MLV Lab Diffusion Study

Align, Adapt and Inject: Sound-guided Unified Image Generation

Presenter: Jiwon Jeong

jjwon4086@korea.ac.kr

Paper

Align, Adapt and Inject: Sound-guided Unified Image Generation

Yue Yang[1, 2] Kaipeng Zhang[2]✉ Yuying Ge[3] Wenqi Shao[2]
Zeyue Xue[3] Yu Qiao[2] Ping Luo[2, 3]✉

[1] Shanghai Jiao Tong University [2]Shanghai AI Laboratory [3]The University of Hong Kong
yang-yue@stju.edu.cn {zhangkaipeng, shaowenqi, qiaoyu}@pjlab.org.cn
yuyingge@hku.hk xuezeyue@connect.hku.hk pluo@cs.hku.hk

Abstract

- Text-guided image generation -> Sound-guided?
  - Sound is vital element with the sphere of human perception
  - Scarcity of datasets
- Unified framework ‘Align, Adapt and Inject’ (AAI)
  - Adapts input sound into a sound token like ordinary word
- Outperforms other text and sound-guided state-of-the-arts methods

Introduction

- Text-to-Image (T2I) diffusion models
  - Only vision and text modalities:

    discrete form of text -> challenging to convey distinctions and vivid properties
  - Sound (continuous form) -> vivid signals
    - ex) different expressions when lion starts roaring and finishes
- Hard to implement Audio-to-image model like T2I model
  - Impractical to train a large-scale sound-guided generative model
  - Audio-image dataset is small-scale -> Hard to integrate audio into T2l model

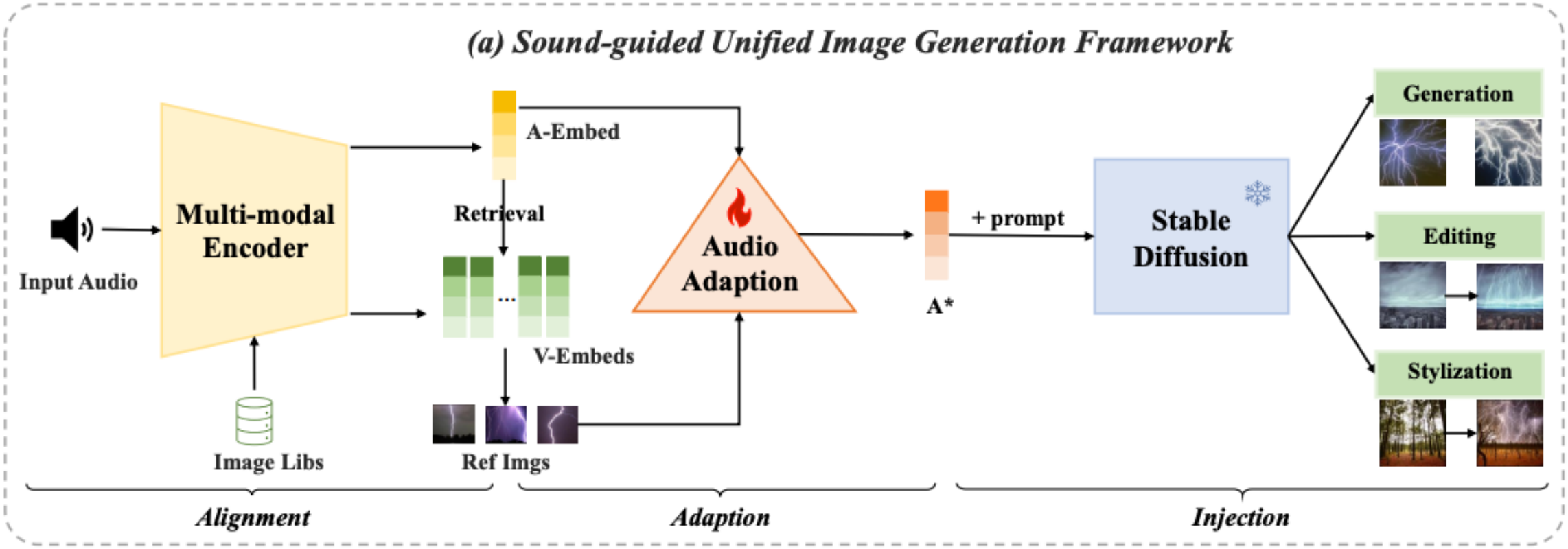
Introduction

- Recent approaches
  - Require training the whole model -> high computation cost
  - Drastic quality degradation when new audio comes
- Unified strategy: Align, Adapt and Inject (AAI)
  - Align audio representations with visual and textual modality
  - Adapt each audio representation into an audio token
  - Inject the audio token into T2I models

Introduction

Figure 1: Examples from our unified strategy AAI. A^* is the audio token for each audio. Our method provides various capabilities based on sound inputs. The user can use the input sound for image generation (left), image editing (middle), and image stylization (right) flexibly.

