SoundBrush: Sound as a Brush for Visual Scene Editing

Kim Sung-Bin[1], **Kim Jun-Seong**[1], **Junseok Ko**[2], **Yewon Kim**[1], **Tae-Hyun Oh**[1,3,4]

[1]Department of Electrical Engineering, POSTECH, Korea
[2]Department of Statistics, Inha University, Korea
[3]Graduate School of Artificial Intelligence, POSTECH, Korea
[4]Institute for Convergence Research and Education in Advanced Technology, Yonsei University, Korea
sungbin@postech.ac.kr, junseong.kim@postech.ac.kr, thoh.kaist.ac.kr@gmail.com

Abstract

We propose SoundBrush, a model that uses sound as a brush to edit and manipulate visual scenes. We extend the generative capabilities of the Latent Diffusion Model (LDM) to incorporate audio information for editing visual scenes. Inspired by existing image-editing works, we frame this task as a supervised learning problem and leverage various off-the-shelf models to construct a sound-paired visual scene editing dataset for training. This richly generated dataset enables SoundBrush to learn to map audio features into the textual space of the LDM, allowing for visual scene editing guided by diverse in-the-wild sound. Unlike existing methods, SoundBrush can accurately manipulate the overall scenery or even insert sounding objects to best match the input sound semantics while preserving the original content. Furthermore, by integrating with novel view synthesis techniques, our framework can be extended to edit 3D scenes, facilitating sound-driven 3D scene manipulation.

Figure 1: Sound-guided visual scene editing. We propose SoundBrush, a model that can edit visual scenes by adding sounding objects or manipulating the scene to align with the input sound (top). Furthermore, it can be extended to edit 3D visual scenes using the input sound (bottom).

Introduction

Visual scene editing extends beyond image generation by modifying existing images to meet specific user requirements. With its promising applications, this field has evolved to include tasks, such as style transfer (Gatys, Ecker, and Bethge 2016) and translation between image domains (Zhu et al. 2017; Huang et al. 2018). Recently, the remarkable capabilities of text-based generative models have significantly broadened the scope of visual scene editing to include not only 2D images (Tumanyan et al. 2023; Hertz et al. 2023; Brooks, Holynski, and Efros 2023) but also 3D scenes (Haque et al. 2023), guided by text descriptions as control signals.

Although text descriptions are effective, sound has become a unique control signal for visual scene editing (Lee et al. 2022; Li et al. 2022; Li, Singh, and Grover 2023). Unlike text, sound shares the natural association with visual scenes and can provide rich information that text may overlook. For example, the intensity of rainfall can range from light to heavy in sound, whereas a text description might simply note it as "raining." Despite these advantages, existing sound-guided editing methods face significant challenges, such as being limited to generating entirely new scenes rather than editing existing ones (Sung-Bin et al. 2023; Yariv et al. 2023; Qin et al. 2023; Sung-Bin et al. 2024), struggling to insert new objects while only manipulating styles (Lee et al. 2022; Li et al. 2022), and altering the structure of the source image after editing (Li, Singh, and Grover 2023).

Addressing these limitations requires overcoming several challenges. First, effectively capturing informative audio signals to determine how to manipulate or edit the given image is difficult due to the significant information gap between audio and visual signals. Secondly, the scarcity of sound-paired image editing datasets and the difficulty in acquiring them further amplify these challenges, becoming a bottleneck for learning the capabilities of sound-guided visual scene editing.

In this work, we propose SoundBrush, a model that tackles these challenges and uses sound as a brush to edit visual scenes. Specifically, we extend the generative capabilities of the text-to-image model, Latent Diffusion Model (LDM) (Rombach et al. 2022), by incorporating audio information into visual scene editing. To utilize sound as an editing control signal, we augment textual token spaces of LDM to include diverse auditory features and design a mapping network that converts sound into these tokens. Inspired by existing image-editing works (Brooks, Holynski, and Efros 2023), we employ various off-the-shelf models, such as sound source localization (Park, Senocak, and Chung 2024), image inpainting (Suvorov et al. 2022), and Prompt-to-Prompt model (Hertz et al. 2023), to generate a sound-paired visual scene editing dataset for training. This richly generated dataset facilitates joint training of the mapping network and LDM in SoundBrush, enabling the model to edit visual scenes guided by various in-the-wild sound cues.

We validate the efficacy of our proposed SoundBrush by comparing it with existing sound-guided visual scene editing models (Yariv et al. 2023; Qin et al. 2023; Li, Singh, and Grover 2023). Unlike previous methods, SoundBrush can accurately insert sounding objects and edit the overall scenery to reflect the sound semantics, as shown in Fig. 1. Furthermore, by integrating with a novel view synthesis method (Mildenhall et al. 2020), our framework can be extended to edit 3D scenes, enabling sound-guided 3D scene editing. Our main contributions are summarized as follows:

- Proposing SoundBrush, a model that effectively incorporates auditory information to manipulate visual scenes.
- Generating a comprehensive dataset paired with sound cues and visual data, which facilitates the training of models for sound-guided visual scene editing.
- Demonstrating SoundBrush's ability to accurately insert objects or manipulate the overall visual scenes based on sound cues, including extensions to 3D scene editing.

