Supplementary Material

Susan Liang[1], Chao Huang[1], Yapeng Tian[1],
Anurag Kumar[2], and Chenliang Xu[1]

[1] University of Rochester, Rochester NY 14627, USA
[2] Meta Reality Labs Research, Redmond WA 98052, USA

1 Implementation Details

Multimodal One-Shot Adaptation. We design two MLPs to project the $\mathbf{f}_{AV}$ to the embedding space of τ_A and τ_V, respectively. These MLPs share the same architecture, each comprising two linear layers with a width of 1024. We apply a ReLU activation function between the linear layers. During multimodal one-shot adaptation, in addition to updating the parameters of the MLPs, we also update the parameters of key and value matrices in cross-attention layers within diffusion models. We set the learning rate of the audio branch as $5e{-}5$, the vision branch as $5e{-}5$, and the MLP as $1e{-}4$. We use the Adam optimizer to optimize parameters, performing 300 steps with a batch size of 1. We use the DDPM scheduler [2] (1000 diffusion steps) for training and the DDIM scheduler [5] (50 steps) for inference.
Cross-Modal Semantic Enhancement. We apply cross-model semantic enhancement to cross-attention layers in the image diffusion model. After we calculate an attention map M between the query matrix of image patches and the key matrix of text tokens, we scale this attention map M according to Eq. (7). Then we multiply the updated attention map M^* and the value matrix of text tokens. We perform cross-modal semantic enhancement for all inference steps.
Prompt Template for Inference. To benchmark the performance of language-guided audio-visual editing methods on the OAVE dataset, we collect 25 prompt templates (see Table 1). For each prompt template, we design a vision prompt and an audio prompt. Vision and audio prompts have the same editing target but different leading words — vision prompts start with "*an image of {}*" and audio prompts begin with "*a recording of {}*". When we use templates to edit an audio-visual sample, we replace "*{}*" with the category name of this sample, such as "*bird*" or "*bell*."

These prompts can instruct a language-guided audio-visual editing model to add a new object to the user-provided data, assessing the audio-visual composition ability of an editing method. For example, "*{} with a dog barking*" needs the model to insert the sound of a dog barking into the original audio and add the image of a barking dog to the original photo. Additionally, these prompts can demand a model to alter the environment of the user-provided sounding object. For instance, a model should generate an image depicting a cathedral background and an audio clip with noticeable reverberation following the prompt "*{} in a cathedral*."

Table 1: Prompt templates for inference. We design these prompt templates to edit user-provided audio-visual samples.

Vision Prompt	Audio Prompt
An image of {} with a dog barking.	A recording of {} with a dog barking.
An image of {} with a child laughing.	A recording of {} with a child laughing.
An image of {} with birds chirping.	A recording of {} with birds chirping.
An image of {} with waves crashing.	A recording of {} with waves crashing.
An image of {} with people chatting.	A recording of {} with people chatting.
An image of {} with a car passing by.	A recording of {} with a car passing by.
An image of {} with raindrops falling.	A recording of {} with raindrops falling.
An image of {} with leaves rustling.	A recording of {} with leaves rustling.
An image of {} with a train whistle.	A recording of {} with a train whistle.
An image of {} with a cat meowing.	A recording of {} with a cat meowing.
An image of {} in a small room.	A recording of {} in a small room.
An image of {} in a large room.	A recording of {} in a large room.
An image of {} in a cathedral.	A recording of {} in a cathedral.
An image of {} in a big crowd.	A recording of {} in a big crowd.
An image of {} at a bustling marketplace.	A recording of {} at a bustling marketplace.
An image of {} at a lively carnival.	A recording of {} at a lively carnival.
An image of {} under water.	A recording of {} under water.
An image of {} in the rain.	A recording of {} in the rain.
An image of {} in a serene forest.	A recording of {} in a serene forest.
An image of {} on a peaceful beach.	A recording of {} on a peaceful beach.
An image of {} by a crackling fireplace.	A recording of {} by a crackling fireplace.
An image of {} on a tranquil lake.	A recording of {} on a tranquil lake.
An image of {} in a bustling city street.	A recording of {} in a bustling city street.
An image of {} in a mysterious cave.	A recording of {} in a mysterious cave.
An image of {} on a serene mountaintop.	A recording of {} on a serene mountaintop.

2 Conclusion

This paper investigates the novel language-guided joint audio-visual editing problem and proposes a new diffusion-based editing framework. We incorporate multimodal one-shot adaptation and cross-modal semantic enhancement to achieve superior editing quality. We present both quantitative and qualitative results, demonstrating the advantages of our approach over existing methods.

Our current focus lies in image-level audio-visual editing. However, it is imperative to explore the video-level audio-visual editing in future research. Video diffusion models, such as Sora [1], have shown the potential to generate realistic videos mimicking real-world scenarios. Expanding audio-visual editing to the video level would yield promising outcomes. Nevertheless, video editing presents greater challenges compared to our current task, as it requires maintaining temporal consistency and audio-visual synchronization.

Moreover, our framework is built upon two independently trained diffusion models [3, 4]. It is worth utilizing a jointly trained audio-visual model as the foundation for editing audio-visual content, as these models typically produce audio-visual samples characterized by high cross-modal consistency.

References

1. Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., Ramesh, A.: Video generation models as world simulators (2024), https://openai.com/research/video-generation-models-as-world-simulators
2. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
3. Liu, H., Tian, Q., Yuan, Y., Liu, X., Mei, X., Kong, Q., Wang, Y., Wang, W., Wang, Y., Plumbley, M.D.: Audioldm 2: Learning holistic audio generation with self-supervised pretraining. arXiv preprint arXiv:2308.05734 (2023)
4. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022)
5. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2020)