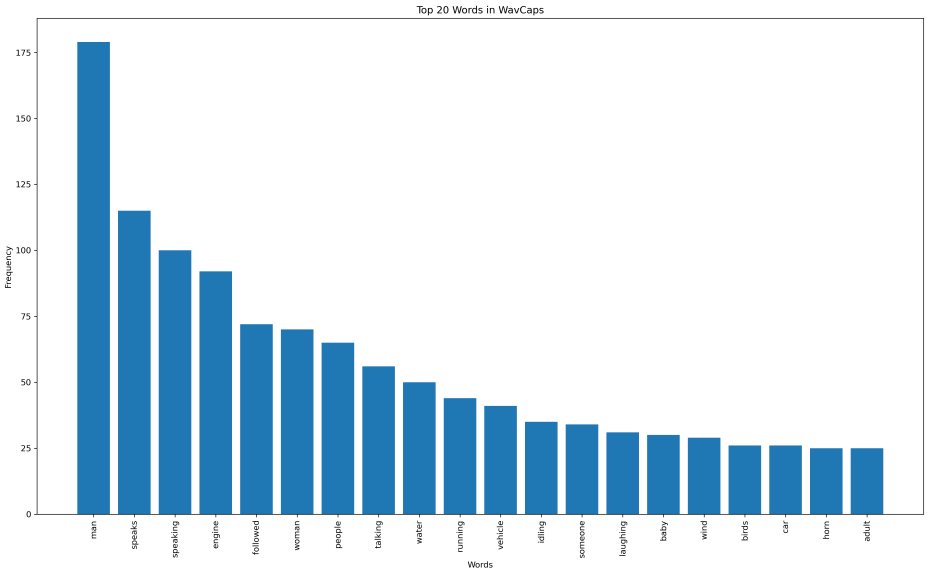


Fig. 4. Top 20 words from the vocabulary of the WavCaps samples (not including stop words)

	Diffusion (Denoising Network)				Classification	
Configuration	DiffSound[24]	AudioGen[8]	AudioLDM[11]	Ours	AS-20K	AS-2M
Optimizer	AdamW	Adam	-	AdamW	AdamW	AdamW
Optimizer β_1 - β_2	0.9 - 0.94	-	-	0.9 - 0.999	0.9 - 0.999	
Base learning rate	3.0e-6	5.0e-4	1.0e-4	2.56e-4	0.001	2e-4
LR schedule	Constant	Inv Sqrt	Constant	CosDecay	CosDecay	CosDecay
Noise schedule	Sc-Linear	-	Sc-Linear	Cosine	-	-
ChannelMultiplier	-	1,2,4,8	1,2,3,5	1,2,4,4	-	-
Diffusion Steps	-	-	1K	1K	-	-
Warm-up epochs	-	-	-	-	1	4
Training Epochs	600	-	-	-	60	10
Warm-up steps	-	3K	-	1K	1K	1K
Training Steps	-	200K	1.5M	40K	-	-
Batch size	16	256	8	192	256	32
GPUs	32	128	1	8	4	4
GPU Type	V100	A100	A100	A100	V100	V100
SpecAug [16]	-	-	-	-	192/48	192/48
Mixup [25]	-	-	-	-	0.5	0.5
Loss Function	-	ℓ_1,ℓ_2,CE	MSE	MSE	BCE	BCE
Sampler	DDIM[20]	-	DDIM	PNDM[12]	-	-
Sample Steps	100	-	200	45	-	-
Guidance Scale	-	1 - 5	4.5	3.5 - 7.5	-	-
Normalization	-	-	-	Channel	(-4.27, 4.57)	(-4.27, 4.57)

Table A8. Table comparing training hyperparameters between SOTA audio generation models, our model, and our classification models. All "-" values are either unknown or not applicable to the given model. All values are from respective papers and appendices sections on training. Inv Sqrt = Inversed Square root; Sc-Linear = Scaled Linear; CosDecay = Cosine with Decay. For normalization we include a "-" if the values are unknown, channel for our per-channel normalization, or (μ, σ) for the dataset.

