POLARIZATION-AWARE EXPERTS FOR UNIFIED VIDEO-TEXT-TO-AUDIO GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Unified video-text-to-audio generation supports Text-to-Audio (T2A), Video-to-Audio (V2A), and joint Video-Text-to-Audio (VT2A) within a single model. However, joint optimization introduces cross-task competition due to their heterogeneous conditioning requirements, as well as intra-task competition in VT2A, where excessive reliance on one modality can weaken alignment with the other. Existing systems primarily rely on shared dense backbones and progressive training, without explicitly preserving task-associated computation or directly regulating condition dominance in VT2A samples. We propose a two-stage Mixture-of-Experts framework that independently specializes T2A- and V2A-associated expert pools before composing them for unified multi-task training. Task-conditioned candidate restriction separates the computation paths of the single-condition tasks while allowing VT2A to access both pools. We further characterize excessive routing preference for one pool as expert-pool polarization and introduce a training-only intervention that reserves one routing slot for the under-selected pool while leaving the other unrestricted. Experiments demonstrate competitive T2A performance and strong V2A and VT2A results. Ablations validate the proposed specialization and candidate restriction, while the intervention improves targeted VT2A cases and mitigates inference-time expert-pool polarization. Our code will be publicly released.

1 INTRODUCTION

Audio generation has become an essential component of multimedia creation, enabling immersive experiences across films, video games, and user-generated content. Early methods predominantly addressed single-condition generation using either text or video. Text-to-Audio (T2A) models Liu et al. (2023); Kreuk et al. (2023); Ghosal et al. (2023); Majumder et al. (2024); Huang et al. (2023); Hung et al. (2026); Evans et al. (2025) achieve semantic alignment with textual descriptions but lack temporal grounding, whereas Video-to-Audio (V2A) Luo et al. (2023); Wang et al. (2024a); Xing et al. (2024); Wang et al. (2024b); Zhang et al. (2025) models support synchronization with visible events but may struggle with ambiguous scenes and off-screen sounds Kushwaha & Tian (2025). The complementary roles of text and video have motivated Video-Text-to-Audio (VT2A) methods that jointly condition on both modalities for semantic fidelity and temporal alignment, but their reliance on both modalities may degrade performance when either is absent, limiting their applicability to single-condition scenarios Dai et al. (2026); Kushwaha & Tian (2025); Liu et al. (2025). This motivates a unified VT2A framework that supports T2A, V2A, and VT2A within a single model while preserving performance across heterogeneous conditioning settings.

However, integrating these tasks introduces two major challenges: cross-task and intra-task competition Dai et al. (2026). Cross-task competition occurs when T2A and V2A, despite their distinct conditioning requirements, compete for shared model capacity during joint optimization, potentially inducing a performance trade-off. Intra-task competition arises in VT2A when excessive dominance of one condition suppresses complementary information from the other Wei et al. (2024). Although the relative utility of text and video varies across samples, text dominance may weaken audio–visual synchronization, whereas video dominance may suppress text-specified content, particularly off-screen sounds. Therefore, a unified model should preserve task-associated T2A and V2A knowledge while mitigating excessive dominance without enforcing uniform condition usage across VT2A samples.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Recent studies primarily address these challenges through multimodal fusion, data scaling, and progressive training. MMAudio Cheng et al. (2025a) jointly learns text- and video-conditioned generation with conditional synchronization, AudioX Tian et al. (2026) supports broader multimodal conditions, and Omni2Sound Dai et al. (2026) uses progressive training to address cross-task and intra-task competition. Despite these advances, existing systems predominantly use shared dense backbones without separate T2A- and V2A-specialized pathways or a mechanism for adaptively composing them under joint conditioning, while relying on data or training strategies rather than directly regulating sample-level condition dominance. These limitations motivate a unified model that preserves task-associated computation pathways and selectively mitigates excessive modality dominance.

