Audio-Guided Visual Editing with Complex Multi-Modal Prompts

Hyeonyu Kim[1]
hykim@maum.ai

Seokhoon Jeong[2]
shjd0246@unist.ac.kr

Seonghee Han[2]
seonghee@unist.ac.kr

Chanhyuk Choi[2]
chan4184@unist.ac.kr

Taehwan Kim[2]
taehwankim@unist.ac.kr

[1] MAUM AI Inc.
Republic of Korea

[2] Artificial Intelligence Graduate School
UNIST
Republic of Korea

Abstract

Visual editing with diffusion models has made significant progress but often struggles with complex scenarios that textual guidance alone could not adequately describe, highlighting the need for additional non-text editing prompts. In this work, we introduce a novel audio-guided visual editing framework that can handle complex editing tasks with multiple text and audio prompts without requiring additional training. Existing audio-guided visual editing methods often necessitate training on specific datasets to align audio with text, limiting their generalization to real-world situations. We leverage a pre-trained multi-modal encoder with strong zero-shot capabilities and integrate diverse audio into visual editing tasks, by alleviating the discrepancy between the audio encoder space and the diffusion model's prompt encoder space. Additionally, we propose a novel approach to handle complex scenarios with multiple and multi-modal editing prompts through our separate noise branching and adaptive patch selection. Our comprehensive experiments on diverse editing tasks demonstrate that our framework excels in handling complicated editing scenarios by incorporating rich information from audio, where text-only approaches fail.

1 Introduction

Recent advancements in text-to-image diffusion models have facilitated visual editing tasks, which require precise manipulation of images or videos while preserving essential elements [10, 15, 34, 48]. However, relying solely on text prompts limits their effectiveness in complex editing scenarios where descriptions may be ambiguous or insufficient [5, 14, 20, 57]. To address this limitation, researchers have incorporated non-text conditions—such as edge map, segmentation mask, and depth map—to provide richer and more precise control over the editing process [14, 26, 33, 57]. Audio is one of the widely used modalities and can provide

rich and dynamic information, thus being effective for visual editing. However, in audio-guided visual editing, existing methods often require additional training on specific datasets to align audio with the text [3, 24, 55], limiting their generalization to diverse real-world scenarios.

We address this limitation by leveraging a pre-trained aligned multi-modal encoder trained on large-scale text-audio and audio-visual datasets [46], which demonstrates strong zero-shot capabilities. However, integrating this encoder with diffusion models such as Stable Diffusion [41] faces challenges due to the discrepancy between the encoder's output space and the CLIP text encoder space used in Stable Diffusion, as illustrated in Figure 1. We overcome this incompatibility by introducing a simple mapping technique involving only a matrix inversion, effectively incorporating audio features in visual editing without extensive retraining.

Achieving richer and more complex visual editing often requires integrating multiple editing prompts [14, 26, 33, 57], such as combining audio and text prompts. Existing approaches typically enhance diffusion models with additional modules to handle multiple editing prompts, requiring further training. In contrast, we propose a novel separate noise branching and adaptive patch selection as shown in Figure 1, which can effectively integrate multiple editing signals without training.

To comprehensively evaluate our framework, we construct new benchmark datasets, *PIEBench-multi* and *DAVIS-multi*, by extending existing text-guided visual editing benchmarks to include audio editing prompts. Through comprehensive experiments and analysis, we demonstrate that our framework effectively handles complex editing scenarios by effectively incorporating rich information from audio, which existing text-only methods struggle with. We will publicly release our benchmarks to promote subsequent research.

Our contributions can be summarized as follows:

- We present a novel zero-shot approach to integrate audio into visual editing, which delivers strong performance across diverse types of audio by effectively integrating a pre-trained aligned multi-modal encoder without any further training.

- We propose a novel approach of separate noise branching and adaptive patch selection, which can handle complex visual editing scenarios with multiple editing prompts without further training.

- Comprehensive experiments and analysis on our new benchmark datasets, *PIEBench-multi* and *DAVIS-multi*, demonstrate that our framework effectively handles complex editing scenarios involving both text and audio prompts, where existing text-only methods struggle. We will also release our benchmarks to promote subsequent research.

2 Related Work

2.1 Diffusion Models in Multi-modal Generation

With the groundbreaking achievements in text-to-image diffusion models [36, 40, 41, 42], researchers have actively explored the application of diffusion models in various output modalities, including text [13, 29, 53], audio [28, 38, 43] and video [4, 18, 30]. Furthermore, incorporating non-text conditions, such as video-to-audio [8, 29], image-to-video [12, 55], and audio-to-image [23, 55] has led to impressive results. Notably, the integration of visual, textual, and auditory information as either condition or output [46, 52, 54] enables us to handle a broader range of tasks within a unified framework.
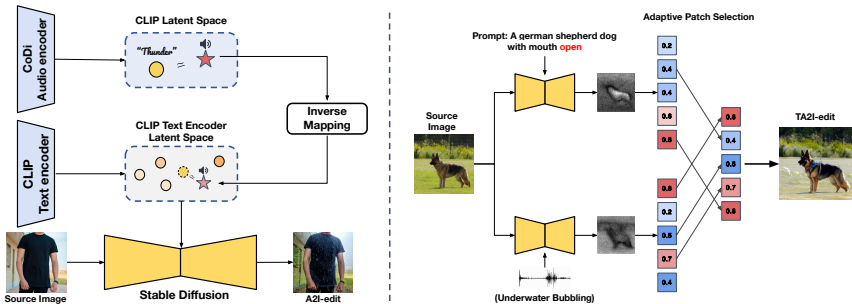
Figure 1: **Left**: To address the limited generalization of previous methods that require additional training to align audio data with text, our framework leverages an aligned multimodal encoder pretrained on large-scale datasets without extra training. **Right**: Instead of enhancing diffusion models to handle multiple editing prompts through additional training, we propose a novel method, separate noise branching and adaptive patch selection, which can effectively combine multiple and multi-modal prompts, while preserving each editing effect.