Related Work

Multimodal-guided image editing. Image editing aims to modify source images according to diverse user requirements. Recent advancements, including image-to-image translation (Isola et al. 2017; Liu, Breuel, and Kautz 2017; Huang et al. 2018), object insertion (Li, Singh, and Grover 2023; Yariv et al. 2023), and 3D scene editing (Haque et al. 2023), have leveraged various modalities to meet diverse needs of applications. Driven by the generative power and flexibility of Latent Diffusion Models (LDMs) (Rombach et al. 2022), integrating multiple modalities within LDMs has become foundational for manipulating images using different user inputs, such as images (Yang et al. 2023), drag-and-click actions (Shi et al. 2024), and textual descriptions (Huberman-Spiegelglas, Kulikov, and Michaeli 2024).

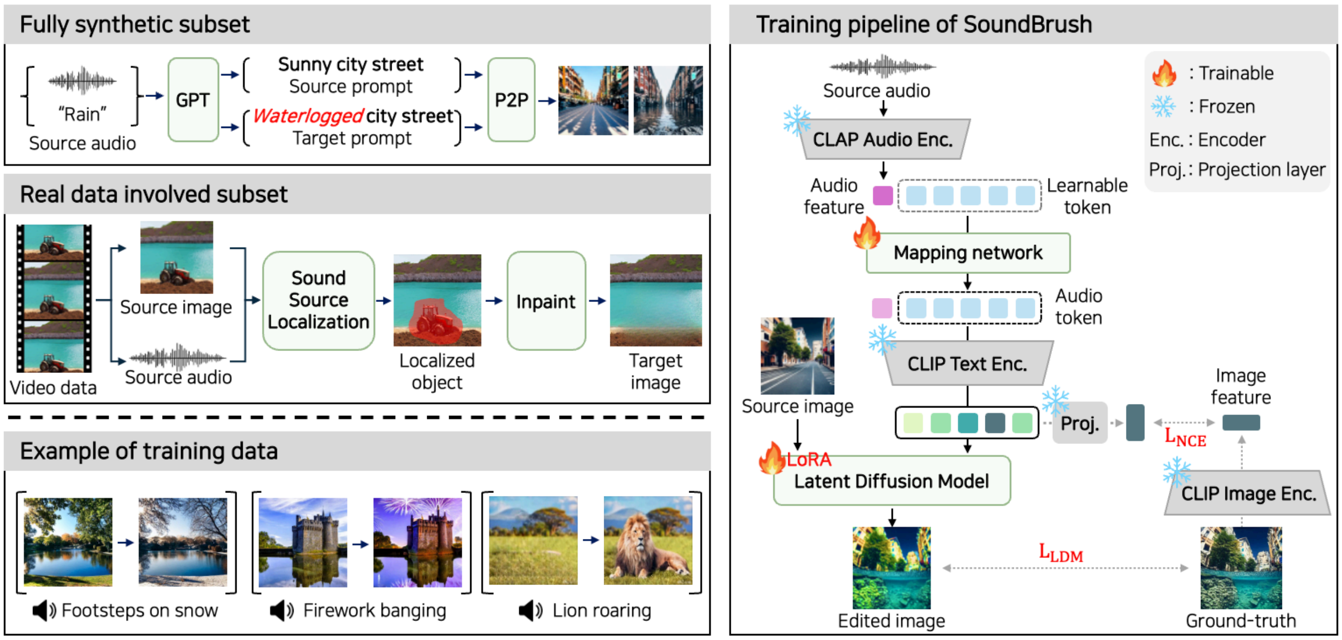
Among these modalities, textual description is a predominant signal for editing, allowing users to describe their editing intentions using words. The introduction of the CLIP model (Radford et al. 2021) has significantly advanced this area by leveraging high-level text-visual embeddings to guide image editing. However, achieving precise control over image changes and ensuring consistent results for specific prompts remain challenging. To address these challenges, methods such as Prompt-to-Prompt (Hertz et al. 2023) and Plug-and-Play (Tumanyan et al. 2023) have shown that controlling the attention map within a diffusion model—originally trained for image generation—can effectively edit images without additional training. These approaches enhance the precision and consistency of edits based on user inputs. Building on the success of text-based methods (Brooks, Holynski, and Efros 2023; Saharia et al. 2022; Kim, Kwon, and Ye 2022), the expansion to other input modalities continues to progress by aligning each modality to these text-based models (Li, Singh, and Grover 2023; Kawar et al. 2023). In this trend, our main goal is to extend the existing text-based editing model to include in-the-wild sound as a control modality. By leveraging expressive semantics in audio inputs, we aim to significantly advance multimodal image editing.

Sound-guided image synthesis. Recent studies in sound-guided image synthesis have shown that the cross-modal association between sight and sound provides important information effectively leveraged for image synthesis. Utilizing this fact, one line of research focuses on generating images from sound inputs. Initially, early works targeted specific sound categories, such as musical instruments (Hao, Zhang, and Guan 2018; Narasimhan et al. 2022), bird sounds (Shim, Kim, and Kim 2021), or speech data (Oh et al. 2019), for image synthesis. With advancements in generative models, more recent methods have achieved remarkable results using Generative Adversarial Networks (GANs) (Sung-Bin et al. 2023) or text-conditioned diffusion models (Yariv et al. 2023; Qin et al. 2023). However, since their main goal is generation rather than editing, these methods often struggle to preserve the original content when used to edit images with sound.

