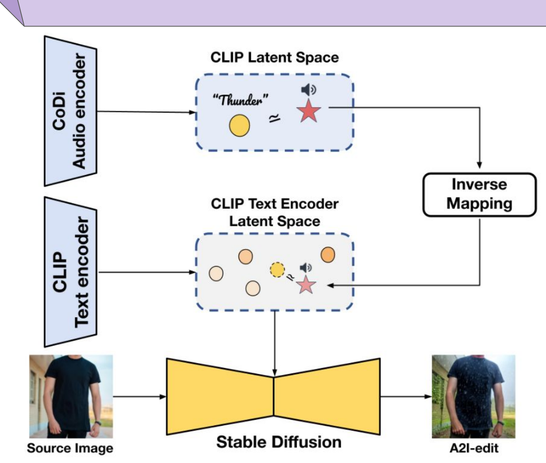
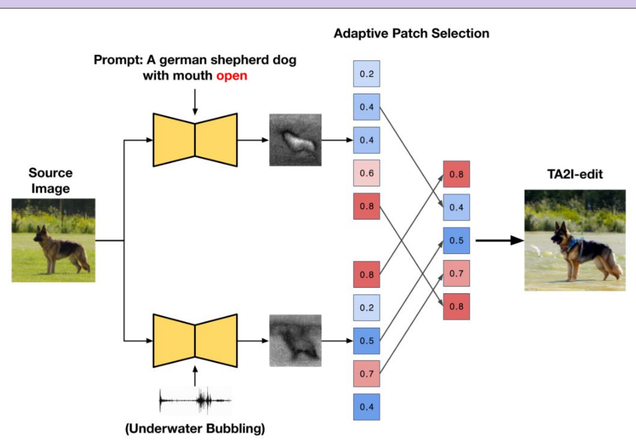
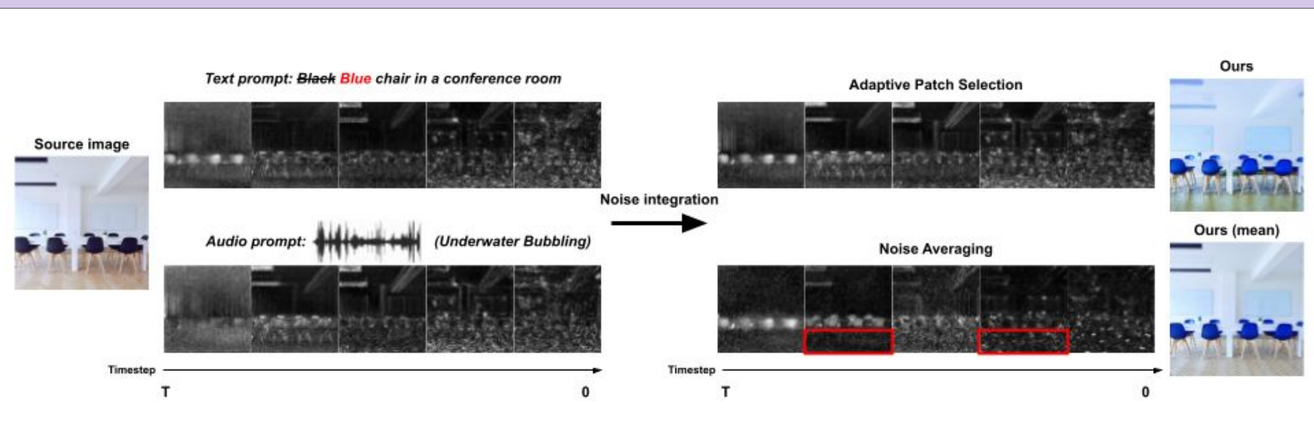
Audio-Guided Visual Editing with Complex Multi-Modal Prompts

Hyeonyu Kim, Seokhoon Jeong, Seonghee Han, Chanhyuk Choi, Taehwan Kim







1. Introduction

- Diffusion models enable powerful visual editing, but text prompts alone are often ambiguous and insufficient.
- Non-text conditions (edge, depth, segmentation) improve control, yet audio remains underutilized despite its rich and dynamic information.
- Existing audio-guided methods require additional training, which limits generalization to diverse real-world scenarios.
- We propose a zero-shot framework that integrates a pre-trained multimodal encoder with diffusion models, enabling effective audio-guided visual editing without retraining.

