AudioControl: Efficient Audio-to-Image Generation via Global Latent Modulation and Audio–Visual Alignment

Junbin Zhang[1,*], Tao Fang[2], Peng Jin[3]
[1]*School of Electronic and Computer Engineering, Shenzhen Graduate School, Peking University, China*
[2]*Institute of International Language Services Studies, Macau Millennium College, Macao, China*
[3]*School of Environment and Energy, Shenzhen Graduate School, Peking University, China*
junbinzhang@pku.edu.cn, taofang@mmc.edu.mo, jinpeng@pkusz.edu.cn

Abstract—Recent advances in diffusion and flow-matching models have enabled high-fidelity text-to-image synthesis. However, conditioning these models on audio remains challenging, as token-level fusion through cross-attention increases the attention length and introduces substantial computational overhead. We present AudioControl, an efficient audio-to-image conditioning module for flow-matching generators that adapts global affine conditioning for audio guidance with minimal overhead. AudioControl extracts a compact global audio embedding with a pre-trained Audio Spectrogram Transformer (AST) and maps it through a lightweight projection to obtain modulation vectors. These vectors are fused additively with pooled text embeddings to form a unified global condition, which is then injected into Transformer blocks via adaptive layer normalization. By modulating feature statistics instead of concatenating audio tokens, AudioControl preserves the original attention sequence length and avoids the quadratic scaling associated with longer contexts, while still steering the full visual latent space. Experiments on VGGSound and GreatestHits demonstrate improved audio-visual alignment with competitive image quality, while substantially reducing the parameter footprint and inference latency relative to token-based conditioning strategies.

***Index Terms*—Audio-to-image generation, global latent modulation, audio–visual alignment, flow-matching, efficient conditioning**

I. Introduction

Modern text-to-image (T2I) generators—from latent diffusion backbones such as Stable Diffusion [1] to more recent flow-matching [2], [3] and rectified-flow architectures such as SD3 [2] and FLUX [4]—have achieved remarkable photorealism, shifting the bottleneck toward controllability. Accordingly, a large body of work has focused on incorporating control signals through lightweight adapters or condition encoders (e.g., ControlNet [5], IP-Adapter [6], and EasyControl [20]) while keeping the pretrained backbone intact. In parallel, parameter-efficient fine-tuning (PEFT) methods [7] such as LoRA [8] have become a practical way to adapt large generative models without updating all parameters. Audio-driven control, however, remains comparatively underexplored [9]–[11]. Audio conveys environmental and material cues (e.g., the timbre of wood impacts or the ambience of rain) that are often difficult to describe precisely in text yet are highly informative about scene appearance. Most existing audio-to-image approaches, such as AudioToken [10], SonicDiffusion [11], and Seeing Sound [12], encode audio as additional pseudo-tokens and fuse them with text via cross-attention or token concatenation. Although flexible, this design lengthens the attention context and introduces non-trivial training and inference overhead. These observations motivate a pluggable audio-conditioning mechanism that improves audio-visual alignment while preserving the computational profile of the base T2I backbone.

We argue that acoustic signals often describe the overall atmosphere or material properties of a scene rather than precise spatial layouts. As a result, representing audio as a sequence of localized tokens can be both computationally inefficient and semantically redundant. To address this, we introduce AudioControl, a lightweight adapter for efficient audio-visual alignment in flow-matching generators. Rather than proposing a new modulation primitive, AudioControl adapts established global conditioning machinery (Global Latent Modulation) to the audio modality: a pre-trained Audio Spectrogram Transformer (AST) [13] extracts global acoustic embeddings, which are mapped by a lightweight projection layer into the generator's global conditioning space. The resulting vector is then additively fused with the pooled text embedding and injected into the backbone through AdaLN-style [14], [15] conditioning, thereby modulating the visual latent space.

Because the token sequence is unchanged, AudioControl is substantially faster than concatenation-based baselines. We also observe that lower-rank LoRA does not always reduce wall-clock latency, so we tune the adapter based on measured latency rather than GFLOPs alone. Our main contributions are:

- We instantiate an audio-conditioning design for flow-matching generators that replaces token-level audio fusion with global latent modulation, thereby preserving the backbone attention length.
- We propose a simple additive global fusion mechanism that combines a projected audio embedding with the pooled text embedding, injecting audio cues (e.g., material and ambience) without extending the attention context.
- We provide an empirical efficiency analysis, including the observed non-monotonic rank-latency behavior of LoRA configurations, and show that AudioControl improves audio-image alignment while maintaining fast inference.