2.2 Diffusion Models in Editing

Building on the observation that DDIM sampling can be reversed to reconstruct or interpolate visual data [44], numerous studies have focused on diffusion-based editing. These works primarily target image editing with text prompts [15, 17, 34, 48], aiming to achieve faithful inversion for enhanced editing capabilities. Additional control mechanisms have been incorporated alongside DDIM inversion, such as cross-attention [15], unconditional embeddings [34], and spatial and self-attention features [48].

Text-guided video editing has also drawn significant interest [2, 10, 39, 50], with an emphasis on preserving temporal consistency across edited frames. Moreover, the field of diffusion-based editing has expanded to include sound, which is utilized either as a prompt [3, 22, 24, 27, 55] or as the subject of editing [28, 31, 49].

Recently, it has been recognized that text prompts alone are insufficient for handling complex editing scenarios [5, 14, 20, 57]. To address this limitation, researchers have explored incorporating non-text conditions—such as edge map, segmentation mask, and depth map—to achieve richer and more precise control [14, 26, 33, 57]. However, most of these approaches expand the capabilities of diffusion models by adding and training additional modules to the text-to-image diffusion model, which requires extra training with paired datasets. CoDi-2 [45] demonstrates instructional editing capability with multiple and multi-modal prompts, but requires extensive training on paired datasets of original data [15]. In contrast, our approach does not require training with paired data and has strong zero-shot editing ability.

3 Method

3.1 Background

A family of Latent Diffusion Model [28, 41] transforms a high-dimensional input data x into a lower-dimensional latent $\mathbf{z}$ using the encoder $\mathcal{E}$ and decoder $\mathcal{D}$, where $\hat{x} = \mathcal{D}(\mathcal{E}(x))$. Subsequently, a time-conditioned U-net, denoted as θ, conducts a sequential diffusion process to predict $\mathbf{z}_0$ from random noise $\mathbf{z}_T$, considering the textual condition $\mathcal{C} = \psi(\mathcal{P})$ as Equation 1, where ψ refers to the prompt encoder and $\mathcal{P}$ is editing prompt.

$$\mathcal{L}_{ldm} = \mathbb{E}_{\mathcal{E}(x), \varepsilon \sim \mathcal{N}(0,1), t} \|\varepsilon - \varepsilon_\theta(\mathbf{z}_t, t, \mathcal{C})\|_2^2, \tag{1}$$
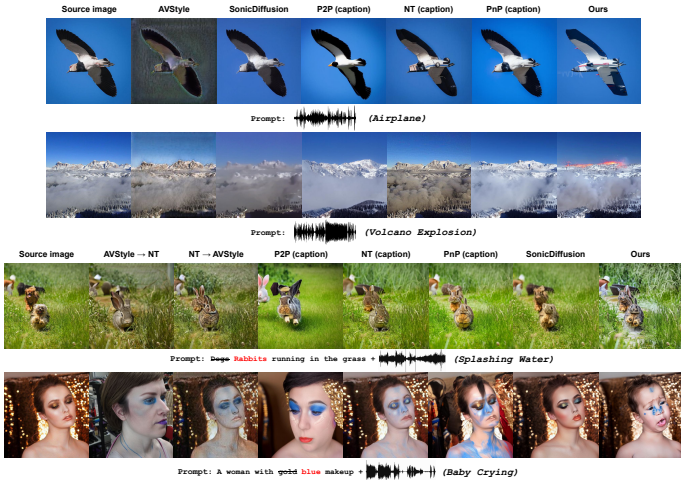
Figure 2: Editing results for Audio-guided Image Editing (**A2I-edit**) and Text and Audio-guided Image Editing (**TA2I-edit**)

DDIM inversion [44] allows the original data to be encoded through an invertible transformation as expressed in Equation 2, resulting in initial noise $\mathbf{z}_T^*$ capable of reconstruction and interpolation. In editing tasks, conditional DDIM inversion with textual description $\mathcal{P}_{inv}$ is commonly employed, where the initial noise $\mathbf{z}_T^*$ is utilized along with the editing prompt $\mathcal{P}$ in sampling stage to maintain the structure and semantic layout. We use $*$ to distinguish the diffusion trajectory of DDIM inversion and editing.

$$\mathbf{z}_{t+1}^* = \sqrt{\frac{\alpha_{t+1}}{\alpha_t}}\mathbf{z}_t^* + \left(\sqrt{\frac{1}{\alpha_{t+1}}-1} - \sqrt{\frac{1}{\alpha_t}-1}\right) \cdot \varepsilon_\theta\left(\mathbf{z}_t^*, t, \mathcal{C}_{\text{inv}}\right) \quad (2)$$

However, this approach often leads to undesired changes [34], prompting the additional controls over the diffusion process for fine-grained editing [7, 32, 34, 48]. PnP-Diffusion [48] has demonstrated the effectiveness of injecting spatial features and self-attention to preserve the original image content. Specifically, spatial features of the fourth decoder layer and the self-attention matrix are extracted during the DDIM inversion, which are then injected in the editing stage. TokenFlow [10], which we adopt for video editing, adapts the PnP-Diffusion to video editing. It first randomly samples a subset of keyframes from the input video and jointly edits them via an extended self-attention mechanism as in Tune-A-Video [50]. It then propagates the edited tokens from each keyframe's self-attention map to its neighboring frames to ensure temporal coherence.