In response to this, another line of research has emerged that proposes editing existing images using sound-based inputs. Lee et al. (2022) expand the embedding space of text-based image manipulation models to include sound inputs, and Li et al. (2022) utilize conditional GANs to adjust the visual style of images to match sounds by learning from unlabeled audio-visual data. More recently, InstructAny2Pix (Li, Singh, and Grover 2023) shows significant improvement in image editing by introducing a model that takes multimodal instructional inputs, including sound and text. Despite their success, these existing sound-guided image editing models primarily focus on style manipulation rather than the insertion of sounding objects (Lee et al. 2022; Li et al. 2022) and often fail to preserve the content of the source image (Li, Singh, and Grover 2023). Our work aligns with the latter approach, proposing a model that can edit and manipulate visual scenes based on sound inputs. Furthermore, we aim to address the limitations of existing methods by accurately inserting sounding objects and editing scenes based on sound, including extensions to 3D scene editing (Haque et al. 2023).

Method

Our goal in this work is to edit visual scenes based on the given in-the-wild audio. We are motivated by the fact that sound contains useful information that can directly influence the editing of an image. This information in the sound may be related to a sound-producing object, *e.g.*, a "Dog barking" and a "Train whistling," or it could indicate general changes in the scene, *e.g.*, "Raining" and "Stream burbling."



(a) Dataset construction pipeline (b) Overall architecture of SoundBrush

Figure 2: Our proposed approach. We start by designing an automatic dataset-construction pipeline as in (a). The dataset is constructed with a fully synthetic subset, involving synthetically generated image pairs paired with audio, and real data involved subset, involving real audio and images. Using this dataset, we train SoundBrush to effectively translate audio features into audio tokens, which can then be used as control signals for the image editing latent diffusion model as described in (b).

Inspired by the prior arts (Brooks, Holynski, and Efros 2023; Li, Singh, and Grover 2023), we approach this as a supervised learning problem. We begin by generating a paired training dataset that includes audio along with images both before and after editing. This dataset is then used to train a conditional diffusion model that uses audio as a control signal, enabling to learn how to accurately edit the visual scene based on the perceived sound. The overall pipeline of training data generation is described in Fig. 2 (a).

Constructing the Training Dataset

Leveraging synthetic datasets has proven effective for training in generative tasks (Brooks, Holynski, and Efros 2023; Li, Singh, and Grover 2023). Given that our dataset requires the incorporation of sound, we have extended these prior works to construct a rich dataset that pairs a sound with an input source image of a visual scene and a target image that reflects auditory information. Specifically, we develop two subsets: one fully synthetic, consisting of generated image pairs with corresponding audio, and the other synthetically generated but incorporating real images and audio.

Fully synthetic subset. The generation of the fully synthetic subset involves creating source and target prompts based on sound events, followed by generating paired images from these prompts (Fig. 2 (a)-row 1). For example, given the source prompt "Sunny city street" and the sound-related keyword "Waterlogged," we create the target prompt "Waterlogged city street," which is then used to generate paired images. For this, we design an automated pipeline using GPT-4 (Achiam et al. 2024) and a text-to-image diffusion model (Rombach et al. 2022). This process includes (1) selecting categories from the sound dataset, (2) creating source and target prompts with sound-related keywords, and (3) generating the corresponding before and after images.

Initially, we extract sound categories from the VGGSound dataset (Chen et al. 2020), focusing on environmental sounds like "Waterfall burbling" and "Thunder." We then generate source prompts by combining various locations, objects, weather conditions, and environments, utilizing GPT-4's extensive language capabilities. Subsequently, we create source-target prompts in large quantities through few-shot prompting with GPT-4, integrating sound-related keywords into the target prompts. This allows for a natural integration of various sound events into the target prompts, beyond simply concatenating keywords. The constructed prompt pairs are used for generating image pairs. Following previous studies (Brooks, Holynski, and Efros 2023), we employ Prompt-to-Prompt (Hertz et al. 2023) to create before and after editing samples. This method adjusts the cross-attention in the denoising process to enhance content similarity while reflecting changes specified in the target prompt.

After generating the samples, we adopt several CLIP-based metrics, including directional similarity and feature similarity between two images in CLIP space, to ensure the quality of the generated subset. Images that fall below a specified threshold are filtered out. Additionally, we incorporate ImageBind (Girdhar et al. 2023) to validate whether the target image accurately represents the sound event by measuring the audio-visual feature similarity in the ImageBind space and excluding those that do not meet the required threshold.

Real data involved subset. Along with the fully synthetic one, we also construct a subset that incorporates real-world data (Fig. 2 (a)-row 2). We begin by extracting paired audio and images from the VGGSound dataset. We then use a sound source localization model (Park, Senocak, and Chung 2024) to detect and localize the sounding object with a coarse mask. Given this coarse mask on the sounding object, we employ LaMa (Suvorov et al. 2022) to inpaint over the mask of the sounding object. The inpainted image serves as the "before editing" image, while the original image is used as the "after editing" image, thus creating triplet pairs for the dataset. However, as neither the sound source localization model nor the inpainting model may be sufficiently reliable, we use ImageBind (Girdhar et al. 2023) to filter out noisy inpainted pairs. We remove pairs where the feature similarity between the audio features and the inpainted image in the ImageBind space is above the threshold, *i.e.*, indicating that the inpainted image still consists the sounding object.