The data and code are available at: https://github.com/wolf-bailang/AudioControl.

II. Related Work

Audio-to-Image Generation. Most audio-to-image pipelines bridge the modality gap by converting audio into extra tokens (or pseudo-tokens) and mixing them with text through token concatenation or cross-attention [10]–[12], [16]–[18]. AudioToken [10] and SonicDiffusion [11], for instance, append audio-derived tokens to the prompt or context and rely on the backbone to learn cross-modal alignment. This strategy is expressive, but it also expands the attention context and increases computational cost as the context grows. In contrast, we encode audio as a single global vector and inject it through global feature modulation, in the spirit of FiLM/AdaIN/style-based conditioning [14], [15], [19]. Practically, this design aligns with the observation that many sounds act as scene-level cues, such as ambience, material, or event intensity, and therefore allows conditioning without incurring the sequence-length penalty of token-based fusion.

Efficient Adaptation of Generative Models. With the shift from latent diffusion backbones [1] to recent flow-matching and rectified-flow generators [2]–[4], efficient adaptation of large pretrained models has become increasingly important. PEFT methods such as LoRA [8] and lightweight conditioning adapters (e.g., ControlNet [5], IP-Adapter [6], EasyControl [20]) are widely used to avoid full fine-tuning. Global AdaLN-based conditioning is already a standard design in diffusion Transformers such as DiT [25]. AudioControl, therefore, does not introduce a new normalization primitive; instead, it investigates how this conditioning interface can be repurposed for audio guidance and compared with token-based audio fusion under unified efficiency metrics. We keep the generative backbone frozen and train only a small audio projection layer together with LoRA, while reusing strong pretrained encoders (AST [13], T5 [21], CLIP [22]) to provide stable audio/text representations. This design keeps the trainable footprint small and, importantly, preserves fast inference.

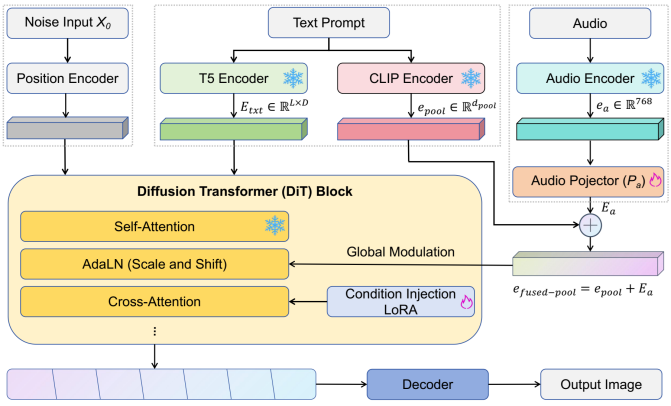

Fig. 1. AudioControl architecture.

III. Methodology

A. Preliminaries and Notation

a) Image Latent Space: A target image I is encoded by a VAE and packed into a sequence of visual tokens $X \in \mathbb{R}^{d\times D}$, where d denotes the number of visual tokens and D the token embedding dimension.

b) Dual Text Encoders: We adopt a dual-encoder text representation. A CLIP encoder [22] produces a pooled global text feature $e_{pool} \in \mathbb{R}^{d_{pool}}$ used for global conditioning. In parallel, a T5 encoder [21] outputs a sequence of token embeddings $E_{txt} \in \mathbb{R}^{L\times D}$, which serve as Key/Value inputs in the cross-attention layers.

c) Flow-Matching: Flow-matching [3] learns a time-dependent velocity field $v_t(x_t)$ for $t \in [0,1]$ that transports samples from a simple prior p_0 (e.g., $\mathcal{N}(0,I)$) to the target data distribution p_1. The flow is defined by the ODE

$$\frac{dx_t}{dt} = v_t(x_t), \quad x_0 \sim p_0. \tag{1}$$

A common construction defines an interpolation between a data sample $x_1 \sim p_1$ and a noise sample $z \sim p_0$:

$$x_t = (1-\alpha_t)\, x_1 + \alpha_t\, z, \quad \alpha_0 = 1,\ \alpha_1 = 0. \tag{2}$$

Differentiating this trajectory yields the analytic target velocity along the path. A parametric model $v_\theta(x_t, t)$ is trained via velocity regression (optionally with a time-dependent weight $w(t)$).

d) Conditional Velocity Field: For conditional generation, we learn a conditional velocity $v_\theta(x_t, t, C)$, where C denotes the conditioning signals. In AudioControl, $C = \{e_{fused\_pool}, E_{txt}\}$: the fused global audio-text condition used for modulation and the text sequence embeddings used in cross-attention.