To this end, we propose a two-stage Mixture-of-Experts (MoE) framework that explicitly addresses cross-task and intra-task competition in unified VT2A generation. In the first stage, we replace selected dense layers in a pre-trained audio generation backbone based on a Diffusion Transformer (DiT) Peebles & Xie (2023) with MoE blocks and independently specialize separate expert pools for T2A and V2A. In the second stage, we compose the specialized expert pools into a unified model and jointly optimize it across all three tasks. Task-conditioned candidate restriction preserves separate routed computation paths for T2A and V2A, reducing direct competition over experts, while allowing VT2A to access both pools. Within the resulting VT2A routing space, intra-task competition can manifest as expert-pool polarization, where excessive preference for one pool limits routing opportunities for the other. Conventional load-balancing objectives regulate aggregate expert utilization but do not directly address within-sample polarization, whereas uniform cross-pool allocation would suppress legitimate sample-dependent preferences.

We therefore introduce a *Polarization-Aware Cross-Pool Routing Intervention*. For detected polarized VT2A sample-layers, the training-only intervention stochastically reserves one top-2 routing slot for the under-selected pool while leaving the other slot unrestricted. This design provides the under-selected pool with a direct routing and optimization opportunity, discouraging excessive expert-pool polarization without enforcing uniform pool usage. At inference, the intervention is removed to retain unrestricted top-2 routing and sample-adaptive expert selection.

Experiments show that our framework maintains competitive T2A performance while achieving strong V2A and VT2A results within a single unified model. Structural ablations demonstrate the complementary benefits of independent task specialization and task-conditioned candidate restriction, while the polarization-aware intervention improves polarized and off-screen VT2A cases and reduces expert-pool polarization at inference. These results support the effectiveness of structuring task-associated expert pathways in mitigating direct cross-task competition over routed experts and intra-task expert-pool polarization in VT2A. Our main contributions are summarized as follows:

- We propose a two-stage MoE framework that independently specializes T2A- and V2A-associated expert pools before composing them for unified T2A, V2A, and VT2A generation. Task-conditioned candidate restriction preserves separate routed computation paths to mitigate direct cross-task competition.
- We characterize excessive expert-pool polarization as a routing-level form of intra-task competition and meaningful criterion for identifying VT2A cases that benefit from cross-pool intervention. We introduce a training-only intervention that selectively promotes the under-selected pool while retaining unrestricted inference.
- Experiments demonstrate strong performance across all three tasks and validate the effectiveness of task-associated specialization, candidate restriction, and polarization-aware routing.

2 Related Work

2.1 Unified Video-Text-to-Audio Generation

Recent unified systems support T2A, V2A, and VT2A within a single model Cheng et al. (2025a); Tian et al. (2026); Dai et al. (2026). MMAudio Cheng et al. (2025a) jointly trains on video–audio and text–audio data with a conditional synchronization module, while AudioX Tian et al. (2026) supports diverse multimodal conditions through large-scale training and adaptive condition fusion.

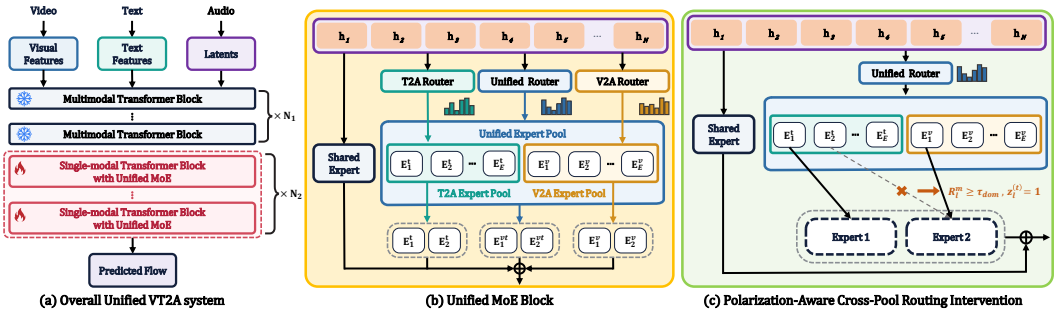

Figure 1: Overview of the proposed Unified MoE framework. (a) Overall architecture for unified T2A, V2A, and VT2A generation. (b) Independently specialized task-associated expert pools are composed through task-conditioned routing in each unified MoE block. (c) For a polarized VT2A sample-layer, the training-only intervention reserves one routing slot for the under-selected pool while retaining one unrestricted slot.