In total, we construct a dataset consisting of 83,614 pairs, with 27,056 fully synthetic and 56,558 involving real data. The example pairs are shown in the Fig. 2 (a)-row 3.

Learning to Edit Visual Scene with Sound

We leverage the strong generative capabilities of the existing Latent Diffusion Model (LDM) (Rombach et al. 2022) and extend this to accommodate sound as a condition for editing. To achieve this, we design a mapping network that translates audio features into a sequence of tokens within the textual spaces of LDM. These tokens are then fed into the LDM to manipulate the visual scene. We train this model with our constructed dataset to jointly learn how to map audio into sequences of tokens in the textual space, while also learning to edit visual scenes using these tokens. The overall pipeline of our proposed SoundBrush is illustrated in Fig. 2 (b).

Preliminary of the text-guided image editing model. We base our model on InstructPix2Pix (Brooks, Holynski, and Efros 2023), which utilizes text instructions for image editing. InstructPix2Pix is built on the LDM (Rombach et al. 2022), which learns the underlying probabilistic model of image data x within the latent space $p(z)$ of the variational autoencoder with a encoder $\mathcal{E}(\cdot)$ and a decoder $\mathcal{D}(\cdot)$, where $z = \mathcal{E}(x)$, and $\hat{x} = \mathcal{D}(\mathcal{E}(x))$. Learning such probabilistic model involves learning the reverse Markov process over a sequence of T timesteps in the latent space. For each timestep $t = 0, \dots, T$, the denoising function $\epsilon_\theta : R^d \rightarrow R^d$, where d is the dimension of the latent, is trained to predict a denoised version of the perturbed z_t at timesetp t, as $\epsilon_\theta(z_t, t)$.

InstructPix2Pix incorporates the conditioning text input c^V during the denoising process by employing the CLIP text encoder as $c^V = \text{CLIP}_\text{T}(V)$, where V represents the textual tokens of the input text instruction. Additionally, InstructPix2Pix adds extra input channels to the first convolutional layer of the LDM to facilitate image conditioning by concatenating z_t with $\mathcal{E}(c^I)$, where c^I is the conditioning image. The corresponding objective can be simplify written as follows:

$$L_\text{LDM} = \mathbb{E}_{z_t, t, \epsilon \in \mathcal{N}(0,I)}[||\epsilon - \epsilon_\theta(z_t, t, \mathcal{E}(c^I), c^V)||_2^2]. \quad (1)$$

Translating audio into textual spaces. While the conditional input c^V in InstructPix2Pix is derived from text instructions, our aim is to use sound solely as the conditional input to guide the model in editing and manipulating visual scenes. To facilitate this, we design a mapping network that translates audio features into tokens within the textual spaces of the LDM. Specifically, given the input audio A, we extract audio features $f^A = \text{CLAP}(A)$, where $\text{CLAP}(\cdot)$ is a pretrained CLAP audio encoder (Wu et al. 2023). These features f^A are then fed into the mapping network $M(\cdot)$, which converts f^A into a sequence of audio tokens in the textual spaces as $V^A = M(f^A)$. The mapping network, built with Transformer encoder layers, takes two different inputs: the audio features and a sequence of learnable tokens. These learnable tokens are designed to extract useful information from the audio features through multi-head attention, while adjusting the tokens to formulate a signal that enables the model to perform editing. Finally, we feed the sequence of audio tokens into the CLIP text encoder to extract text-aligned audio conditions as $c^A = \text{CLIP}_\text{T}(V^A)$. The extracted c^A can then replace c^V for sound-guided visual scene editing.

Learning objectives. One straightforward way to train the mapping network is by optimizing Eq. (1) with c^V replaced by c^A. However, we find that soley optimizing Eq. (1) is insufficient for the mapping network to learn to translate audio into semantically meaningful tokens. To extract useful information from the audio features and effectively map them into textual space in a continuous form, we propose incorporating a contrastive loss between the features derived from audio tokens and the visual features of the ground-truth image. The audio conditions c^A are fed into the projection layer of the CLIP text encoder to extract features, $q^V = P(c^A)$, where $P(\cdot)$ is the projection layer. The ground-truth image I is fed into the CLIP visual encoder to extract visual features, $q^I = \text{CLIP}_\text{I}(I)$. We employ the InfoNCE loss (Oord, Li, and Vinyals 2018), treating pairs of q^V and q^I as positive and those from different pairs as negative in the batch N. The objective is formulated as:

$$L_\text{NCE} = \frac{1}{N} \sum_{j=1}^{N} -\log \frac{\exp(\langle q_j^V, q_j^I \rangle)}{\sum_{k=1}^{N} \exp(\langle q_j^V, q_k^I \rangle)}, \quad (2)$$

where $\langle \cdot \rangle$ is the cosine similarity function. Furthermore, although InstructPix2Pix has demonstrated rich editing capabilities, we find that the model struggles to insert new objects. Therefore, we jointly optimize the mapping network and LDM using Low-Rank Adaptation (LoRA) (Hu et al. 2022), while the CLIP text encoder and the CLAP audio encoder remain fixed. Finally, we add an ℓ_1 regularization to the audio tokens to encourage a more even distribution. Thus, our final objective for training SoundBrush is as follows:

$$L_\text{total} = L_\text{LDM} + \lambda_\text{NCE} L_\text{NCE} + \lambda_{\ell_1} |V^A|_1. \quad (3)$$