3.2   Training-Free Integration of Audio Prompts

With the rapid advancement of visual editing, research has evolved to integrate more diverse conditions beyond simple text prompts [14, 26, 33, 45, 57]. Most audio-conditioned approaches fine-tune the diffusion model to align sound with text [3, 24, 55], but such adaptation couples the model to the specific training data and generalizes poorly to in-the-wild audio.

To overcome this limitation, we adopt a large-scale multimodal encoder such as ImageBind [11] and CoDi [46]. Specifically, we utilize CoDi because (i) it relies on the same OpenCLIP-L/14 backbone as Stable Diffusion 1.x, common in visual-editing pipelines, and (ii) it is trained on the largest public audio corpus, giving robust audio representations.
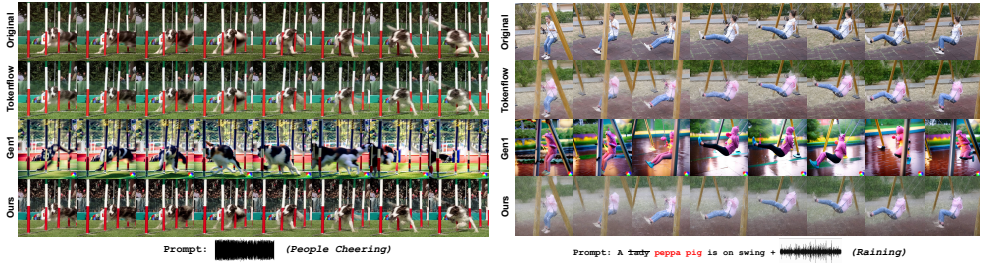
Figure 3: Editing results for Audio-guided Video Editing (**A2V-edit**) and Text and Audio-guided Video Editing (**TA2V-edit**)

However, integrating CoDi with Stable Diffusion is non-trivial, since they operate in different feature spaces. Stable Diffusion directly conditions on the CLIP text encoder's output, while CoDi's audio encoder produces embeddings in the aligned CLIP space. We denote the Stable Diffusion space by $\mathcal{C}_{\text{SD}}$ and the aligned CLIP space by $\mathcal{C}_{\text{CLIP}}$. Given a text prompt $\mathcal{P}$, the CLIP text encoder ψ_{text} generates an SD-space vector $c = \psi_{\text{text}}(\mathcal{P}) \in \mathcal{C}_{\text{SD}}$. Applying pooling to this output yields the special-token embedding $c_{\text{SD,pooled}} \in \mathcal{C}_{\text{SD}}$. We then project and normalize:

$$c_{\text{CLIP}} = \frac{M c_{\text{SD,pooled}}}{\left\| M c_{\text{SD,pooled}} \right\|}, \tag{3}$$

where $M : \mathcal{C}_{\text{SD}} \to \mathcal{C}_{\text{CLIP}}$ is a learned linear mapping with no bias.

Given an audio embedding $c_A \in \mathcal{C}_{\text{CLIP}}$ from CoDi's encoder ψ_{audio}, we invert Eq. 3 and estimate its SD-space counterpart by solving a Tikhonov-regularized least squares problem:

$$\tilde{c}_{\text{SD}} = \left(M^\top M + \lambda I \right)^{-1} M^\top \left(\| c_{\text{inv}} \| c_A \right), \tag{4}$$

where $c_{\text{inv}} = \psi_{\text{text}}(\mathcal{P}_{\text{inv}}) \in \mathcal{C}_{\text{SD}}$, $\lambda = 10^{-5}$, P_{inv} is the inversion prompt, and I is the identity matrix. Finally, we replicate the inverted feature $\tilde{c}_{\text{SD}}$ certain times and concatenate it with the first and last special tokens from c_{inv}. Despite minor precision errors that may arise from matrix inversion, the semantic content is effectively preserved in visual editing.

3.3 Separate Noise Branching with Adaptive Patch-wise Noise Selection

Multi-modal editing commonly requires multiple guidance signals—text, audio, depth, masks, and so on—to be combined within a single diffusion process [14, 26, 33, 57]. Existing works often enlarge the network or attach learned adapters, both of which incur additional training on specific paired dataset. We instead treat multi-prompt fusion as a *variance-preserving aggregation* of the model's native noise estimates, yielding a parameter-free, training-free procedure that can be plugged into any pre-trained latent diffusion model.