Omni2Sound Dai et al. (2026) further addresses cross-task and intra-task competition through tightly aligned video–audio–text data and progressive multi-task training.

Despite using different training and condition-fusion strategies, these approaches largely share the same generation backbone across tasks without explicitly separating task-associated computation. Consequently, modality competition is primarily addressed through data construction, condition fusion, or training schedules rather than task-structured conditional computation. These limitations motivate a conditional computation framework that can preserve task-associated specialization while adapting its computation across T2A, V2A, and VT2A.

2.2 Mixture-of-Experts for Generative Models

MoE enables conditional computation by replacing a dense Feed-Forward Network (FFN) in a transformer block with multiple expert networks and a router that activates a sparse subset of experts for each token. This design increases model capacity without proportionally increasing active computation and enables input-dependent specialization. Because sparse routing can result in highly imbalanced expert utilization, conventional MoE architectures commonly employ auxiliary load-balancing objectives to distribute routing assignments more evenly across experts.

Recent studies have incorporated MoE into DiTs for image generation Fei et al. (2024); Cheng et al. (2025b); Sun et al. (2025); Yuan et al. (2025). DiT-MoE Fei et al. (2024) studies expert specialization across spatial positions and diffusion timesteps, while Diff-MoE Cheng et al. (2025b) introduces time-aware and space-adaptive experts. EC-DiT Sun et al. (2025) and Race-DiT Yuan et al. (2025) further allocate computation according to token- or generation-level complexity. These methods primarily employ MoE for capacity scaling and adaptive computation across tokens, spatial regions, or diffusion stages, whereas task-associated specialization for unified audio generation remains underexplored. Our work extends this direction to unified audio generation by composing independently specialized T2A and V2A expert pools and regulating sample-layer expert-pool polarization during joint VT2A training.

3 Methods

3.1 Stage 1: Task-Specific MoE Specialization

We build upon a pretrained MMAudio backbone that supports audio generation under text-only, video-only, and joint video–text conditioning. Within selected transformer blocks, we replace the dense FFNs with MoE blocks comprising the original FFN as a shared expert, separate T2A- and V2A-associated expert pools, and task-specific routers that select the top-2 experts for each hidden token. Beyond adding task-specific capacity, this structure provides a shared routing interface through which separately learned task-associated transformations can be selectively composed for VT2A.

To learn task-dependent transformations, we independently specialize the two expert pools on their corresponding T2A and V2A tasks. For each task, only the associated experts and router are optimized, while the shared expert and all other pretrained backbone parameters remain frozen. The resulting T2A- and V2A-associated expert pools are subsequently composed in Stage 2 for unified multi-task training.

3.2 Stage 2: Unified Expert Composition and Multi-Task Training

After task-specific specialization, we integrate the T2A- and V2A-associated expert pools obtained in Stage 1 to construct the final unified model. For each MoE block at layer l, let $\mathcal{E}_l^{\mathrm{T}}$ and $\mathcal{E}_l^{\mathrm{V}}$ denote the specialized T2A- and V2A-associated expert pools, respectively. We concatenate them into the ordered unified expert pool $\mathcal{E}_l^{\mathrm{U}} = \mathcal{E}_l^{\mathrm{T}} \parallel \mathcal{E}_l^{\mathrm{V}}$, where $\parallel$ denotes ordered concatenation along the expert dimension. While the pretrained shared expert remains common to all tasks, a new unified router r_l^{U} is initialized for each MoE block and jointly optimized across T2A, V2A, and VT2A.