Introduction

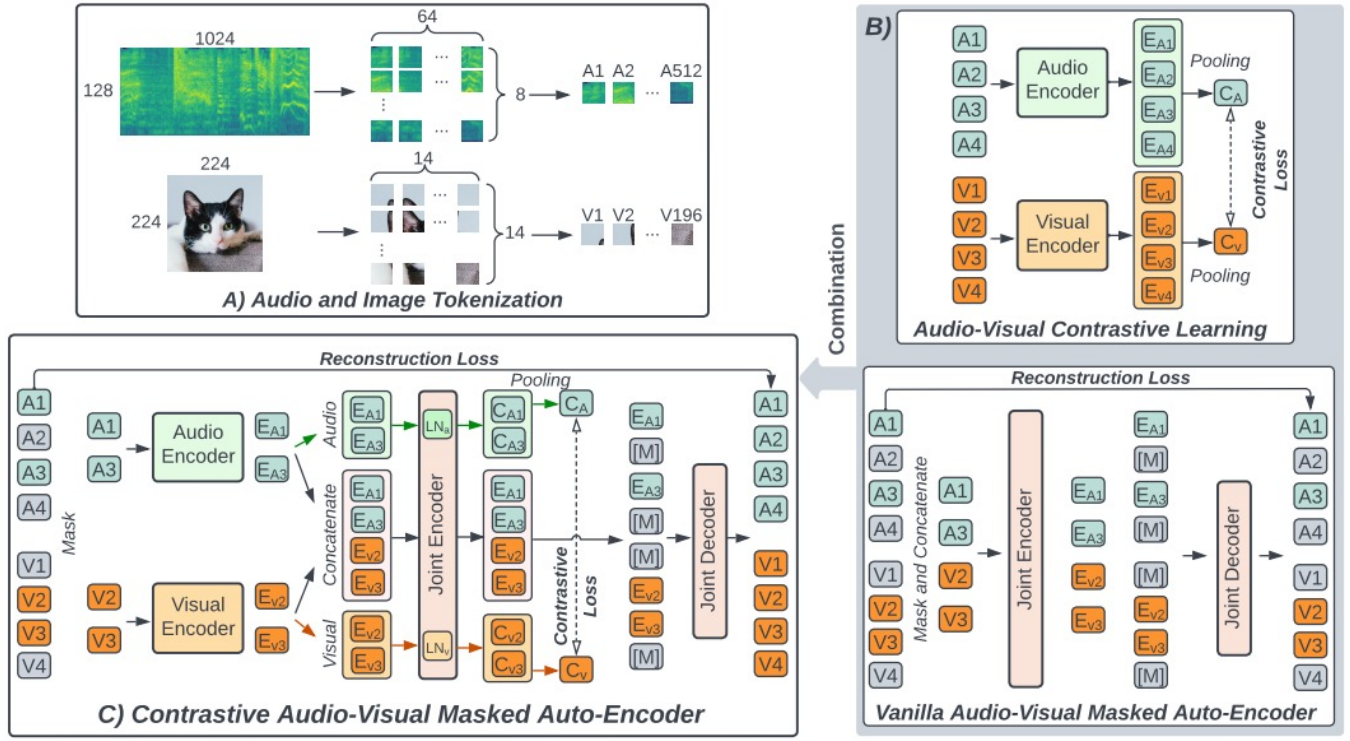


Method

- Use audio as a clue to guide image generation
- Propose a sound-guided image generation framework (AAI)
  - Align: Multi-modal alignment
  - Adapt: Audio-representation adaption for T2Is
  - Inject: Audio-representation injection into T2Is