Separate noise estimation. Let $z_t \in \mathbb{R}^{c \times h \times w}$ be the latent at diffusion step t and $\mathcal{C} = \{\mathcal{C}_1, \ldots, \mathcal{C}_N\}$ the set of N encoded prompts (possibly of different modalities). For each prompt we query the frozen denoiser

$$\varepsilon_i = \varepsilon_\theta(z_t, t, \mathcal{C}_i), \qquad i = 1, \ldots, N. \tag{5}$$

Limitations of naïve averaging. A naïve fusion, $\hat{\varepsilon}_\theta = \frac{1}{N} \sum_i \varepsilon_i$, reduces variance in proportion to $1/N$. Because diffusion process relies on scaled noise to reconstruct detail, this variance collapse produces suboptimal results, as observed in Fig. 4 (top) and quantified in Sec. 4.
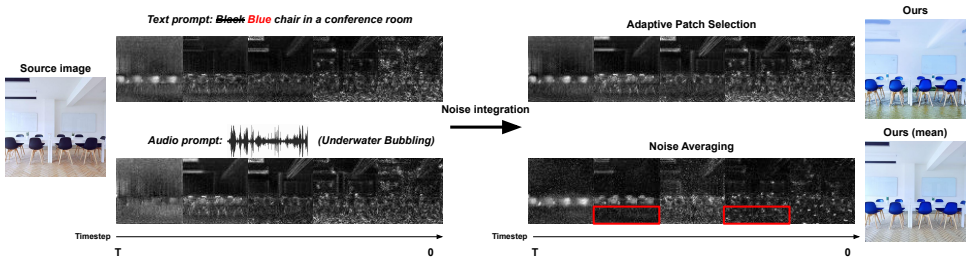
Figure 4: Comparison of different noise integration methods. Directly averaging each noise prediction diminishes important high-frequency components, as evident in the red box. In contrast, our separate noise branching and adaptive patch selection preserve these details while effectively incorporating multiple editing effects.

Adaptive patch-wise Noise Selection. We utilize the inversion prompt $\mathcal{C}_{\text{inv}}$ and its noise prediction $\varepsilon_{\text{inv}} = \varepsilon_\theta(\mathbf{z}_t, t, \mathcal{C}_{\text{inv}})$ to migitage this limitation. Subtracting ε_{inv} from each editing prompt yields residual feature maps $\Delta\varepsilon_i = \varepsilon_i - \varepsilon_{\text{inv}}$, where each $\Delta\varepsilon_i \in \mathbb{R}^{c \times h \times w}$ represents the difference in noise predictions for each prompt. Let p index spatial locations on the $h \times w$ grid. We retain, for every patch, the residual with the largest magnitude:

$$\hat{\varepsilon}_\theta(p) = \varepsilon_{\text{inv}}(p) + \Delta\varepsilon_{i^\star(p)}(p), \quad i^\star(p) = \arg\max_i \|\Delta\varepsilon_i(p)\| \tag{6}$$

, where $\|\cdot\|$ is the channel-wise ℓ_2 norm. Equation (6) preserves high-frequency detail and prevents destructive interference between heterogeneous conditions, as illustrated in Fig. 4 (bottom).

4 Experiments

In this section, we rigorously validate the effectiveness of our framework by constructing various evaluation datasets and baselines that incorporate audio editing prompts. Specifically, quantitative evaluations are conducted across four tasks: Audio-guided Image Editing (**A2I-edit**), Text and Audio-guided Image Editing (**TA2I-edit**), Audio-guided Video Editing (**A2V-edit**), and Text and Audio-guided Video Editing (**TA2V-edit**). Additionally, we explore more complex editing scenarios in Section 4.7.

The results from the A2I-edit and A2V-edit tasks assess whether our audio mapping can produce appropriate editing results while preserving the diverse information contained in the audio. The TA2I-edit and TA2V-edit tasks demonstrate that our noise integration method yields suitable results in complex editing scenarios that existing visual editing approaches cannot handle. The zero-shot capability of our framework can be further verified in the Sections 3, 4, 5, and 6 of the supplementary material.

4.1 Implementation Details

We incorporate diffusion editing techniques such as PnP-Diffusion [43] and TokenFlow [10]. These techniques demonstrate superior edit fidelity and do not require word-level configurations, such as specifying which words to replace or emphasize. We employ Stable Diffusion v1.5 for fairness and utilize A100 with 40GB memory for inference with the seed set to 1 for all experiments.