Experiments

We validate the editing power of our proposed SoundBrush both qualitatively and quantitatively. We begin by outlining the experimental setup, which includes the dataset, metrics, and competing methods. We then present comparisons of

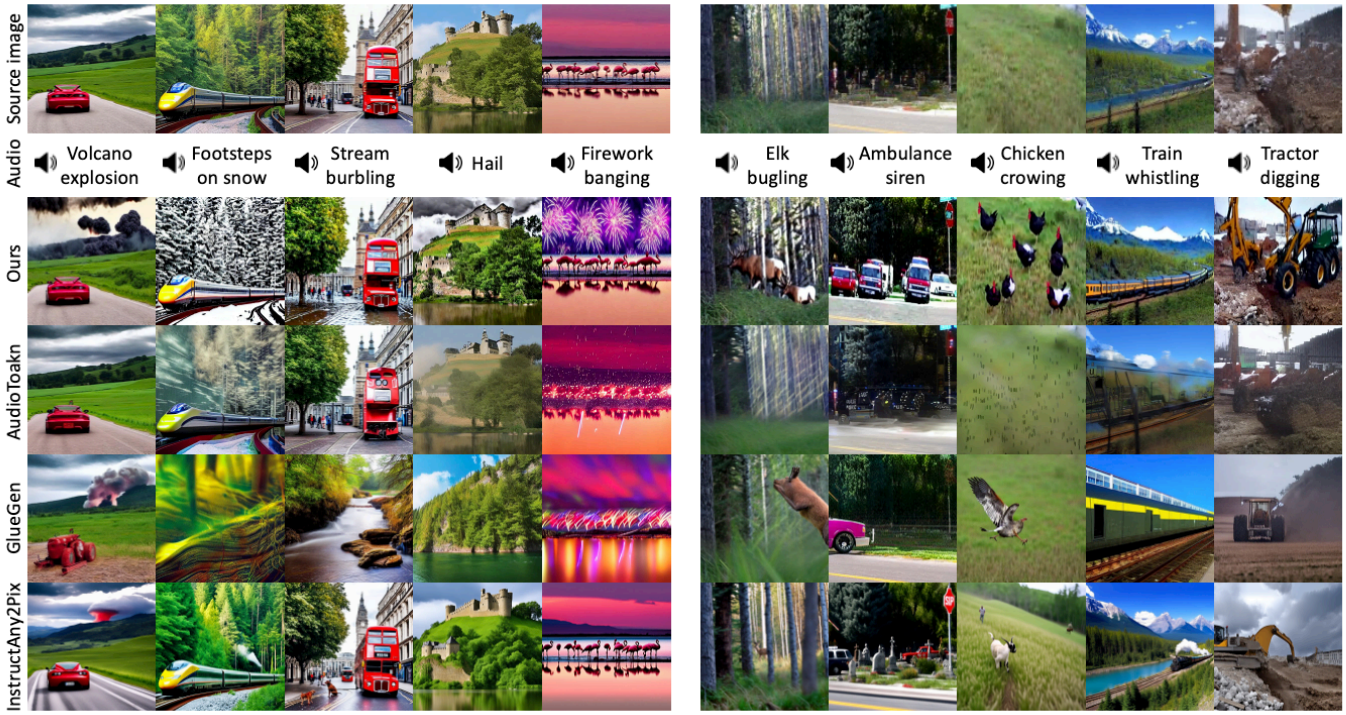


Figure 3: Qualitative comparison. We compare our model with existing editing methods and demonstrate that our model can edit visual scenes using diverse in-the-wild audio while preserving the rest of the content unchanged.

sound-guided 2D visual scene editing between SoundBrush and existing methods. Finally, we demonstrate how SoundBrush can be extended to edit 3D visual scenes.

Experimental Setup

Dataset. We construct the evaluation dataset following the previously described dataset construction pipeline. All the audio files are sourced from VGGSound (Chen et al. 2020), which is an audio-visual dataset containing around 200K videos from 309 different sound categories. We select 20 categories from these and use the provided test splits for constructing the evaluation dataset.

Evaluation metrics. We assess our method using both objective and subjective metrics. For objective metrics, we measure audio-visual similarity (AVS) using ImageBind (Girdhar et al. 2023) space by computing the feature similarity between the input audio and the edited image. We also evaluate image-image similarity (IIS) in the CLIP (Radford et al. 2021) space by comparing the feature similarity between edited and ground-truth images. Text-visual similarity (TVS) is assessed in the CLIP space, evaluating the similarity between the audio category's name and the edited image. Additionally, we measure the Fréchet Inception Distance (FID), which quantifies the distance between the features obtained from real and synthesized images using a pre-trained Inception-V3 (Szegedy et al. 2016). For subjective metrics, we conduct human evaluations to analyze performance from a human perception perspective, focusing on evaluating whether the structure of the edited images remains similar to the original images while accurately reflecting the conditioned sound.

Competing methods. We compare the editing capabilities of our method against three different methods: AudioToken (Yariv et al. 2023), GlueGen (Qin et al. 2023), and InstructAny2Pix (Li, Singh, and Grover 2023). AudioToken and GlueGen are originally targeted for generating images from sound and have demonstrated significant image generation performance. To adapt these models for image editing, we employ a training-free Plug-and-Play (PnP) method (Tumanyan et al. 2023) that enables them to edit visual scenes with sound input. Additionally, as GlueGen is initially trained with UrbanSound8K (Salamon, Jacoby, and Bello 2014), we fine-tune this model using the VGGSound dataset to ensure a fair comparison. We also benchmark against InstructAny2Pix, which utilizes sound and instructional text for image editing.