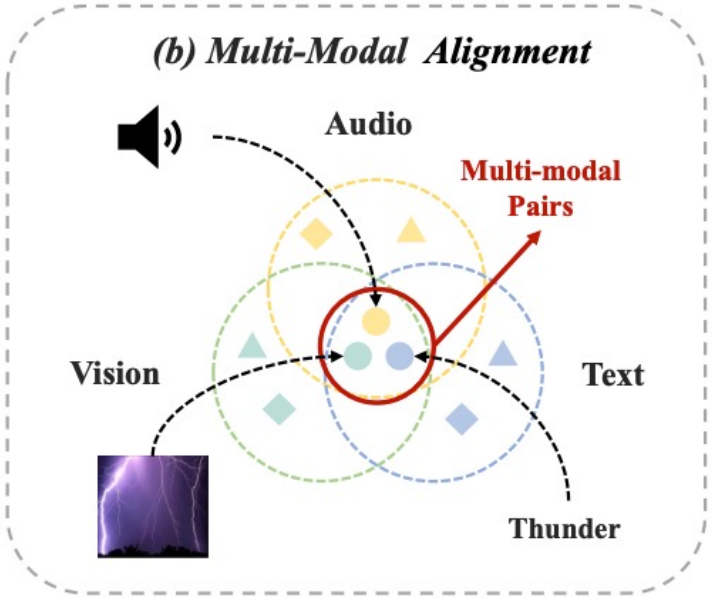
Multi-modal Alignment

- Contrastive Audio-Visual Masked Autoencoder (CAV-MAE)



Multi-modal Alignment

- Learn audio representation aligned with the paired visual and textual content
  - audio, vision encoder -> CAV-MAE (transformer)
  - text encoder -> Diffusion
- Text encoder does not align with the image encoder
  - encoded label embedding works well in our CAVT-MAE
  - CAVT-MAE: fine-tune CAV-MAE



Multi-modal Alignment

- CAVT-MAE
  - from CAV-MAE
  - freeze visual and textual encoder weights and only tune audio encoder
  - fine-tuned by the alignment between audio and text, audio and vision

$$\mathcal{L} = \mathcal{L}_c(A, T) + (1 - \alpha)\mathcal{L}_c(A, V) + \frac{\alpha}{2}\mathcal{L}_t(A, V)$$

Multi-modal Alignment

- CAVT-MAE
  - fine-tuned by the alignment between audio and text, audio and vision
  - total loss:

$$\mathcal{L} = \mathcal{L}_c(A, T) + (1 - \alpha)\mathcal{L}_c(A, V) + \frac{\alpha}{2}\mathcal{L}_t(A, V)$$

    - A, T, V denotes modalities for audio, text, vision
    - $L_c(,)$ is the InfoNCE loss between two modalities -> align multi-modal representations
    - $L_t(A, V)$ is a momentum contrastive loss to align audio and visual representation
    - α is a hyper-parameter

Multi-modal Alignment

- InfoNCE Loss $L_c(A, T/V)$
  - pulling the positive pairs together and pushing the negative pairs apart
    - a_i : i-th audio sample in a mini-batch N
    - m^+ : its positive paired data (m can be v and t for image and text)
    - m_j^-: N negative paired data, $j \in \{1,2, \dots, N\}$
    - audio encoder F_a extracts the audio feature f_a
    - specific encoder F_m extracts the feature f_m
    - sim is the cosine similarity function

Multi-modal Alignment

- InfoNCE Loss $L_c(A, T/V)$
  - Similarity of positive pairs from audio to the other modality

$$p_i^{\mathrm{a2m}} = \frac{\exp(\mathrm{sim}(f_{a_i}, f_{m^+})/\tau)}{\sum_{j=1}^{N} \exp(\mathrm{sim}(f_{a_i}, f_{m_j^-})/\tau)}$$

  - Similarity of positive pairs from the other modality to the audio

$$p_i^{\mathrm{m2a}} = \frac{\exp(\mathrm{sim}(f_{m_i}, f_{a^+})/\tau)}{\sum_{j=1}^{N} \exp(\mathrm{sim}(f_{m_i}, f_{a_j^-})/\tau)}$$

Multi-modal Alignment

- InfoNCE Loss $L_c(A, T/V)$
  - Total InfoNCE Loss:

$$\mathcal{L}_c(A, T/V) = \frac{1}{2}\mathbb{E}_{(A,T/V)\sim D}[\mathrm{H}(\mathbf{y}^{\mathrm{a2m}}, \mathbf{p}^{\mathrm{a2m}}) + \mathrm{H}(\mathbf{y}^{\mathrm{m2a}}, \mathbf{p}^{\mathrm{m2a}})]$$