		Str. Dist(↓)	LPIPS(↓)	CLIP_Audio(↑)	CLIP_Text(↑)	CLIP_integrated(↑)	Structural Preservation(↑)			Audio Alignment(↑)			Text Alignment(↑)			Overall Alignment(↑)		
							Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose
A2I-edit	AVStyle	0.0396	0.2219	14.5596	-	-	**88.40**	7.60	4.00	**90.06**	2.59	7.34	-	-	-	-	-	-
	Sonic Diffusion	0.0475	0.2915	13.4908	-	-	**77.38**	4.96	17.66	**91.27**	4.15	4.59	-	-	-	-	-	-
	P2P (caption)	0.0617	0.3161	12.2536	-	-	**89.66**	2.87	7.47	**82.31**	11.13	6.56	-	-	-	-	-	-
	NT (caption)	**0.0097**	**0.1145**	12.7248	-	-	**78.85**	10.04	11.11	**86.41**	8.70	4.88	-	-	-	-	-	-
	PnP (caption)	0.0227	0.1866	12.8969	-	-	**74.02**	3.52	22.46	**75.04**	8.56	16.40	-	-	-	-	-	-
	Ours	0.0341	0.2539	**17.0472**	-	-	-	-	-	-	-	-	-	-	-	-	-	-
TA2I-edit	AVStyle → NT	0.0668	0.4174	13.1312	25.2388	21.1613	**68.69**	5.05	26.26	**70.42**	5.52	24.06	**58.92**	11.42	29.66	**70.88**	13.65	15.48
	NT → AVStyle	0.0369	0.2508	14.4867	**25.5348**	21.6772	**56.59**	14.10	29.31	**71.73**	20.90	7.36	**46.12**	14.88	38.99	**59.15**	17.41	23.44
	P2P (caption)	0.0591	0.3278	12.3059	24.9569	21.1932	**56.47**	8.04	35.49	**69.48**	11.69	18.83	**48.64**	14.40	36.96	**76.60**	7.35	16.05
	NT (caption)	0.0192	0.1802	12.8379	24.3057	21.6057	**56.59**	14.10	29.31	**71.73**	20.90	7.36	**46.12**	14.88	38.99	**59.15**	17.41	23.44
	PnP (caption)	0.0264	0.2108	13.0695	24.9375	21.3103	**50.83**	16.39	32.78	**76.88**	7.89	15.23	**48.99**	17.31	33.70	**48.12**	27.33	24.55
	Sonic Diffusion	**0.0179**	**0.1551**	12.9011	23.6654	20.1574	**49.80**	9.92	40.28	**63.12**	14.17	22.71	**51.22**	17.64	31.14	**69.61**	12.53	17.87
	Ours (mean)	0.0204	0.1729	14.0025	24.2257	20.1574	39.69	18.51	41.79	55.33	22.79	21.88	27.36	48.69	23.94	39.71	39.92	20.37
	Ours	0.0333	0.2392	**15.0699**	24.1283	**23.0356**	-	-	-	-	-	-	-	-	-	-	-	-

Table 1: Evaluation results for **A2I-edit** and **TA2I-edit**. Left: quantitative scores; right: user study. **Str. Dist** is structure distance. **CLIP_Audio**, **CLIP_Text**, and **CLIP_integrated** are CLIP scores w.r.t. audio, text, and integrated captions. User ratings—Structural Pres., Audio Align., Text Align., Overall—compare ours with the baseline; bold indicates the best result.

4.2 Evaluation Dataset

To perform a comprehensive evaluation of our framework, we introduce *PIEBench-multi* and *DAVIS-multi*, which augment existing text-guided visual editing benchmarks, PIEBench [17] and DAVIS [37] to involve audio editing prompts. We select 28 classes of audio suitable for editing by filtering out low-quality content from VGGSound [6] and choose three audio clips for each category. We utilize 100 images in the PIEBench dataset and use all 89 videos from DAVIS, excluding one unsuitable video (gunshot). We then match appropriate audio editing prompts to each visual data, resulting in 300 editing pairs in PIEBench-multi and 267 editing pairs in DAVIS-multi. Detailed examples of our editing pairs used in the benchmarks can be found in Section 1 of the supplementary material.

4.3 Baselines

Complex audio-guided visual editing is underexplored and lacks baselines, so we first compare our method to representative text-guided approaches: Prompt-to-Prompt (P2P) [15], Null-Text Inversion (NT) [34], and PnP-Diffusion (PnP) [48] for image editing; and TokenFlow [10] and Gen-1 [9] for video editing.

We initially consider a straightforward baselines that uses audio class names either alone (A2I-edit, A2V-edit) or appended to text editing prompts (TA2I-edit, TA2V-edit). Since this approach does not incorporate the textual description $\mathcal{P}_{inv}$ of original data, it induces severe modifications and produces suboptimal results. Section 2 of the supplementary material presents complete performances and examples for these baselines.

To better integrate audio information, we employ an audio captioning model [19] to convert audio into text. Then we refine $\mathcal{P}_{inv}$ with audio caption using GPT-4o [1]. Section 2 of the supplementary material provides details on constructing these refined baselines, which we denote as X(caption).

For A2I-edit task, we compare our approach with AVStyle [24], Sonic Diffusion [3], and caption-based baselines. We exclude SGSIM [21] as it is trained with highly specific domains and tends to produce inappropriate results on our benchmark. For TA2I-edit we include caption-based baselines, Sonic Diffusion, and 12 cascaded pipelines that sequentially run a T2I-edit (P2P, NT, or PnP) and then an A2I-edit (AVStyle or Sonic Diffusion). For A2V-edit and TA2V-edit tasks, we select baselines only using caption-based approaches (TokenFlow and Gen-1) since there are no suitable audio-guided video editing methods on our benchmarks.