The task label determines the candidate expert set available to each input. For a task $a \in \{\mathrm{T2A}, \mathrm{V2A}, \mathrm{VT2A}\}$, we define

$$\mathcal{C}_l^a = \begin{cases} \mathcal{E}_l^{\mathrm{T}}, & a = \mathrm{T2A}, \\ \mathcal{E}_l^{\mathrm{V}}, & a = \mathrm{V2A}, \\ \mathcal{E}_l^{\mathrm{U}}, & a = \mathrm{VT2A}. \end{cases}$$

Thus, each single-condition task is restricted to its associated pool, whereas VT2A routes over the complete composed pool.

Given a hidden token $\mathbf{h}_l^i$, the unified router produces logits over the complete expert pool as $\mathbf{s}_{l,i}^{\mathrm{U}} = r_l^{\mathrm{U}}(\mathbf{h}_l^i)$. Applying the task-dependent candidate set gives the pre-selection routing distribution

$$\pi_{l,i,e}^a = \begin{cases} \dfrac{\exp(s_{l,i,e}^{\mathrm{U}})}{\sum_{e' \in \mathcal{C}_l^a} \exp(s_{l,i,e'}^{\mathrm{U}})}, & e \in \mathcal{C}_l^a, \\ 0, & e \notin \mathcal{C}_l^a. \end{cases}$$

The router selects the top-2 candidate experts and renormalizes their probabilities:

$$\mathcal{S}_{l,i}^a = \mathrm{TopK}\big(\boldsymbol{\pi}_{l,i}^a, 2\big)\,,\ p_{l,i,e}^a = \frac{\pi_{l,i,e}^a}{\sum_{e' \in \mathcal{S}_{l,i}^a} \pi_{l,i,e'}^a}, \quad e \in \mathcal{S}_{l,i}^a.$$

The resulting task-conditioned MoE output is

$$\mathrm{MoE}_{\mathrm{U}}^a(\mathbf{h}_l^i) = f_{l,\mathrm{shared}}(\mathbf{h}_l^i) + \sum_{e \in \mathcal{S}_{l,i}^a} p_{l,i,e}^a f_{l,e}^{\mathrm{U}}(\mathbf{h}_l^i),$$

where $f_{l,e}^{\mathrm{U}}$ denotes the e-th routed expert in $\mathcal{E}_l^{\mathrm{U}}$.

The composed model is jointly optimized on T2A, V2A, and VT2A. The candidate restriction preserves the task-associated routed pathways for T2A and V2A, while VT2A retains access to both specialized pools through unrestricted top-2 selection over $\mathcal{E}_l^{\mathrm{U}}$. By separating the routed candidate spaces of the two single-condition tasks while enabling their flexible composition under joint conditioning, this design provides an architectural basis for alleviating direct cross-task competition over routed experts.

3.3 Polarization-Aware Cross-Pool Routing Intervention

For VT2A samples, the unified router performs top-2 selection over the composed T2A- and V2A-associated expert pools. Because the relative routing preference between the two pools may legitimately vary across samples, an asymmetric routing preference does not by itself indicate a routing failure. The problem arises when this preference becomes excessive, causing one task-associated pool to dominate VT2A routing and limit routing opportunities for the other. An aggregate expert load-balancing objective does not directly regulate such sample-level pool polarization, whereas enforcing a uniform pool allocation would constrain sample-dependent routing behavior.

To regulate excessive pool dominance while preserving legitimate sample-dependent preferences, we introduce a *Polarization-Aware Cross-Pool Routing Intervention* during Stage 2. At a selected MoE layer, the intervention identifies the under-selected pool for each VT2A sample from its sample-level pre-selection routing distribution. To avoid constraining the unified router during its initial adaptation, we disable the intervention for an initial warm-up period and then progressively increase the intervention probability $q_{\text{int}}^{(t)}$ to its target value. When a polarized sample-layer is stochastically activated, one of the two routing slots is reserved for an expert from the under-selected pool, while the other remains unrestricted. Non-polarized and non-activated sample-layers retain the original top-2 routing.