Results on 2D Image Editing

Comparisons. Figure 3 shows a qualitative comparison between existing methods (Yariv et al. 2023; Qin et al. 2023; Li, Singh, and Grover 2023) and our proposed SoundBrush. As demonstrated, our model exhibits superior capability in manipulating and editing visual scenes based on audio inputs. AudioToken struggles to reflect the audio signal in the edited image while preserving the original structure, and GlueGen reflects the audio signal but modifies the original content beyond recognition. InstructAny2Pix shows favorable editing results; however, if we take a closer look, the original content has been altered. For instance, InstructAny2Pix changes the design of the car and the road when editing with the "Volcano explosion" sound, as shown in the first column of Fig. 3. Using the "Tractor digging" sound for editing (last column) can

Method	AVS (↑)	IIS (↑)	TVS (↑)	FID (↓)
AudioToken	0.204	0.663	0.168	133.3
GlueGen	0.238	0.628	0.202	214.2
InstructAny2Pix	0.222	0.729	0.172	**126.5**
SoundBrush (Ours)	**0.261**	**0.772**	**0.203**	131.7

Table 1: Quantitative comparison. We compare SoundBrush with existing methods and demonstrate that our model outperforms them in overall metrics

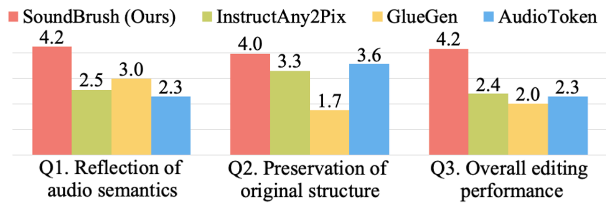


Figure 4: User study. We provide human evaluation results that align with the quantitative comparisons of Table 1.

successfully insert the tractor but also significantly changes the structure of the original image. In contrast, our proposed SoundBrush successfully manipulates the overall scenery or inserts a sounding object into the scene, all while maintaining the original structure of the given image.

We also provide quantitative comparison in Table 1. As summarized, our model consistently outperforms other methods in overall metrics, particularly excelling in the audio-visual similarity (AVS) and image-image similarity (IIS). While InstructAny2Pix achieves better performance on the FID, it shows degradation in other metrics, indicating that it can generate realistic images but often fails to preserve the original structure or accurately reflect the sound semantics in the edited images. These results highlight our model's ability to reflect the audio condition effectively while maintaining the original content and structure of the given image.

User study. We analyze the performance of our proposed SoundBrush from a human perception perspective. We evaluate it based on three criteria, each scored from 1 to 5: Q1. how well the input audio is reflected in the edited image, Q2. how well the original structure and content remain unchanged, and Q3. how well the audio information is reflected while the original structure is unchanged. To conduct this evaluation, we recruit 47 participants to respond to questions about 20 different randomly ordered samples. Figure 4 summarizes the results. As shown, our model is preferred by humans in all the criteria. Especially Q3, which entails both of Q1 and Q2, our model significantly outperforms the other methods which is aligned with the quantitative evaluations.

More qualitative results. Figure 5 demonstrates additional results on sound-guided 2D visual scene editing. Unlike being restricted to a limited number of categories, our model can utilize diverse types of in-the-wild sounds to edit visual scenes. For example, in the second row, the overall city scene is manipulated using environmental sounds, such as "Hail," or by inserting objects that produce sounds, such as "Chicken crowing" and "Ambulance siren."

Figure 5: Additional qualitative results of SoundBrush. From the same source image, SoundBrush can edit the given image using a diverse range of in-the-wild sounds, including environmental sounds, animals, and vehicles sounds.

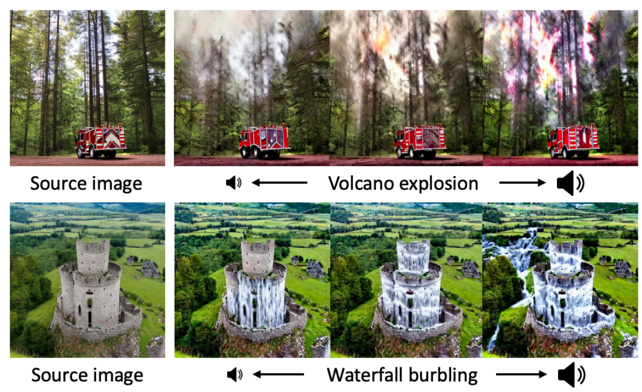


Figure 6: Editing results by different audio intensities. SoundBrush captures the intensity differences in the audio and reflects these changes in the edited images.

In addition, although not explicitly modeled, we observe that our model can reflect the intensity variations inherent in sound during editing, as shown in Fig. 6. This is a distinct property compared to text, where sound uniquely carries intensity information that can be used for editing images. We assess this by manipulating the volume of the reference audio and feeding these varied-volume audios into the model for visual scene editing. For example, as the "Volcano explosion" sound increases, the fire and smoke become larger and more dynamic. These results support that our model not only has a category-specific understanding of audio but also perceives the relationship between audio volume and visual changes.