		CLIP_frame(↑)	CLIP_Audio(↑)	CLIP_Text(↑)	CLIP_integrated(↑)	Structural Preservation(↑)			Audio Alignment(↑)			Text Alignment(↑)			Overall Alignment(↑)		
						Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose	Win	Tie	Lose
A2V-edit	Tokenflow (caption)	**0.9060**	15.2095	-	-	**61.98**	8.89	29.14	**61.25**	10.47	28.29	-	-	-	-	-	-
	Gen1 (caption)	0.8854	14.5752	-	-	57.07	2.31	40.62	**70.77**	6.52	22.71	-	-	-	-	-	-
	Ours	0.9029	**18.1445**	-	-	-	-	-	-	-	-	-	-	-	-	-	-
TA2V-edit	Tokenflow (caption)	**0.9162**	13.2351	23.6890	22.4227	43.50	19.28	37.22	54.84	9.30	35.86	49.78	21.34	28.88	58.40	9.02	32.58
	Gen1 (caption)	0.8780	14.8816	**23.7246**	22.8530	54.46	0.48	45.06	48.41	10.72	40.87	50.39	5.17	44.44	49.79	8.58	41.63
	Ours (mean)	0.9082	15.4571	22.9909	22.7029	48.67	10.40	40.93	50.54	37.91	11.55	35.46	39.38	25.15	38.19	38.56	23.25
	Ours	0.9043	**16.1833**	22.8336	**23.2885**	-	-	-	-	-	-	-	-	-	-	-	-

Table 2: Evaluation results for Audio-guided Video Editing (A2V-edit) and combined Text and Audio-guided Video Editing (TA2V-edit). **CLIP_frame** represents the frame-wise CLIP similarity score. All other abbreviations are consistent with those used in Table 1.

4.4 Evaluation Metrics

To evaluate our framework, we employ metrics for content preservation and edit fidelity following existing works [10, 15, 34, 48]. For content preservation, we use Structure Distance [47] and LPIPS [56] for image editing tasks on PIEBench-multi (A2I-edit and TA2I-edit). For video editing tasks on DAVIS-multi (A2V-edit and TA2V-edit), we utilize frame-wise CLIP similarity (CLIP_frame) in accordance with the TGVE competition [51].

To evaluate edit fidelity, we apply CLIPScore [16] across all tasks. Specifically, we calculate CLIPScore for each individual editing prompt (CLIP_audio and CLIP_text) and use category names for audio editing prompts. However, in editing tasks with multiple prompts, it is crucial to evaluate how well each prompt is integrated and applied together. To evaluate this, we manually create integrated captions in text form for these tasks and report the results as CLIP_integrated.

Although these metrics assess the preservation of original data and the effectiveness of editing effects, they do not provide a comprehensive assessment. To address this limitation, we conducted a user study on Amazon Mechanical Turk with five participants per question, using a total of 100 samples for image editing tasks and all available samples for video editing tasks. Each participant compared our method against a baseline, selecting the better model according to structural preservation, audio alignment, text alignment, and overall condition consistency. To prevent cases where the absence of edits was rated highly for structural preservation, participants were instructed to give the lowest score for structural preservation if no edits were observed.

4.5 Evaluation Results

Table 1 shows the quantitative evaluation results of A2I-edit and TA2I-edit on PIEBench-multi, and Table 2 presents the quantitative evaluation results of A2V-edit and TA2V-edit on DAVIS-multi. We also show qualitative comparisons in Figure 2 and 3.

Audio-guided editing. Our model significantly outperforms the baselines in A2I-edit and A2V-edit in terms of CLIP_audio. Our method does not show competitive scores in content preservation metrics (Structure Distance, LPIPS, and CLIP_frame) since the baselines tend to preserve the original data excessively and do not well reflect the editing prompt.

Qualitative samples in Figure 2 and Figure 3 allow us to observe these characteristics in detail. In Fig. 2 (top), our method converts a bird to an airplane and adds a volcanic blast to a mountain, while baselines fail. In Fig. 3 (left), our model inserts the crowd following the audio prompt, whereas baselines scarcely change. More samples of the zero-shot capabilities of our framework can be found in Section 3 and Section 5 of the supplementary material. These results demonstrate that our framework effectively incorporates the rich information from audio and achieves diverse editing effects that are difficult to capture using text alone.

Audio and Text-guided editing. Similarly, our model outperforms the baselines in the
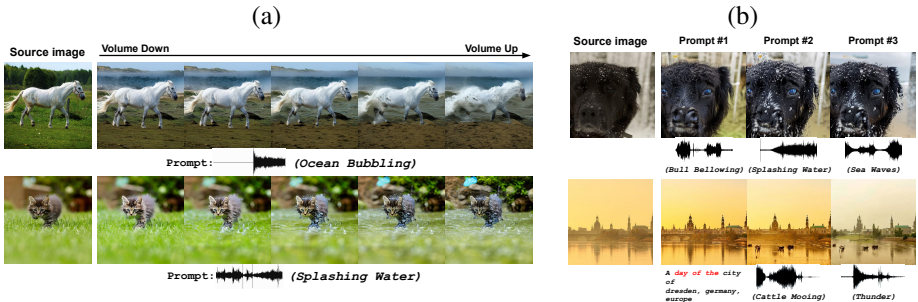
Figure 5: Audio-guided visual editing examples with complex settings—(a) magnitude guided effects and (b) multi-prompt editing.

CLIP_audio and CLIP_integrated. Baselines score well in CLIP_text but drop sharply in other metrics, showing that they capture only part of the desired edits in complex cases, a phenomenon also reported by [5, 20].

In Figure 2 (row 3), our method converts the dog into a rabbit and adds the splashing-water effect that baselines miss. In row 4, it morphs the woman into a blue-made-up baby, while baselines merely recolor the makeup. In Figure 3 (right) our model replaces the lady with Peppa Pig under rainfall, while baselines either omit the rain or over-distort the scene. Additional examples appear in Section 4 and Section 6 of the supplementary material, underscoring how separate-noise branching and adaptive patch selection effectively applies editing effects in complex scenarios.