Specifically, using the pre-selection distribution $\boldsymbol{\pi}_{l,i}^{\text{VT2A}}$ defined in Stage 2, we compute the sample-level routing mass assigned to each task-associated pool at layer l:

$$R_l^m = \frac{1}{N_l} \sum_{i=1}^{N_l} \sum_{e \in \mathcal{E}_l^m} \pi_{l,i,e}^{\text{VT2A}}, \qquad m \in \{\text{T}, \text{V}\},$$

where N_l is the number of hidden tokens at layer l. Since VT2A inputs use the complete unified expert pool and T2A- and V2A-associated pools partition the unified routed expert pool, the pool-level routing masses satisfy $R_l^{\text{T}} + R_l^{\text{V}} = 1$.

From these masses, the under-selected pool and its expert set are

$$m_l^- = \arg\min_{m \in \{\text{T},\text{V}\}} R_l^m, \qquad \mathcal{E}_l^- = \mathcal{E}_l^{m_l^-}.$$

A sample is considered polarized at layer l when $\max_{m \in \{\text{T},\text{V}\}} R_l^m \geq \tau_{\text{dom}}$, where $\tau_{\text{dom}} \in (0.5, 1]$ is the dominance threshold. For a polarized sample-layer, the intervention indicator is sampled as $z_l^{(t)} \sim \text{Bernoulli}\left(q_{\text{int}}^{(t)}\right)$, where $q_{\text{int}}^{(t)}$ denotes the scheduled intervention probability at training step t.

For each token, the reserved slot selects the highest-probability expert from the under-selected pool, while the unrestricted slot selects the highest-probability expert from the unified pool excluding the reserved expert:

$$e_{l,i}^{\text{res}} = \text{Top1}\left(\boldsymbol{\pi}_{l,i}^{\text{VT2A}}; \mathcal{E}_l^-\right),$$
$$e_{l,i}^{\text{free}} = \text{Top1}\left(\boldsymbol{\pi}_{l,i}^{\text{VT2A}}; \mathcal{E}_l^{\text{U}} \setminus \{e_{l,i}^{\text{res}}\}\right).$$

Here, $\text{Top1}(\boldsymbol{\pi}; \mathcal{E})$ returns the expert in candidate set $\mathcal{E}$ with the highest probability under $\boldsymbol{\pi}$. Thus, the unrestricted slot may still select an expert from either pool.

Let $\mathcal{S}_{l,i}^{\text{VT2A}}$ denote the original unrestricted top-2 expert set defined in Stage 2. The final expert set and its renormalized routing weights are

$$\widehat{\mathcal{S}}_{l,i}^{\text{VT2A}} = \begin{cases} \{e_{l,i}^{\text{res}}, e_{l,i}^{\text{free}}\}, & \max_{m \in \{\text{T},\text{V}\}} R_l^m \geq \tau_{\text{dom}} \ \wedge \ z_l^{(t)} = 1, \\ \mathcal{S}_{l,i}^{\text{VT2A}}, & \text{otherwise}, \end{cases}$$

$$\widehat{p}_{l,i,e}^{\text{VT2A}} = \frac{\pi_{l,i,e}^{\text{VT2A}}}{\sum_{e' \in \widehat{\mathcal{S}}_{l,i}^{\text{VT2A}}} \pi_{l,i,e'}^{\text{VT2A}}}, \qquad e \in \widehat{\mathcal{S}}_{l,i}^{\text{VT2A}}.$$

The resulting expert set and routing weights replace $\mathcal{S}_{l,i}^{\text{VT2A}}$ and $p_{l,i,e}^{\text{VT2A}}$, respectively, in the Stage 2 MoE computation. By reserving one slot only for stochastically activated polarized sample-layers while leaving the other unrestricted, the intervention provides the under-selected pool with a direct routing opportunity without enforcing a uniform pool allocation across all samples. The intervention is used only during Stage 2 training, while inference retains the learned unrestricted top-2 routing.