    - $\boldsymbol{y}^{a2m}$: ground-truth one-hot similarity (positive -> 1, negative -> 0)

Multi-modal Alignment

- Momentum Contrastive Loss $L_c(A, V)$
  - audio-vision pairs are collected from the web -> noisy

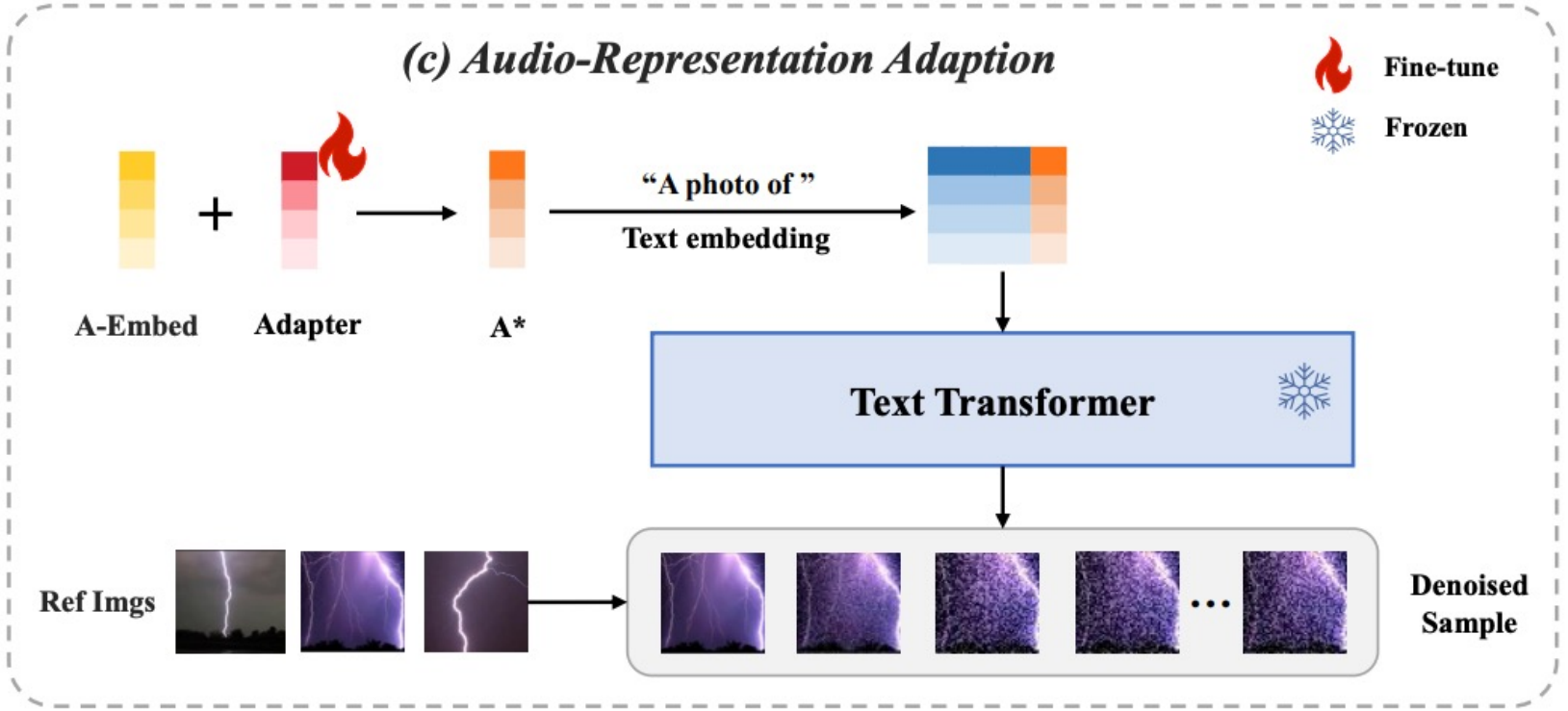
    -> leverage the pseudo-target generated by the momentum visual encoder
  - Loss:

$$\mathcal{L}_t(A, V) = \mathbb{E}_{(A,V)\sim D}[\mathrm{KL}(\mathbf{q}^{\mathrm{a2v}} \| \mathbf{p}^{\mathrm{a2v}}) + \mathrm{KL}(\mathbf{q}^{\mathrm{v2a}} \| \mathbf{p}^{\mathrm{v2a}})$$