2. Method

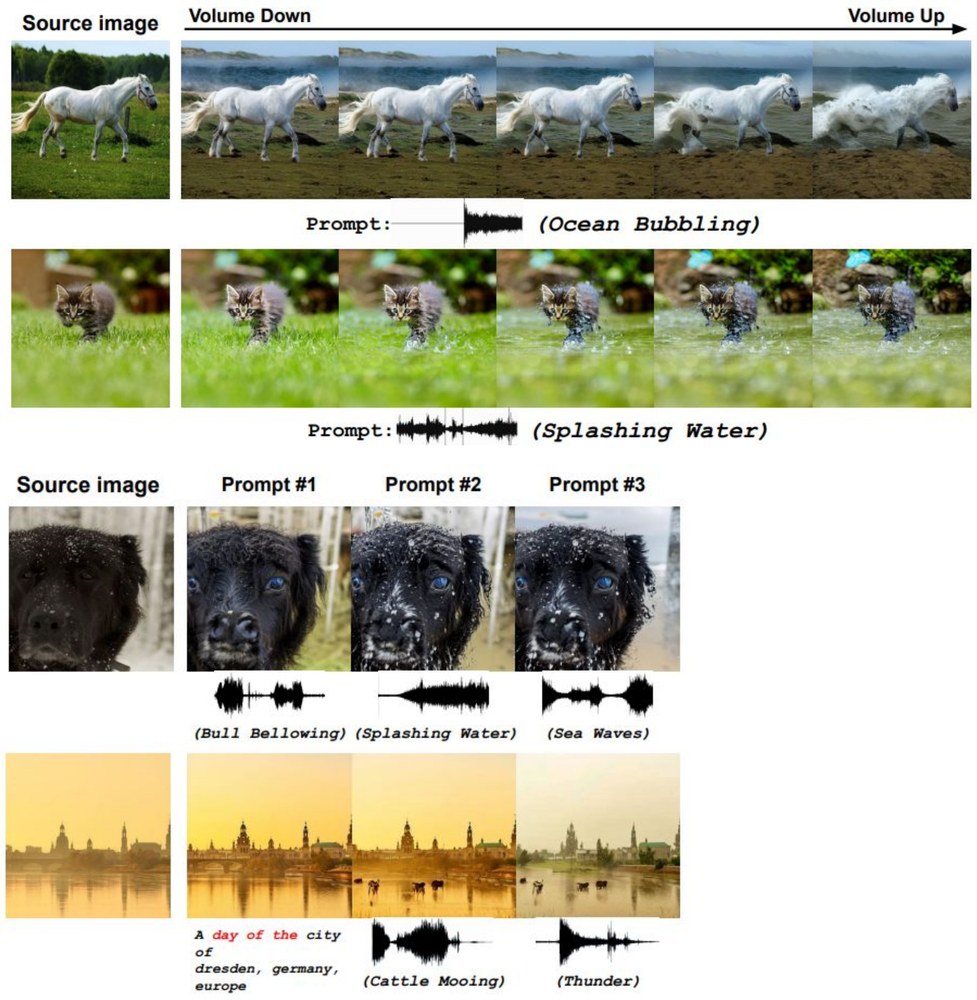
- ***Training-Free Audio Integration***
  - Most audio-guided methods require extra training to align audio with text.
  - We instead leverage a pre-trained multimodal encoder (CoDi) trained on large text–audio–visual data.
  - A simple linear mapping with matrix inversion bridges audio features into Stable Diffusion's space, enabling zero-shot audio-conditioned visual editing without retraining.

$$\mathbf{c}_{\text{CLIP}} = \frac{\mathbf{M}\,\mathbf{c}_{\text{SD,pooled}}}{\left\|\mathbf{M}\,\mathbf{c}_{\text{SD,pooled}}\right\|}, \quad \tilde{\mathbf{c}}_{\text{SD}} = \left(\mathbf{M}^{\top}\mathbf{M} + \lambda I\right)^{-1}\mathbf{M}^{\top}\left(\|\mathbf{c}_{\text{inv}}\|\,\mathbf{c}_{A}\right),$$

- ***Separate Noise Branching + Adaptive Patch Selection***
  - Multi-modal editing often involves complex scenarios that combines multiple multi-modal editing prompts.
  - Naïve averaging of noise predictions reduces detail due to its variance collapse.
  - We propose:
    - Separate noise estimation per prompt.
    - Patch-wise selection of strongest residual features.
    - Preserves high-frequency detail while integrating multiple editing conditions.

3. Experiments

- **Main Results**
  - *Audio-guided editing (A2I-edit / A2V-edit).*
    - Our method achieves substantial gains in CLIP_audio over all baselines.
    - Content preservation metrics (Structure Distance, LPIPS, CLIP_frame) appear higher for baselines, but this is due to over-preservation of baselines, where they fail to reflect the audio prompts.
  - *Text + Audio-guided editing (TA2I-edit / TA2V-edit).*
    - Our approach outperforms in CLIP_audio and CLIP_integrated, demonstrating effective fusion of multiple multi-modal prompts.
    - Baselines maintain high CLIP_text scores but fail to capture full effects, reflecting partial edits only.
- **Ablation Study**
  - Mean noise averaging collapses variance and leads to major drops in CLIP_integrated across all tasks.
  - Our separate noise branching + patch-wise selection preserves high-frequency detail and prevents audio information from being canceled out.
  - Visualization confirms that simple averaging erases effects such as underwater bubbling, while our method retains them.
- **Analysis & Discussion**
  - Audio magnitude control: Increasing audio magnitude naturally strengthens editing effects.
  - Multi-prompt scenarios: Our framework handles more than two prompts simultaneously, deciding which affect objects and which affect background.



		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	0.3261	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**	-	-	-	-	-	-	-	-	-	-	-	-

		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**	-	-	-	-	-	-	-	-	-	-	-	-