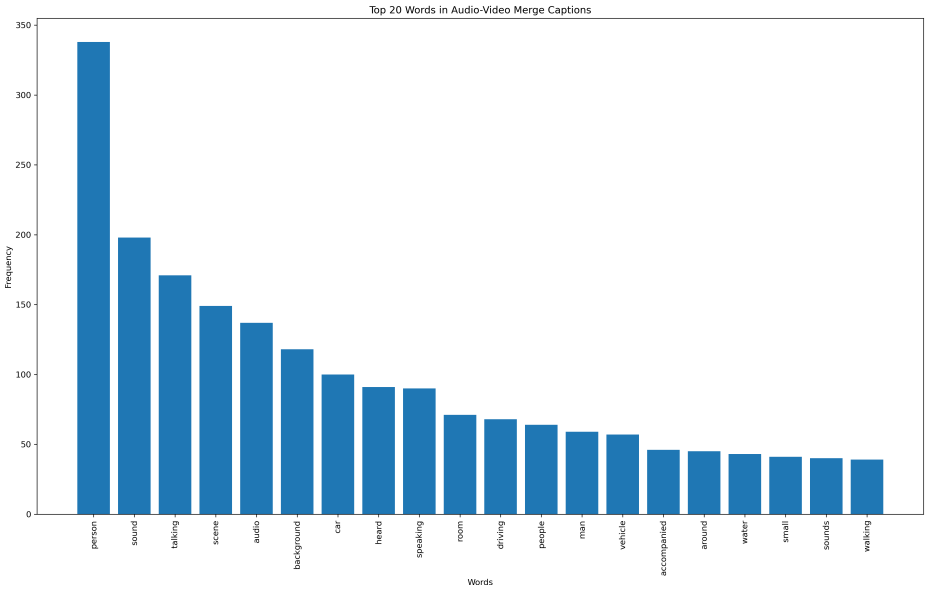


Fig. 5. Top 20 words from the vocabulary of the our audio-video merged samples (not including stop words)

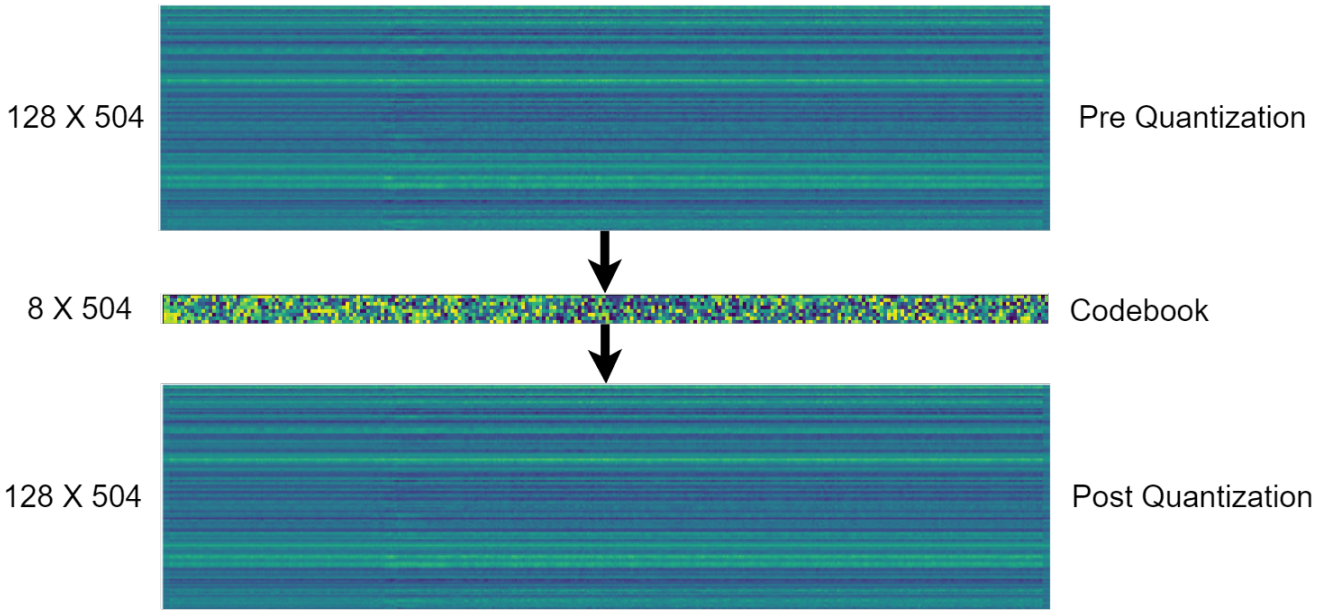


Fig. 6. Encodec [2] quantization process for encoder latent representations. Encoder-decoder pair not included in figure. Codebook stage is cropped in the figure to improve visibility.