B. Additive Fusion for Global Modulation

1) Audio Encoding and Projection

Given an input audio waveform, a pretrained AST [13] encodes it into a global audio embedding $e_a \in \mathbb{R}^{768}$. A learnable linear projection P_a maps e_a to the same dimensionality as the pooled text feature e_{pool}, producing the audio modulation vector E_a:

$$E_a = P_a(e_a) = W_a e_a + b_a, \tag{3}$$

where $W_a \in \mathbb{R}^{d_{pool}\times 768}$ and $b_a \in \mathbb{R}^{d_{pool}}$ are trainable, and d_{pool} is the pooled text embedding dimension.

2) Global Additive Fusion

We form the fused global condition via element-wise addition:

$$e_{fused\_pool} = e_{pool} + E_a. \tag{4}$$

Geometrically, this operation performs a semantic shift within the global conditioning manifold, biasing generation towards audio-consistent imagery without disrupting the text-driven structural layout. For completeness, and for the ablation in Sec. V-C, we also summarize two alternative fusion rules below.

3) Fusion Design

Beyond simple addition, one may consider alternative fusion rules. We summarize three representative strategies that trade off expressivity and efficiency.

a) Additive (ours): We use the pooled fusion in Eq. (4), which preserves the original cross-attention context length L and injects audio through a global conditioning vector.

b) Gated Fusion: A lightweight extension learns gates to balance modalities:

$$e_{fused\_pool} = \alpha\, e_{pool} + \beta\, E_a, \tag{5}$$

where (α, β) can be learnable scalars or produced by a small MLP.

c) Token-level Concatenation (Cross-Attention): A token-level baseline appends an audio token t_a to the text context:

$$\tilde{E}_{txt} = [E_{txt};\, t_a], \tag{6}$$

which increases attention cost as quantified in Eq. (8).

C. Formalization: Global Modulation via Conditional Affine Normalization

AudioControl injects audio conditions through global affine transformations inside Transformer blocks (e.g., AdaLN), a conditioning interface widely used in diffusion Transformers such as DiT [25]. Let z denote an intermediate hidden state. We modulate it as

$$\hat{z} = \mathrm{AdaLN}(z, c) = \gamma(c) \odot \frac{z - \mu(z)}{\sigma(z)} + \beta(c), \tag{7}$$

where c is the global conditioning vector (here $c = e_{fused\_pool}$), $\mu(\cdot)$ and $\sigma(\cdot)$ denote feature-wise statistics, and $\gamma(\cdot)$ and $\beta(\cdot)$ are learned linear heads that produce scale and shift vectors.

Crucially, the cross-attention context remains the original text token sequence E_{txt}. AudioControl therefore introduces no increase in attention sequence length and preserves the computational profile of the base model.

D. Audio–Visual Modality Inductive Bias

A key modeling choice in multimodal generation is *where* and *how* to inject the control signal. Token-level cross-attention can align modalities locally, but it implicitly assumes a fine-grained correspondence between modality tokens and spatial or semantic regions. For audio, this assumption is often mismatched: many sounds are scene-level descriptors (e.g., "wind", "rain", or "crowd noise"), affecting overall ambience and material impression rather than a specific object region.

a) Structure vs. Texture: We interpret this as a *structure-texture* inductive bias. Text typically specifies *structure*: objects, relations, layout, and compositional constraints, which are naturally expressed by the token sequence $E_{txt} \in \mathbb{R}^{L\times D}$ used in cross-attention. In contrast, audio often correlates with *texture/style*: material cues (wood/metal/water), global illumination mood, weather ambience, and event intensity. These attributes tend to be broadcast across the generated image and therefore do not require a longer context.

b) Why AdaLN is a Matched Injection Site in FLUX.1: FLUX.1-like diffusion and flow Transformer uses conditional normalization or modulation (e.g., AdaLN/FiLM-style heads) to control global activation statistics across blocks. In our formulation, the global condition $c = e_{fused\_pool}$ produces affine parameters $(\gamma(c), \beta(c))$ that modulate hidden states via Eq. (7). This mechanism acts as a global, block-wise style controller: it translates and rescales the feature manifold across tokens, which is well matched to expressing texture, material, and ambience attributes.