User Study. Tables 1 and 2 present the results of the user study, demonstrating that our approach significantly outperforms the baselines in all cases, except for slightly lower performance than averaged noise prediction in terms of structural preservation. Although many baselines fail to produce any editing effects in several cases, users were instructed to account for such overly preserved images with the guideline: *EXCEPTION: if an image is unchanged from the original image, consider it as the worst case.* The results demonstrate that our method effectively edits images and videos in alignment with both text and audio while maintaining the original structure.

Note that in the main paper, only AVStyle → NT and NT → AVStyle were included among the 12 cascaded approach baselines for TA2I-edit, as these two methods demonstrate strong performance in A2I-edit and also due to space. We include the results of all baselines in Section 7 of the supplementary material. Also, we observe that CLIP_text reports higher scores than CLIP_audio since the text description of input data is included in the text editing prompt.

4.6   Ablation Study

To demonstrate the effectiveness of our noise integration approach, we conduct an ablation study. Specifically, we compare our method to one that integrates multiple noise predictions by simply averaging them. We report these results as 'Ours (mean)' in Tables 1 and Table 2. Simply averaging (mean) each noise prediction results in a drastic performance drop in CLIP_integrated scores across all tasks, which indicates that our separate noise branching and adaptive patch selection play a crucial role in handling complex multiple editing prompts.

Furthermore, Figure 4 visualizes the differences between the two methods when integrating multiple editing prompts. Specifically, we plot the magnitudes of the separate noise

predictions from each editing prompt and the integrated noise at each diffusion timestep. As indicated by the red boxes, simply averaging the noise predictions causes the information from the audio editing prompt to be canceled out, resulting in the *underwater bubbling* effect not appearing in the final output.

4.7   Analysis and Discussions

Figure 5(a) illustrates the impact of modulating the magnitude of the audio, showing that an increase in magnitude enhances the editing effect. This demonstrates that our framework effectively integrates rich audio information such as magnitude, which is difficult to capture using text alone.

Figure 5(b) illustrates examples of more complex visual editing scenarios involving more than two editing prompts. We add text or audio editing prompts one at a time and visualize each result. We observe that our framework effectively determines which prompts should influence the objects and which should affect the background. Additional examples are provided in Section 8 of our supplementary material.

As for limitations, our frameworks inherit certain constraints from PnP-Diffusion and TokenFlow, particularly a tendency to preserve composition and content excessively or to yield flickering outputs. Although we did not address this in the present study, combining our framework with more powerful diffusion models and editing methods would also be worthwhile. Additionally, our mapping function assumes that audio features align closely with textual counterparts, which may not always be accurate [25]. Utilizing a variety of multi-modal encoders and diffusion models also appears to be a promising direction for future research.

5   Conclusion

In this paper, we introduce a framework for visual editing that seamlessly integrates audio conditions into diffusion models without requiring additional training. By leveraging a pre-trained multi-modal encoder with strong zero-shot capabilities, our approach mitigates the limitations of editing methods with only textual guidance, especially in complex scenarios where textual descriptions would be insufficient or ambiguous. We further introduce separate noise branching and adaptive patch selection, a novel method to handle multiple and multi-modal editing prompts without further training. Extensive experiments on newly established benchmark datasets demonstrate that our framework surpasses existing methods, effectively capturing rich audio information and coherently combining multiple prompts to produce sophisticated and accurate visual editing.

6   Acknowledgement

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Omer Bar-Tal, Dolev Ofri-Amar, Rafail Fridman, Yoni Kasten, and Tali Dekel. Text2live: Text-driven layered image and video editing. In *European conference on computer vision*, pages 707–723. Springer, 2022.

[3] Burak Can Biner, Farrin Marouf Sofian, Umur Berkay Karakaş, Duygu Ceylan, Erkut Erdem, and Aykut Erdem. Sonicdiffusion: Audio-driven image generation and editing with pretrained diffusion models. *arXiv preprint arXiv:2405.00878*, 2024.

[4] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023.

[5] Manuel Brack, Felix Friedrich, Katharia Kornmeier, Linoy Tsaban, Patrick Schramowski, Kristian Kersting, and Apolinário Passos. Ledits++: Limitless image editing using text-to-image models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8861–8870, 2024.

[6] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020.

[7] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022.

[8] Yuexi Du, Ziyang Chen, Justin Salamon, Bryan Russell, and Andrew Owens. Conditional generation of audio from video via foley analogies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2436, 2023.

[9] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023.

[10] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023.

[11] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15180–15190, 2023.

[12] Litong Gong, Yiran Zhu, Weijie Li, Xiaoyang Kang, Biao Wang, Tiezheng Ge, and Bo Zheng. Atomovideo: High fidelity image-to-video generation. *arXiv preprint arXiv:2403.01800*, 2024.

[13] Shansan Gong, Mukai Li, Jiangtao Feng, Zhiyong Wu, and LingPeng Kong. Diffuseq: Sequence to sequence text generation with diffusion models. *arXiv preprint arXiv:2210.08933*, 2022.

[14] Xuehai He, Jian Zheng, Jacob Zhiyuan Fang, Robinson Piramuthu, Mohit Bansal, Vicente Ordonez, Gunnar A Sigurdsson, Nanyun Peng, and Xin Eric Wang. Flexecontrol: Flexible and efficient multimodal control for text-to-image generation. *arXiv preprint arXiv:2405.04834*, 2024.