B.3. Training Instability from Frozen Blocks

One warning from the [18] paper for fine-tuning their models is catastrophic forgetting. While we faced this issue when training our CLAP models, it did not noticeably affect the training of our T5 models. Our U-Net consists of a few major components: a conv_in block, down blocks, a mid-block, up blocks, and a conv_out block. During experimentation, we attempted to accelerate training and conserve the pre-trained weights of *Stable Diffusion-1-4* by freezing the weights of all blocks other than conv_in and conv_out for 5,000 training steps. This method proved inferior to simply allowing the entire pipeline to learn together in final loss value and training stability. Figure 8 shows an example training loss graph for one of the aforementioned experiments,

clearly displaying the high degree of instability that emerged after the blocks were unfrozen. This instability alone would not be a reason to abandon this technique; however, the more troubling trend was the loss values plateauing around 0.4 compared to 0.19 in our best models.

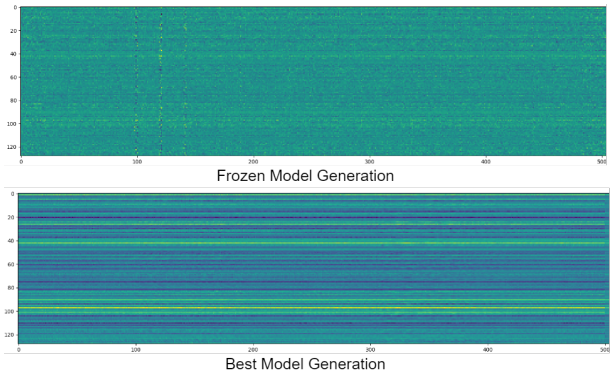


Fig. 7. Generation comparison between model trained with frozen blocks (top) and model trained without (bottom). Comparing this to other examples, such as Figure 6 clearly shows the lack of quality from frozen models.

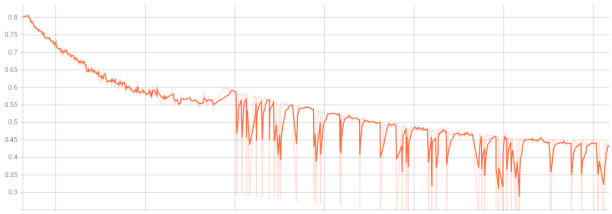


Fig. 8. Training loss graph for model. This model started with all layers other than conv_in and conv_out frozen, then unfroze these blocks after 5,000 training steps.