Therefore, AudioControl's global injection is not merely an efficiency trick; it is a structurally aligned conditioning design that assigns *text* to token-level structure (through E_{txt}) and *audio* to global texture/style (through AdaLN modulation), yielding a principled alternative to sequence-level multimodal fusion.

E. Computational Complexity: Why Zero Sequence Overhead Matters

We formalize the efficiency advantage of global modulation over token concatenation. Let L denote the text context length, and suppose an audio-token fusion method appends K audio tokens to the context, yielding length $L' = L + K$. A standard attention layer has time complexity $\mathcal{O}(L'^2 D)$ (and memory complexity $\mathcal{O}(L'^2)$), so the incremental cost is

$$\Delta\mathcal{C}_{attn} = \mathcal{O}((L+K)^2 D - L^2 D) = \mathcal{O}((2LK + K^2)D). \tag{8}$$

In Table III, the token concatenation baseline uses a single appended audio token, i.e., $K = 1$ in Eq. (8). By contrast, AudioControl keeps the cross-attention context fixed at length L and injects audio only through the pooled global condition $c = e_{pool} + P_a(e_a) \in \mathbb{R}^{d_{pool}}$. The additional computation is dominated by the projection P_a and the affine heads (γ, β), i.e.,

$$\Delta\mathcal{C}_{global} = \mathcal{O}(d_{pool} \cdot 768) + \mathcal{O}(d_{pool} D), \tag{9}$$

which is independent of L and introduces *zero sequence overhead*. Because the same global condition is broadcast across tokens, AudioControl is better suited to scene-level semantics than to spatially localized audio cues; we revisit this limitation in Sec . V-E.

IV. Experiments

a) Datasets: Following prior work [10]–[12], we evaluate AudioControl on two standard audio–visual benchmarks: VGGSound and Greatest Hits.

b) Implementation Details: AudioControl is built on FLUX.1-dev with CLIP pooled text features, T5 token features, and the default FLUX.1 VAE. Audio embeddings come from *MIT/ast-finetuned-audioset-10-10-0.4593* (768-d), and the projector is *nn.Linear(768, 768)*. The backbone is frozen; only the projector and LoRA layers are trained. We use one NVIDIA A800 (80 GB), 512×512 resolution, AdamW with a constant learning rate of 1e-4, global batch size 512, bf16, LoRA rank 128, $\alpha = 12$, and about 30000 steps.

c) Metrics: We report audio–image similarity (AIS) [10], image–image similarity (IIS) [10], and FID [23].

V. Evaluation

A. Main Experimental Results

Table I reports results on VGGSound [24] and GreatestHits [12], together with model complexity and inference cost. Dashes denote statistics unavailable from the original papers; such methods are included as quality references but excluded from strict efficiency comparisons.

AudioControl offers the strongest overall alignment-efficiency trade-off. On VGGSound, it achieves AIS 0.046, outperforming Seeing Sound (0.043) and SonicDiffusion (0.038), while maintaining competitive FID. On GreatestHits, it achieves the best AIS (0.057) and IIS (0.79), together with the lowest reported FID of 85.02. Among methods with available runtime statistics, AudioControl is also markedly more efficient (50.01M parameters, 2.89 GFLOPs, 1.41 s/sample) than AudioToken (1165.40M, 7.17G, 6.63 s/sample) and SonicDiffusion (894.14M, 5.50G, 4.72 s/sample). AudioControl's IIS on VGGSound (0.52) is lower than that of Sound2Scene (0.60) and AudioToken (0.55). We attribute this gap to the different emphases of IIS and AIS: IIS rewards closeness to a reference image, whereas stronger global audio conditioning can shift the background, texture, or ambience toward a more audio-consistent scene. This trade-off is more pronounced on the visually diverse VGGSound benchmark, where better audio alignment does not always translate into higher reference-image similarity. On GreatestHits, whose scenes are more constrained, AudioControl improves both AIS and IIS.