[15] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

[16] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

[17] Xuan Ju, Ailing Zeng, Yuxuan Bian, Shaoteng Liu, and Qiang Xu. Direct inversion: Boosting diffusion-based editing with 3 lines of code. *arXiv preprint arXiv:2310.01506*, 2023.

[18] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15954–15964, 2023.

[19] Jaeyeon Kim, Jaeyoon Jung, Jinjoo Lee, and Sang Hoon Woo. Enclap: Combining neural audio codec and audio-text joint embedding for automated audio captioning. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6735–6739, 2024. doi: 10.1109/ICASSP48485.2024.10446672.

[20] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1931–1941, 2023.

[21] Seung Hyun Lee, Wonseok Roh, Wonmin Byeon, Sang Ho Yoon, Chanyoung Kim, Jinkyu Kim, and Sangpil Kim. Sound-guided semantic image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3377–3386, 2022.

[22] Seung Hyun Lee, Sieun Kim, Innfarn Yoo, Feng Yang, Donghyeon Cho, Youngseo Kim, Huiwen Chang, Jinkyu Kim, and Sangpil Kim. Soundini: Sound-guided diffusion for natural video editing. *arXiv preprint arXiv:2304.06818*, 2023.

[23] Taegyeong Lee, Jeonghun Kang, Hyeonyu Kim, and Taehwan Kim. Generating realistic images from in-the-wild sounds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7160–7170, 2023.

[24] Tingle Li, Yichen Liu, Andrew Owens, and Hang Zhao. Learning visual styles from audio-visual associations. In *European Conference on Computer Vision*, pages 235–252. Springer, 2022.

[25] Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

[26] Han Lin, Jaemin Cho, Abhay Zala, and Mohit Bansal. Ctrl-adapter: An efficient and versatile framework for adapting diverse controls to any diffusion model. *arXiv preprint arXiv:2404.09967*, 2024.

[27] Yan-Bo Lin, Kevin Lin, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Chung-Ching Lin, Xiaofei Wang, Gedas Bertasius, and Lijuan Wang. Zero-shot audio-visual editing via cross-modal delta denoising. *arXiv preprint arXiv:2503.20782*, 2025.

[28] 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. *arXiv preprint arXiv:2301.12503*, 2023.

[29] Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

[30] Zhengxiong Luo, Dayou Chen, Yingya Zhang, Yan Huang, Liang Wang, Yujun Shen, Deli Zhao, Jingren Zhou, and Tieniu Tan. Videofusion: Decomposed diffusion models for high-quality video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10209–10218, 2023.

[31] Hila Manor and Tomer Michaeli. Zero-shot unsupervised and text-based audio editing using ddpm inversion. *arXiv preprint arXiv:2402.10009*, 2024.

[32] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. *arXiv preprint arXiv:2108.01073*, 2021.

[33] Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7465–7475, 2024.

[34] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023.

[35] Haomiao Ni, Changhao Shi, Kai Li, Sharon X Huang, and Martin Renqiang Min. Conditional image-to-video generation with latent flow diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18444–18455, 2023.

[36] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

[37] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017.

[38] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, pages 8599–8608. PMLR, 2021.

[39] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15932–15942, 2023.

[40] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[41] 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.

[42] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022.

[43] Kai Shen, Zeqian Ju, Xu Tan, Yanqing Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. *arXiv preprint arXiv:2304.09116*, 2023.

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

[45] Zineng Tang, Ziyi Yang, Mahmoud Khademi, Yang Liu, Chenguang Zhu, and Mohit Bansal. Codi-2: In-context, interleaved, and interactive any-to-any generation. *arXiv preprint arXiv:2311.18775*, 2023.

[46] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. *arXiv preprint arXiv:2305.11846*, 2023.

[47] Narek Tumanyan, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Splicing vit features for semantic appearance transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10748–10757, 2022.

[48] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023.

[49] Yuancheng Wang, Zeqian Ju, Xu Tan, Lei He, Zhizheng Wu, Jiang Bian, et al. Audit: Audio editing by following instructions with latent diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

[50] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023.

[51] Jay Zhangjie Wu, Xiuyu Li, Difei Gao, Zhen Dong, Jinbin Bai, Aishani Singh, Xiaoyu Xiang, Youzeng Li, Zuwei Huang, Yuanxi Sun, et al. Cvpr 2023 text guided video editing competition. *arXiv preprint arXiv:2310.16003*, 2023.

[52] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023.

[53] Tong Wu, Zhihao Fan, Xiao Liu, Hai-Tao Zheng, Yeyun Gong, Jian Jiao, Juntao Li, Jian Guo, Nan Duan, Weizhu Chen, et al. Ar-diffusion: Auto-regressive diffusion model for text generation. *Advances in Neural Information Processing Systems*, 36, 2024.

[54] Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. *arXiv preprint arXiv:2402.17723*, 2024.

[55] Yue Yang, Kaipeng Zhang, Yuying Ge, Wenqi Shao, Zeyue Xue, Yu Qiao, and Ping Luo. Align, adapt and inject: Sound-guided unified image generation. *arXiv preprint arXiv:2306.11504*, 2023.

[56] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[57] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.