C. CROSS ATTENTION MECHANISM

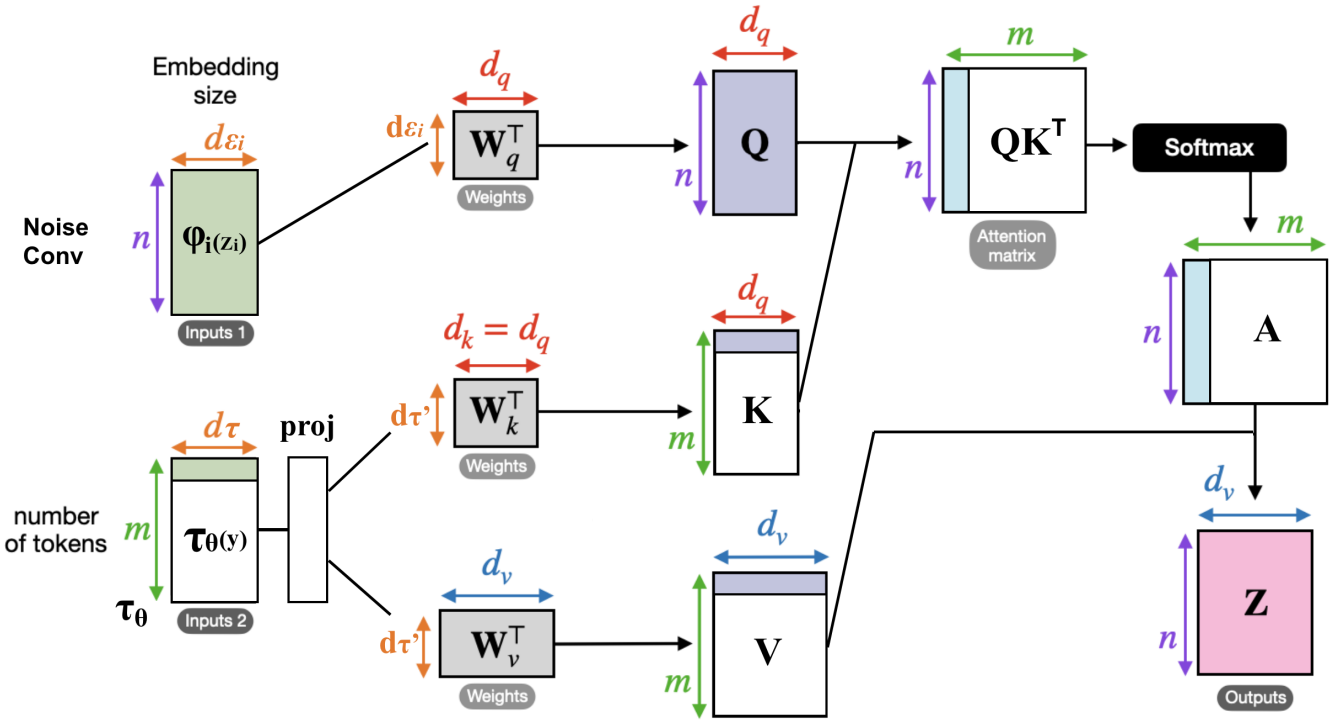


Fig. 9. Illustration of the cross-attention mechanism under masking scenario, the white-colored portions indicate masking.

In Table 2 of our main paper, we observed our *AudioJourney-CLAP* models generally under perform *AudioJourney-T5* models, we believe there could be 2 main reasons:
1) CLAP[23] was trained on 660k samples, which is way smaller dataset than what T5[17] was trained on. Although the CLAP

text encoder demonstrated good performance on audio embedding [23], T5 may be superior in a general setting.
2) Since CLAP [23] output matches with d, we did not adapt its last layer. Using a frozen encoder would solely depend on the W_q, W_k, W_v to learn the mapping, which might be suboptimal.

$$\text{Cross-Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\Big(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}}\Big) \cdot \mathbf{V}$$

$$\text{where } \mathbf{Q} = \mathbf{W}_Q^{(i)} \cdot \varphi_i(\mathbf{z}_i),\ \mathbf{K} = \mathbf{W}_K^{(i)} \cdot \tau_\theta(y),\ \mathbf{V} = \mathbf{W}_V^{(i)} \cdot \tau_\theta(y)$$

$$\text{and } \mathbf{W}_Q^{(i)} \in \mathbb{R}^{d_q \times d_\epsilon^i},\ \mathbf{W}_K^{(i)} \in \mathbb{R}^{d_k \times d_{\tau'}}, \mathbf{W}_V^{(i)} \in \mathbb{R}^{d_v \times d_{\tau'}},\ \varphi_i(\mathbf{z}_i) \in \mathbb{R}^{n \times d_\epsilon^i},\ \tau_\theta(y) \in \mathbb{R}^{m \times d_\tau}$$

in our case: $d_\tau = 1024, d_{\tau'} = 768, d_k = d_v = d_q = d = 768$(initialized from Stable Diffusion [18] weights),

d_ϵ^i is the i^{th}layer of Unet φ_i's output size

(main paper equation (4))

A surprising result in Table 2 of our main paper is that the masked model performed worse than the unmasked model. This is counter-intuitive, given that attention masking is a common mechanism used to handle variable length inputs. Figure 9 explains how masking negatively affects cross-attention in the U-Net. As is illustrated, the white part of the text embedding $\tau_\theta(y)$ indicates the masked out content because the max sequence length is larger than the number of audio caption tokens. Towards the end, you can see if we pass this mask to the U-Net, this would result in a low-rank dot-product of the attention score A and the V tensor, which result in a low-ranked Z. In an extreme case of when token length is 1, this is reduced to a rank-1 vector dot product of column by row. We believe this low-ranked representation of Z is suboptimal to the full-ranked version when unmasked.