Qualitatively, AudioControl more consistently translates acoustic material cues into visual attributes. Fig. 2 shows representative held-out examples in which the baselines often exhibit semantic drift, whereas our global modulation strategy better matches the generated environment to the underlying sound.

Case Studies. We highlight representative rigid and non-rigid material scenarios from Fig. 2. (1) *Rigid Materials (Metal, Wood, Rock)*: For "Metal" and "Rock," AudioControl interprets high-frequency resonance and low-frequency thuds as cues for surface hardness, rendering specular reflections for metallic objects and rugged textures for stones. In the "Wood" case, the model more reliably reproduces grain-like textures, whereas the baselines often drift toward generic backgrounds and miss the material interaction. (2) *Non-Rigid and Stochastic Textures (Plastic-bag, Water)*: The crinkling sound of a "Plastic-bag" and the broadband splashing of "Water" correspond to complex acoustic patterns. AudioControl maps these cues into high-frequency visual details, such as translucent folds of plastic or chaotic droplets and the sense of "wetness" in a watery scene. By contrast, text-only or token-based models often generate visually plausible yet acoustically mismatched scenes. These examples support the view that global latent modulation provides a useful inductive bias for texture and materiality, especially when audio contributes cues that are absent from sparse text prompts.

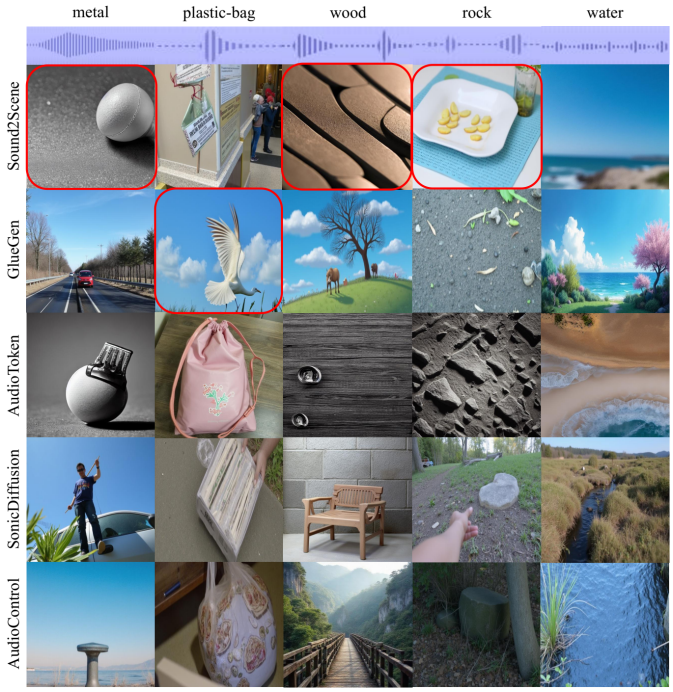

Fig. 2. Qualitative comparisons on GreatestHits. AudioControl better translates holistic audio cues (material, scene ambience, and event semantics) into visual attributes.

B. Ablation Studies and Efficiency Analysis

Table II evaluates the impact of audio injection components and LoRA rank r.

Modality Alignment (P_a). Removing the projection layer P_a (Raw AST) results in a significant AIS drop (0.057 $\rightarrow$ 0.049). This confirms that P_a acts as a modality aligner, mapping AST features into the generator's global conditioning manifold rather than functioning as a mere dimensionality adapter. We retain a linear projector because it adds negligible overhead and isolates the benefit of modality alignment without introducing additional nonlinear capacity; exploring deeper projectors is left for future work.

Empirical Rank-Latency Behavior. We observe that reducing r from 128 to 32 lowers theoretical GFLOPs but slightly increases wall-clock latency (1.41 s $\rightarrow$ 1.47 s). This trend is consistent with a roofline-style interpretation of GPU kernels:

$$t \approx \max\left(\frac{F}{P_{peak}}, \frac{B}{BW}\right), \tag{10}$$

where F denotes FLOPs and B memory traffic. As r decreases, F drops, but B remains dominated by fixed memory-access overheads. The computation therefore shifts from compute-bound to memory-bound ($AI \downarrow$). In practice, a higher rank ($r = 128$) can better saturate the GPU's Streaming Multiprocessors while also providing stronger representational capacity. We treat this as a systems-level interpretation of the observed latency trend rather than a definitive profiling study; a broader sweep over additional rank values is left for future work.