Ablation study. We conduct a series of experiments to verify our design choices as detailed in Table 2. We evaluate the impact of varying the number of audio tokens in the mapping network and the effect of applying the loss function specified

	Num. of tokens	L_{NCE}	AVS (↑)	IIS (↑)	TVS (↑)	FID (↓)
(A)	1	✓	0.235	0.761	0.192	131.8
(B)	5	-	0.258	0.767	0.199	138.3
(C)	5	✓	**0.261**	**0.772**	**0.203**	**131.7**
(D)	10	✓	0.233	0.763	0.191	136.1

Table 2: Ablation studies. We evaluate the effectiveness of various design choices by comparing different configurations of our method, varying the number of audio tokens, and adapting the L_{NCE} loss function during model training.

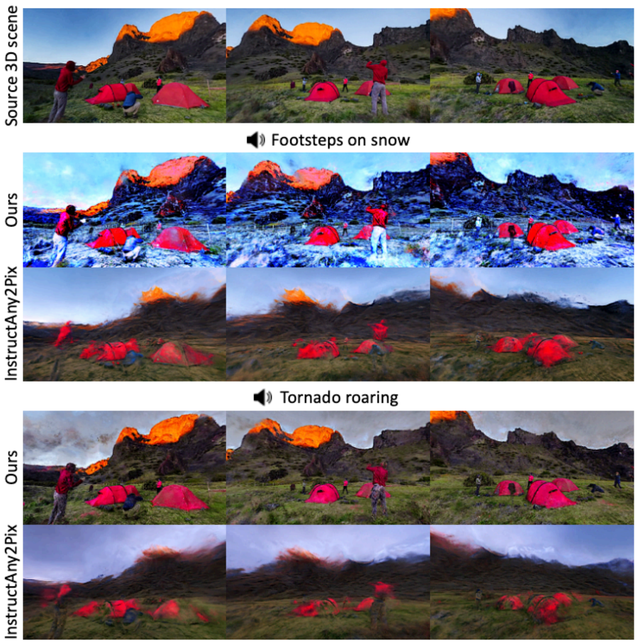


Figure 7: Qualitative comparison of 3D visual scene editing. SoundBrush (Ours) successfully edits the source images while preserving the original geometric properties. In contrast, InstructAny2Pix struggles to maintain the original structure, failing to accurately capture the 3D structure.

in Eq. (2). We find that using a single audio token (A) does not contain sufficient information for sound-guided image editing. Increasing the number of tokens to five (B) results in significant improvements; however, further increasing to ten (C) begins to degrade performance. Additionally, we validate Eq. (2) by comparing results with configurations (B) and (C) and observe that applying Eq. (2) stabilizes model training and leads to improved performance. These insights have guided us to our final model configuration (C).

Results on 3D Image Editing

To further validate sound-guided image editing performance, we conduct a novel experiment called sound-based 3D scene editing. This experiment utilizes the InstructNeRF2NeRF pipeline (Haque et al. 2023), which progressively edits and updates training images using InstructPix2Pix as learning progresses. For this experiment, we modify the InstructNeRF2NeRF model by replacing InstructPix2Pix with either

Figure 8: Examples of failure cases. We observe that our model occasionally inserts sounding objects into the scene without proper spatial understanding.

InstructAny2Pix or our proposed SoundBrush.

Effective 3D scene editing requires making changes only where necessary, without altering the overall structure or content of the image. As illustrated in Fig. 7, InstructAny2Pix leads to inconsistencies in editing across multi-view images, resulting in floating and blurry artifacts in 3D editing. In contrast, our model tends to preserve the original structure while making appropriate changes, achieving accurate 3D editing results with coherent object occupancy. These results demonstrate our model's potential to be further extended for editing 3D visual scenes based on sound cues.

Discussion and Conclusion

Discussion. Although our method demonstrates promising results in editing, it also reveals limitations. As shown in Fig. 8, our model occasionally inserts objects without proper spatial understanding. For instance, a train may appear in the sky, or a chicken might be depicted as large as the building behind it. This issue arises from our design choice, wherein our model does not specify locations for inserting objects but relies solely on audio input. Incorporating text descriptions or learning from multichannel audio could mitigate these issues, which we plan to explore in future work.

Conclusion. In this work, we introduce SoundBrush, a model that leverages audio signals to edit and manipulate visual scenes. By extending the capabilities of Latent Diffusion Models, we demonstrate the potential of using sound as a tool for image editing. Our approach involves constructing a sound-paired visual scene dataset, which is then used to train the model to translate diverse auditory cues into textual spaces, thereby enabling editing based on input sound. As a result, SoundBrush acquires the ability to finely manipulate scenes to reflect the mood of the input audio or to insert sounding objects while preserving the original structure. Moreover, we demonstrate that our model can be integrated with novel view synthesis methods, opening new possibilities for sound-guided 3D scene manipulation.

Acknowledgments

This work was supported by IITP grant funded by Korea government (MSIT) (No.RS-2023-00225630, Development of Artificial Intelligence for Text-based 3D Movie Generation; No.RS-2024-00457882, National AI Research Lab Project; No.RS-2024-00358135, Corner Vision: Learning to Look Around the Corner through Multi-modal Signals; No.RS-2024-00395401, Development of VFX creation and combination using generative AI).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; and et al. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Brooks, T.; Holynski, A.; and Efros, A. A. 2023. Instructpix2pix: Learning to follow image editing instructions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen, H.; Xie, W.; Vedaldi, A.; and Zisserman, A. 2020. VGGSound: A Large-scale Audio-Visual Dataset. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*.

Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2016. Image style transfer using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K. V.; Joulin, A.; and Misra, I. 2023. Imagebind: One embedding space to bind them all. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hao, W.; Zhang, Z.; and Guan, H. 2018. Cmcgan: A uniform framework for cross-modal visual-audio mutual generation. In *AAAI Conference on Artificial Intelligence (AAAI)*.

Haque, A.; Tancik, M.; Efros, A. A.; Holynski, A.; and Kanazawa, A. 2023. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *IEEE International Conference on Computer Vision (ICCV)*.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2023. Prompt-to-prompt image editing with cross attention control. In *International Conference on Learning Representations (ICLR)*.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2022. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Huang, X.; Liu, M.-Y.; Belongie, S.; and Kautz, J. 2018. Multimodal unsupervised image-to-image translation. In *European Conference on Computer Vision (ECCV)*.

Huberman-Spiegelglas, I.; Kulikov, V.; and Michaeli, T. 2024. An edit friendly ddpm noise space: Inversion and manipulations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kawar, B.; Zada, S.; Lang, O.; Tov, O.; Chang, H.; Dekel, T.; Mosseri, I.; and Irani, M. 2023. Imagic: Text-based real image editing with diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kim, G.; Kwon, T.; and Ye, J. C. 2022. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lee, S. H.; Roh, W.; Byeon, W.; Yoon, S. H.; Kim, C.; Kim, J.; and Kim, S. 2022. Sound-guided semantic image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, S.; Singh, H.; and Grover, A. 2023. InstructAny2Pix: Flexible Visual Editing via Multimodal Instruction Following. *arXiv preprint arXiv:2312.06738*.

Li, T.; Liu, Y.; Owens, A.; and Zhao, H. 2022. Learning visual styles from audio-visual associations. In *European Conference on Computer Vision (ECCV)*.

Liu, M.-Y.; Breuel, T.; and Kautz, J. 2017. Unsupervised image-to-image translation networks. *Advances in neural information processing systems*.

Mildenhall, B.; Srinivasan, P.; Tancik, M.; Barron, J.; Ramamoorthi, R.; and Ng, R. 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*.

Narasimhan, M.; Ginosar, S.; Owens, A.; Efros, A. A.; and Darrell, T. 2022. Strumming to the Beat: Audio-Conditioned Contrastive Video Textures. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Oh, T.-H.; Dekel, T.; Kim, C.; Mosseri, I.; Freeman, W. T.; Rubinstein, M.; and Matusik, W. 2019. Speech2face: Learning the face behind a voice. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Park, S.; Senocak, A.; and Chung, J. S. 2024. Can CLIP Help Sound Source Localization? In *IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Qin, C.; Yu, N.; Xing, C.; Zhang, S.; Chen, Z.; Ermon, S.; Fu, Y.; Xiong, C.; and Xu, R. 2023. Gluegen: Plug and play multi-modal encoders for x-to-image generation. In *IEEE International Conference on Computer Vision (ICCV)*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems (NeurIPS)*.

Salamon, J.; Jacoby, C.; and Bello, J. P. 2014. A dataset and taxonomy for urban sound research. In *ACM International Conference on Multimedia (MM)*.

Shi, Y.; Xue, C.; Liew, J. H.; Pan, J.; Yan, H.; Zhang, W.; Tan, V. Y.; and Bai, S. 2024. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shim, J. Y.; Kim, J.; and Kim, J.-K. 2021. S2I-Bird: Sound-to-Image Generation of Bird Species using Generative Adversarial Networks. In *International Conference on Pattern Recognition (ICPR)*.

Sung-Bin, K.; Senocak, A.; Ha, H.; and Oh, T.-H. 2024. Sound2Vision: Generating Diverse Visuals from Audio through Cross-Modal Latent Alignment. *arXiv preprint arXiv:2412.06209*.

Sung-Bin, K.; Senocak, A.; Ha, H.; Owens, A.; and Oh, T.-H. 2023. Sound to visual scene generation by audio-to-visual latent alignment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Suvorov, R.; Logacheva, E.; Mashikhin, A.; Remizova, A.; Ashukha, A.; Silvestrov, A.; Kong, N.; Goka, H.; Park, K.; and Lempitsky, V. 2022. Resolution-robust large mask inpainting with fourier convolutions. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tumanyan, N.; Geyer, M.; Bagon, S.; and Dekel, T. 2023. Plug-and-play diffusion features for text-driven image-to-image translation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wu, Y.; Chen, K.; Zhang, T.; Hui, Y.; Berg-Kirkpatrick, T.; and Dubnov, S. 2023. 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)*.

Yang, B.; Gu, S.; Zhang, B.; Zhang, T.; Chen, X.; Sun, X.; Chen, D.; and Wen, F. 2023. Paint by example: Exemplar-based image editing with diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yariv, G.; Gat, I.; Wolf, L.; Adi, Y.; and Schwartz, I. 2023. Audiotoken: Adaptation of text-conditioned diffusion models for audio-to-image generation. In *Conference of the International Speech Communication Association (INTERSPEECH)*.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*.