Therefore, to compensate for the above issue, we reduce our max token length to 50, and use the unmasked versions in our best-performing recipe.

D. ADDITIONAL SAMPLES AND DEMOS

For overall listening experience, we put our listening samples and spectrogram visualizations to our anonymous website: `https://audiojourney.github.io/` and the code and implementations are at: `https://github.com/audiojourney/audiojourney.github.io`

Our Audio-journey models could serve as the base model similar to Stabel Diffusion [18] in vision, and it would allow a separate smaller network to learn new concepts from new dataset (ControlNet) [26], and would also allow finetuning through low-rank approximation like Dreambooth [19].

E. IMPLICATIONS OF THIS WORK

E.1. Broader Impact

Our work demonstrates the potential of using LLMs and diffusion as a data augmentation pipeline, by showing that classifiers can obtain improved performance when trained on diffusion-generated data. The competitive performance of classifiers trained on data generated by the diffusion model validated the efficacy of diffusion models as valuable tools for enhancing the performance of audio classifiers, thereby fueling more future effort to augment audio datasets. This work could have broader impacts on many domains where auditory information is crucial but datasets are limited. Almost all advances in machine learning bring both opportunities and risks. A privacy concern of more powerful audio classifiers is that may enable privacy violations, in particular when the audio data in question is a person's voice [1, 15]. We hope that future work on privacy-preserving methods can also make use of our diffusion-generated audio data.

E.2. Limitations

Our methodology relies heavily on the use of LLM outputs, and such output will inevitably contain "hallucinations." However, our overall framework accounts for and alleviates such hallucinations by grounding our caption generation with visual context. "Hallucination" is an inevitable property of LLMs, but our approach already greatly alleviated this. Future work could explore fine-tuning our LLMs to discourage excessive hallucinations.

Extracting reliable embeddings that faithfully represent the underlying data from a decoder-only LLM such as Alpaca [21] can require a meticulous and strategic approach. As a result, some insights derived from our embeddings may not always align with the actual data samples. Regarding the utilization of AudioSet [4] for training, there is the potential drawback of inconsistent data quality. The presence of low-quality labels or consistently noisy audio samples could potentially introduce nontrivial noise

into the model's learning, thereby affecting the quality of the generated outputs. While AudioSet [4] is an essential source of data, we hope that our work and future research can continue to improve it and other datasets, enabling a wide range of methods and applications in the audio domain. This isn't to discount the immense value that AudioSet [4] provides as a comprehensive sound library, but rather to highlight the importance of data cleansing and refinement for optimal model performance.

F. REFERENCES

[1] Ranya Aloufi, Hamed Haddadi, and David Boyle. Privacy-preserving voice analysis via disentangled representations. In *Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop*, pages 1–14, 2020.

[2] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[4] 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.

[5] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021.

[6] Yuan Gong, Andrew Rouditchenko, Alexander H Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James Glass. Contrastive audio-visual masked autoencoder. *arXiv preprint arXiv:2210.07839*, 2022.

[7] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *NAACL-HLT*, 2019.

[8] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*, 2022.

[9] Juncheng B Li, Shuhui Qu, Xinjian Li, Po-Yao Bernie Huang, and Florian Metze. On adversarial robustness of large-scale audio visual learning. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 231–235. IEEE, 2022.

[10] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023.

[11] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D. Plumbley. Audioldm: Text-to-audio generation with latent diffusion models, 2023.

[12] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*, 2022.

[13] Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *arXiv preprint arXiv:2303.17395*, 2023.

[14] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34:14200–14213, 2021.

[15] Patrick O'Reilly, Andreas Bugler, Keshav Bhandari, Max Morrison, and Bryan Pardo. Voiceblock: Privacy through real-time adversarial attacks with audio-to-audio models. In *Advances in Neural Information Processing Systems*, 2022.

[16] Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin Dogus Cubuk, and Quoc V. Le. Specaugment: A simple augmentation method for automatic speech recognition. In *INTERSPEECH*, 2019.

[17] 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. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[18] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[19] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022.

[20] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[21] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`, 2023.

[22] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[23] Yusong Wu*, Ke Chen*, Tianyu Zhang*, Yuchen Hui*, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, 2023.

[24] Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[25] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[26] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.