TABLE I
QUALITATIVE COMPARISONS ON GREATESTHITS. AUDIOCONTROL BETTER MAPS HOLISTIC AUDIO CUES TO VISUAL ATTRIBUTES.

Method	Year	VGGSound			GreatestHits			#Params	GFLOPs	Latency
		AIS↑	IIS↑	FID↓	AIS↑	IIS↑	FID↓	(M)	(G)	(s/sample)
Sound2Scene [17]	CVPR 2023	0.037	**0.60**	**85.8**	0.049	0.68	184.2	120.22	3.35	2.13
GlueGen [18]	ICCV 2023	0.030	0.49	140.0	0.022	0.58	192.0	155.68	3.50	2.52
AudioToken [10]	INTERSPEECH 2023	0.034	0.55	115.70	0.053	0.60	190.60	1165.40	7.17	6.63
SonicDiffusion [11]	arXiv2024	0.038	0.47	120.40	0.029	0.72	87.30	894.14	5.50	4.72
Seeing Sound [12]	arXiv2025	0.043	0.51	103.50	0.053	0.61	228.40	-	-	-
AudioControl (ours)		**0.046**	0.52	101.10	**0.057**	**0.79**	**85.02**	50.01	2.89	1.41

TABLE II
ABLATION ON AUDIO INJECTION, PROJECTION LAYER P_a, AND LORA RANK. GFLOPS REPORTS TOTAL THEORETICAL COMPUTE PER INFERENCE STEP (BACKBONE + ADAPTERS). TRAINABLE PARAMS INCLUDE LORA AND P_a (IF USED). TIME IS MEASURED AS THE AVERAGE WALL-CLOCK TIME FOR A SINGLE IMAGE GENERATION (50 STEPS).

Variants	Audio	P_a	LoRA Rank r	#Params (M)	GFLOPs (G) ↓	Latency (s) ↓	AIS ↑
AudioControl (Ours)	Yes	Yes	128	50.01	2.89	1.41	0.057
Text-only Baseline	No	No	128	46.12	2.84	1.12	0.042
Raw AST (w/o P_a)	Yes	No	128	46.12	2.87	1.39	0.049
AudioControl (low-rank)	Yes	Yes	32	15.48	2.81	1.47	0.053

Fusion Strategy. As shown in Table III, additive fusion ($e_{pool} + E_a$) offers the best efficiency, maintaining zero sequence overhead while achieving competitive alignment.

C. Fusion Strategy: Efficiency vs. Alignment

We compare the fusion strategies defined in Sec . III-B3. Token concatenation uses one appended audio token ($K = 1$), so its extra attention cost follows Eq. (8). Table III evaluates three representative paradigms: additive fusion, gated fusion, and token-level concatenation ($K = 1$). Although token concatenation yields a marginal AIS gain (0.058 vs. 0.057), it incurs a 20% increase in GFLOPs and a 31% increase in wall-clock latency (1.41 s $\rightarrow$ 1.85 s). Thus, even a single appended audio token noticeably enlarges the computational footprint of attention. By contrast, additive fusion preserves zero sequence-length growth and provides the best practical throughput. The gated variant does not improve over simple addition (0.056 vs. 0.057), suggesting that an additive bias is already sufficient to synchronize holistic audio semantics with the generator's global conditioning space.

TABLE III
FUSION STRATEGY ABLATION (GREATESTHITS). TOKEN CONCATENATION USES ONE APPENDED AUDIO TOKEN ($K = 1$).

Fusion	AIS↑	GFLOPs↓	Time (s)↓
Additive (Eq. (4))	0.057	2.89	1.41
Gated (Eq. (5))	0.056	2.90	1.44
Token concat (Eq. (6))	0.058	3.45	1.85

D. Inference-time Noise Robustness

To evaluate robustness without retraining, we inject Gaussian noise into the input spectrograms at varying signal-to-noise ratios (SNRs). Fig. 3 compares AudioControl with representative baselines under the same perturbation protocol. AudioControl degrades more gracefully as noise increases:

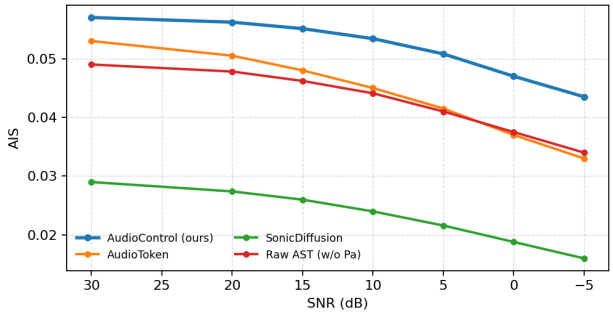

Fig. 3. Noise robustness analysis. AIS vs. SNR under spectrogram Gaussian noise perturbations.

whereas competing methods are more sensitive to local spectral corruption, our method maintains comparatively stable AIS even at lower SNRs. We attribute this robustness to the global aggregation behavior of latent modulation: the projection layer P_a and global pooling act as a semantic filter, extracting the dominant acoustic gist (e.g., core timbre and event category) while suppressing uncorrelated high-frequency noise. This is one of the clearest practical advantages of global conditioning over fine-grained token attention.

E. Representation Alignment and Limitations

a) Feature Space Visualization (t-SNE): We jointly embed projected audio features and the corresponding image features and visualize them with t-SNE in Fig. 4. The result shows category-consistent clusters with substantial overlap between audio and image points, indicating that the projection P_a narrows the audio-image modality gap in the pooled conditioning space. Small within-cluster offsets remain, reflecting unavoidable cross-modal variance, but they do not prevent coherent global semantic alignment.

b) Limitations and Failure Cases: Because AudioControl injects a single fused global condition via additive affine

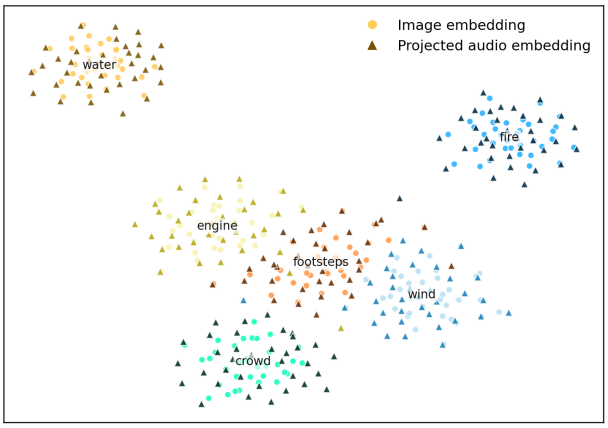

Fig. 4. t-SNE visualization of projected audio embeddings and image embeddings. Overlapping clusters suggest semantic alignment in the learned representation space.

modulation (Eq. (4)), it reliably captures scene-level attributes such as material, ambience, and event class, but it does not explicitly encode spatially localized cues such as interaural time or level differences. As illustrated in Fig. 5, directional audio can lead the model to predict the correct object category while placing it near the center rather than at the implied bearing. This trade-off—strong global semantics at the expense of spatial specificity—motivates future work on spatially aware conditioning mechanisms (e.g., conditioned token maps, spatial-audio encoders, or hybrid global+token schemes) that may recover localization without reintroducing quadratic attention costs. A formal human perceptual study of audio-visual faithfulness would also be a valuable direction for future evaluation.

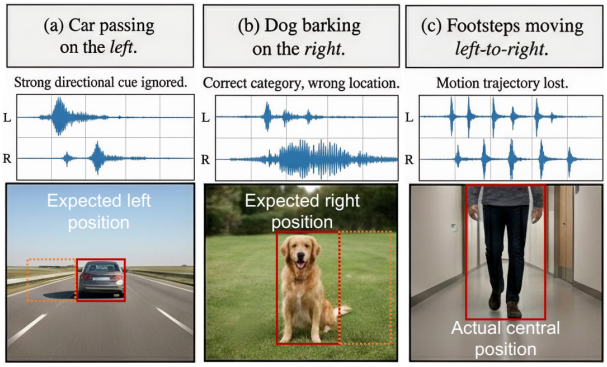

Fig. 5. Failure case illustration. Global modulation captures event/category semantics but may miss spatial localization cues from binaural audio.

VI. Conclusion

We presented AudioControl, which injects audio into flow-matching generators through additive global fusion. By avoiding sequence expansion, the method adds only negligible computational overhead while effectively capturing acoustic semantics such as material and ambience. Our results indicate that adapting global latent modulation for audio conditioning provides a strong efficiency-alignment trade-off relative to token-based baselines, making AudioControl a practical design choice for scalable multimodal generation.