    - pseudo targets $\boldsymbol{q}$ are obtained with similarity operation with $\boldsymbol{p}$

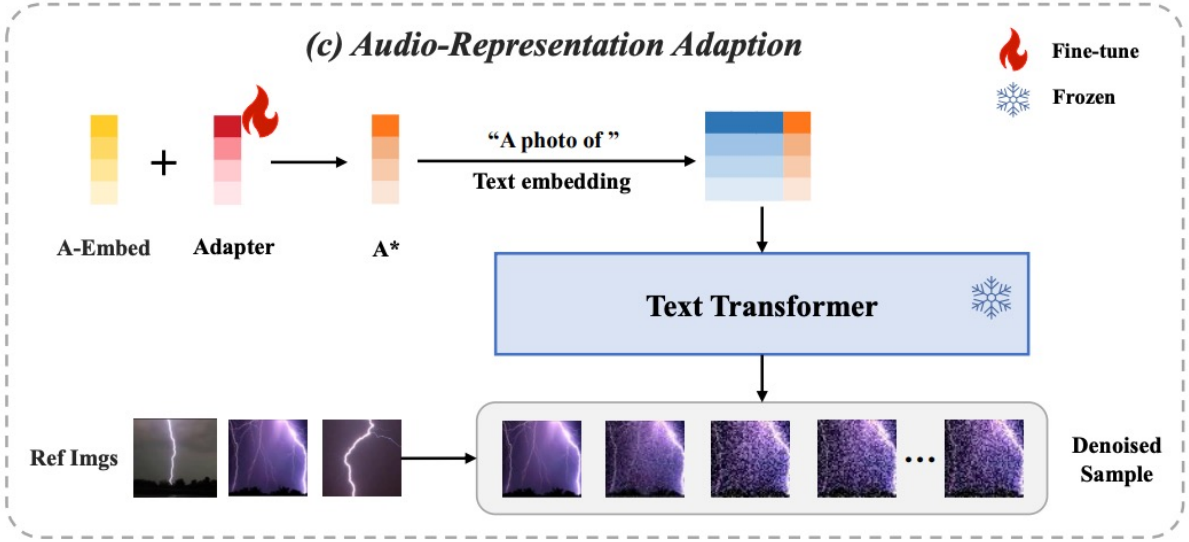
Audio-Representation Adaption

- Audio adapter transforms audio representations into an intermediate form for T2Is



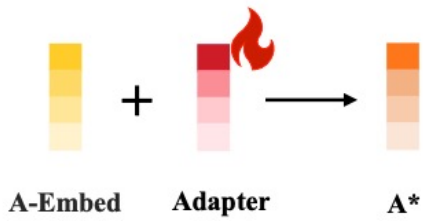
Audio-Representation Adaption

- Goal: Find a new representation A^* which can be used for image-generative tasks
- Adopt the visual reconstruction as the adaption object
  - Minimize the LDM loss with a trainable adapter, $f_{adapter}$, while freezing other parameters



Audio-Representation Adaption

- LDM Loss from Latent Diffusion Model
  - $\mathbb{E}_{z,y,t,\epsilon\sim\mathcal{N}(0,1)}\lfloor\|\epsilon-\epsilon_\theta(z_t,t,c(y))\|^2\rfloor$
- Loss of the audio-representation adaption
  - $\mathcal{L}_A=\mathbb{E}_{z,y,f_a,t,\epsilon\,\sin\mathcal{N}(0,1)}\lfloor\|\epsilon-\epsilon_\theta(z_t,t,c([y,f_a+f_{\text{Adapter}}]))\|^2\rfloor$
    - $f_{Adapter}$ aligns the audio representation f_a within the text embedding space
    - $f_{Adapter}$ is used to make a special pseudo-token A^* for the audio-representation injection



Audio-Representation Injection

- Goal: integrate the pseudo-token A^* into the T2Is framework to enable image generation, editing, and stylization
  - Treat the audio adapter token A^* as a regular word
    - Generation: each A^* is incorporated into new conditioning
    - Editing and stylization: introduce new tokens or replace existing ones in the prompts