References

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. "High-resolution image synthesis with latent diffusion models." in CVPR, pp. 10684-10695, 2022.
[2] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, D. Podell, T. Dockhorn, Z. English and R. Rombach. "Scaling rectified flow transformers for high-resolution image synthesis." in ICML. pp. 12606-12633, July, 2024.
[3] Y. Lipman, R. T. Chen, H. Ben-Hamu, M. Nickel, and M. Le. "Flow matching for generative modeling." arXiv:2210.02747, 2022.
[4] Black Forest Labs. "Flux." 2024.
[5] L. Zhang, A. Rao, and M. Agrawala. "Adding conditional control to text-to-image diffusion models." in ICCV, pp. 3836-3847, 2023.
[6] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang. "IP-Adapter: Text compatible image prompt adapter for text-to-image diffusion models." arXiv:2308.06721, 2023.
[7] V. Lialin, V. Deshpande, and A. Rumshisky. "Scaling down to scale up: A guide to parameter-efficient fine-tuning." arXiv preprint arXiv:2303.15647, 2023.
[8] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, and W. Chen. "LoRA: Low-rank adaptation of large language models." ICLR, 1(2), 3, 2022.
[9] H. Chung, J. Shim, and J. K. Kim. "Cross-Modal Contrastive Representation Learning for Audio-to-Image Generation." arXiv preprint arXiv:2207.12121, 2022.
[10] G. Yariv, I. Gat, L. Wolf, Y. Adi, and I. Schwartz. "AudioToken: Adaptation of text-conditioned diffusion models for audio-to-image generation." arXiv:2305.13050, 2023.
[11] B. C. Biner, F. M. Sofian, U. B. Karakaş, D. Ceylan, E. Erdem, and A. Erdem. "SonicDiffusion: Audio-driven image generation and editing with pretrained diffusion models." arXiv:2405.00878, 2024.
[12] D. Petermann and M. M. Kalayeh. "Seeing Sound: Assembling Sounds from Visuals for Audio-to-Image Generation." arXiv:2501.05413, 2025.
[13] Y. Gong, Y. A. Chung, and J. Glass. "AST: Audio spectrogram transformer." arXiv:2104.01778, 2021.
[14] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville. "FiLM: Visual Reasoning with a General Conditioning Layer." in AAAI, Vol. 32, No. 1, April, 2018.
[15] X. Huang and S. Belongie. "Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization." in ICCV, pp. 1501-1510, 2017.
[16] H. Oh, S. Cha, K. Lee, S. W. Kim, and D. J. Kim. "CatchPhrase: EXPrompt-Guided Encoder Adaptation for Audio-to-Image Generation." in ACM MM, pp. 9773-9782, October, 2025.
[17] K. Sung-Bin, A. Senocak, H. Ha, A. Owens, and T. H. Oh. "Sound to visual scene generation by audio-to-visual latent alignment." in CVPR, pp. 6430-6440, 2023.
[18] C. Qin, N. Yu, C. Xing, S. Zhang, Z. Chen, S. Ermon, Y. Fu, C. Xiong, and R Xu. "Gluegen: Plug and play multi-modal encoders for x-to-image generation." in ICCV, pp. 23085-23096, 2023.
[19] T. Karras, S. Laine, and T. Aila. "A Style-Based Generator Architecture for Generative Adversarial Networks." in CVPR, pp. 4401-4410, 2019.
[20] Y. Zhang, Y. Yuan, Y. Song, H. Wang, and J. Liu. "EasyControl: Adding efficient and flexible control for diffusion transformer." in ICCV, pp. 19513-19524, 2025.
[21] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. "Exploring the limits of transfer learning with a unified text-to-text transformer." JMLR, 21(140), pp. 1-67, 2020.
[22] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. "Learning transferable visual models from natural language supervision." in ICML, pp. 8748-8763, July, 2021.
[23] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. "GANs trained by a two time-scale update rule converge to a local Nash equilibrium." in NeurIPS, 30, 2017.
[24] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman. "Vggsound: A large-scale audio-visual dataset." in ICASSP 2020, pp. 721-725, May, 2020.
[25] Peebles W, Xie S. Scalable diffusion models with transformers. Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195-4205, 2023.