Experiments

- Datasets
  - VGG-Sound contains large-scale audio-visual pairs over 200K
  - 166K for training, 14K for testing

Experiments

- Implementation Details
  - NVIDIA A100 GPUs with PyTorch
  - pre-trained models in CAV-MAE (audio and visual encoder)
  - Bert-like model in the Stable Diffusion Model (text encoder)
  - output dimension of encoders = 768
  - mini-batch AdamW optimizer with a weight decay of 0.02
  - momentum updating $m = 0.995$
  - SDM as the generation baseline

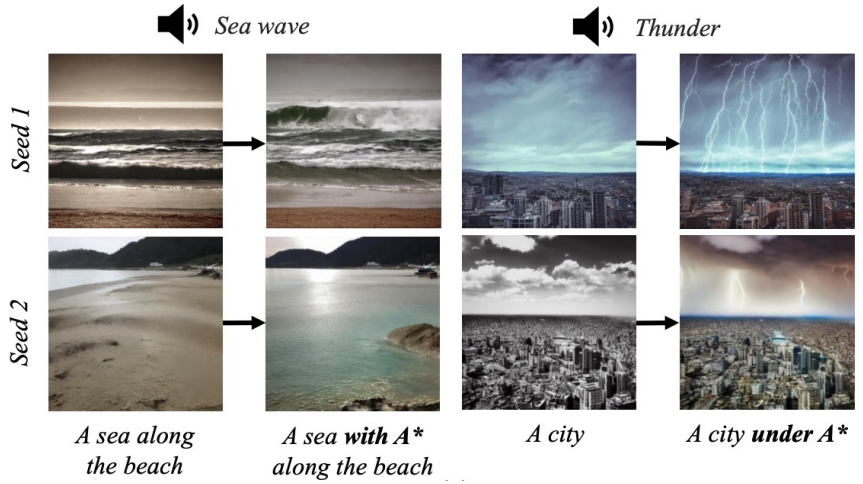
Applications

- Image Generation

  - Each sound token A^* is served as high-level semantic of a scene or object
  - Stereotyped prompts to generate the audible object, like “A photo of A^*”
  - Model can capture the semantics in the audio and provides sufficient details

Applications

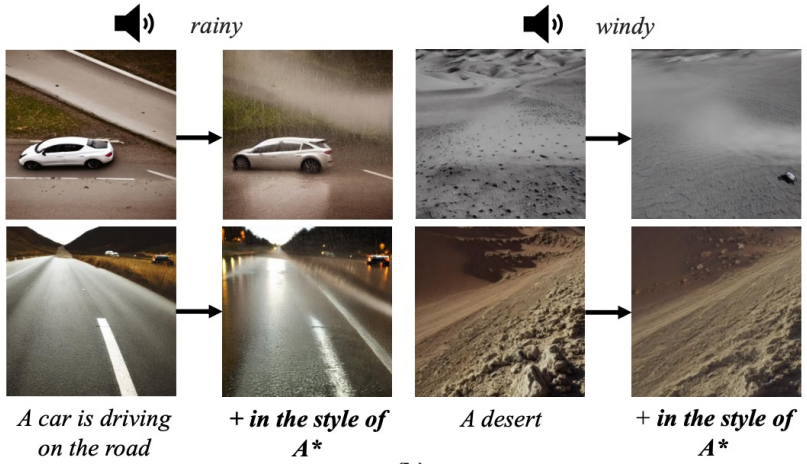
- Image Editing



  - Some pairs of source and target prompts
  - Model can preserve the image composition and edit the image following audio

Applications

- Image Stylization



  - Various properties can be caught by the sound of weather
  - Add audio on the end of any ordinary prompt with ”in the style of A^*”
  - Model can use the audio to represent a specific, vivid style

Quantitative Results

- Audio-Video retrieval results on VGG-Sound
  - w/o m: fine-tuning without the pseudo-targets

	Fine-tune	EvalSubset-1.4K			EvalSubset-2.8K			EvalSubset-14K		
		R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
CAV-MAE [19]	-	0.28	0.77	1.62	0.17	0.60	1.41	0.02	0.15	0.24
AudioCLIP [10]	-	13.84	33.26	44.42	8.36	23.79	34.28	2.69	9.02	14.15
AudioCLIP	✓	21.53	47.24	57.98	14.68	36.82	48.34	5.67	17.17	25.34
CAVT-MAE w/o m	✓	27.04	53.31	64.75	20.76	44.87	55.47	7.96	22.26	31.64
CAVT-MAE$^{\text{Ours}}$	✓	**28.31**	**54.51**	**64.26**	**21.53**	**45.05**	**55.89**	**8.86**	**23.46**	**32.83**

Quantitative Results

- Zero-shot audio classification
  - ESC-50, Urban sound 8k

Model	Modal	Dataset	
		ESC-50	Urban sound 8k
AudioCLIP [10]	A-V-T	4.1 %	18.1 %
AudioCLIP* [10]	A-V-T	6.2 %	19.58 %
Wav2clip [48]	A-V	41.4 %	40.4 %
SGSIM [11]	A-T	**57.8 %**	**45.7%**
CAVT-MAE^{Our}	A-V-T	42.8 %	39.8 %

- Audio-Text retrieval results on VGG-Sound

Model	Fine-tune	Accuracy
AudioCLIP [10]	-	38.94%
AudioCLIP [10]	✓	48.41%
CAVT-MAE w/o m	✓	49.32%
CAVT-MAE^{Our}	✓	**52.16%**

User Study

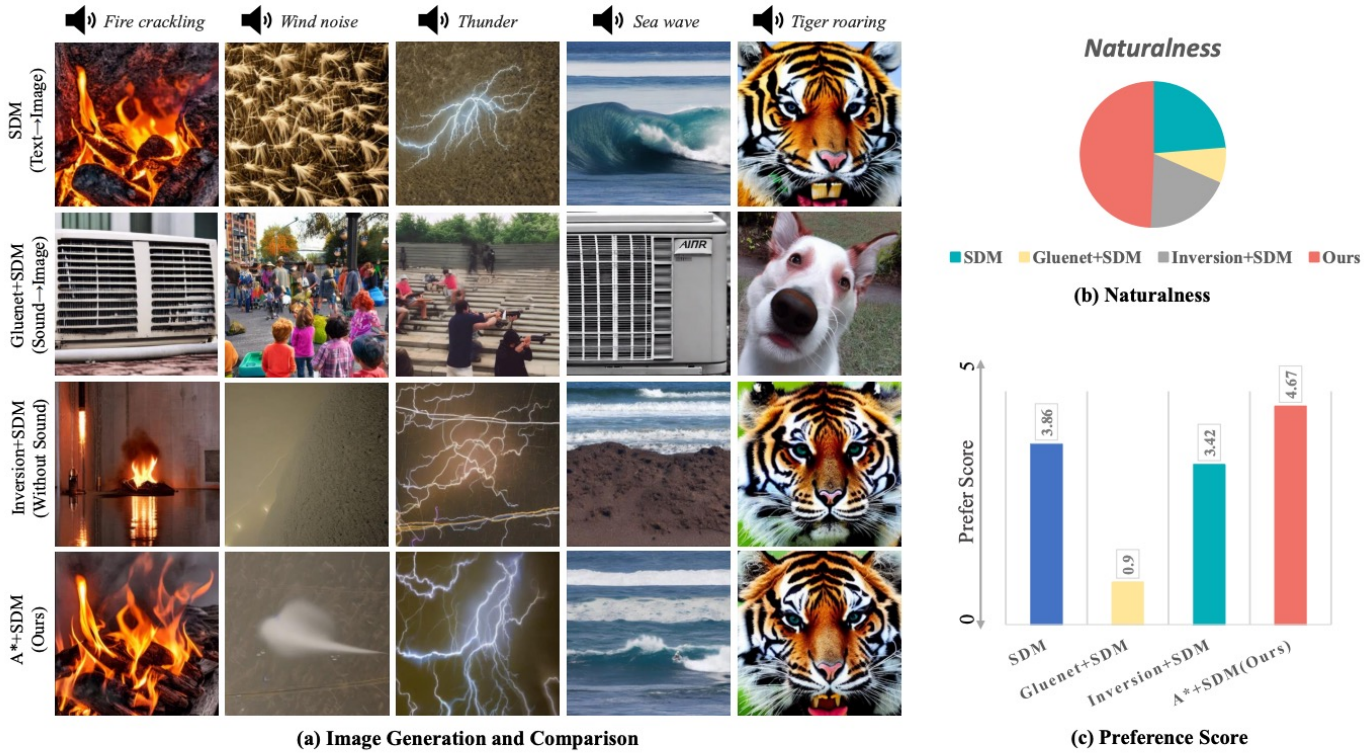
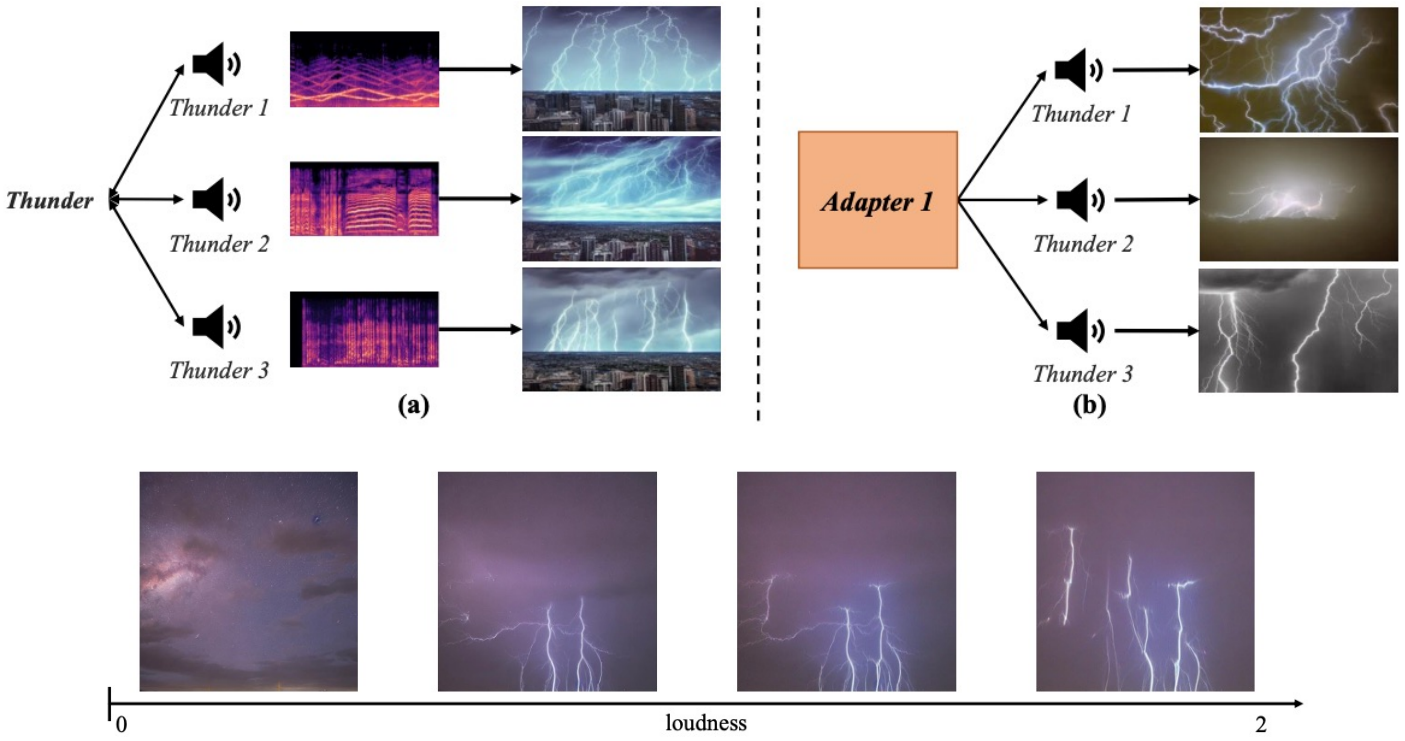


Figure 5: Qualitative comparisons to alternative personalized creation approaches. (a) Our model can more accurately capture the semantics of the audio, enabling generate images which typically more faithful images to the input. (b) Human evaluation results comparing Ours vs. other existing models.

Ablation Study

- Diversity depending on input sounds



Conclusions and Limitations

- Conclusion
  - a unified framework for sound-guided image generation, editing, and stylization
  - aligns the representation of the audio and adapts it into an audio token, which has vivid semantics and can be injected into T2I models flexibly
  - our method successfully captures and utilizes the semantic cue of the audio
- Limitations
  - global audio representation bounds the generative ability
  - the injection of more fine-grained representation will be studied in the future

Questions?

- Thanks to listening!