3.4 Training Objectives

Task-Conditioned Load Balancing. Sparse MoE models commonly employ an auxiliary load-balancing objective to discourage routing collapse. In Stage 2, however, T2A, V2A, and VT2A samples route over different candidate pools. Aggregating routing statistics over the entire mixed-task

Dataset	Method	Task	Unified	KL ↓	FD ↓	IS ↑	IB-Score ↑	CLAP ↑	DeSync ↓
AudioCaps	StableAudio Evans et al. (2025)	T2A	×	2.461	21.647	**5.094**	-	0.141	-
	TangoFlux Hung et al. (2026)	T2A	×	**1.907**	14.691	4.503	-	**0.320**	-
	MMAudio Cheng et al. (2025a)	T2A	✓	2.269	24.450	4.640	-	0.267	-
	AudioX Tian et al. (2026)	T2A	✓	2.006	15.615	4.993	-	0.281	-
	Omni2Sound Dai et al. (2026)	T2A	✓	3.241	**11.424**	4.033	-	0.246	-
	Ours	T2A	✓	1.974	15.481	4.647	-	**0.320**	-
VGGSound	Diff-Foley Luo et al. (2023)	V2A	×	2.491	29.282	2.796	0.182	-	1.015
	FoleyCrafter Zhang et al. (2025)	V2A	×	2.519	19.347	4.374	0.272	-	1.232
	MMAudio Cheng et al. (2025a)	V2A	✓	2.102	8.681	5.359	0.326	-	0.434
	AudioX Tian et al. (2026)	V2A	✓	2.395	14.514	4.952	0.244	-	0.985
	Omni2Sound Dai et al. (2026)	V2A	✓	2.069	9.762	5.035	**0.333**	-	0.513
	Ours	V2A	✓	**2.066**	**6.774**	**5.505**	0.319	-	**0.424**
	MMAudio Cheng et al. (2025a)	VT2A	✓	1.928	8.502	5.572	**0.329**	0.217	0.441
	AudioX Tian et al. (2026)	VT2A	✓	2.164	12.901	5.341	0.273	0.186	0.985
	Omni2Sound Dai et al. (2026)	VT2A	✓	1.997	10.721	5.269	0.325	0.199	0.505
	Ours	VT2A	✓	**1.853**	**6.910**	**5.600**	0.325	**0.218**	**0.435**

Table 1: Objective performance comparison across T2A, V2A, and VT2A tasks. A check mark (✓) in the "Unified" column indicates that a single model checkpoint supports all three tasks. The best and second-best results are shown in bold and underlined, respectively.

batch can therefore obscure within-task under-utilization. We instead compute the load-balancing objective independently for each task over its corresponding candidate expert pool.

Let $\mathcal{A}_\mathcal{B} \subseteq \{\mathrm{T2A}, \mathrm{V2A}, \mathrm{VT2A}\}$ denote the tasks represented in batch $\mathcal{B}$, and let $\mathcal{I}_l^a$ denote the tokens belonging to task a at MoE layer l. For each expert $e \in \mathcal{C}_l^a$, its mean pre-selection probability and selected-slot frequency are

$$P_{l,e}^a = \frac{1}{|\mathcal{I}_l^a|} \sum_{i \in \mathcal{I}_l^a} \pi_{l,i,e}^a, \qquad F_{l,e}^a = \frac{1}{2|\mathcal{I}_l^a|} \sum_{i \in \mathcal{I}_l^a} \mathbf{1}\left[e \in \mathcal{S}_{l,i}^a\right],$$

where the factor of 2 accounts for top-2 routing. The task-conditioned load-balancing objective at layer l is

$$\mathcal{L}_{\mathrm{bal},l} = \frac{1}{|\mathcal{A}_\mathcal{B}|} \sum_{a \in \mathcal{A}_\mathcal{B}} |\mathcal{C}_l^a| \sum_{e \in \mathcal{C}_l^a} F_{l,e}^a P_{l,e}^a.$$

This objective balances aggregate expert utilization within each task-specific candidate pool but does not enforce uniform allocation between the T2A- and V2A-associated pools for individual samples. We average it over the MoE layers as $\mathcal{L}_{\mathrm{bal}} = |\mathcal{M}|^{-1} \sum_{l \in \mathcal{M}} \mathcal{L}_{\mathrm{bal},l}$, where $\mathcal{M}$ denotes the set of MoE layer indices.

Expert-to-Shared Lower-Bound Objective. We additionally observed that selected-expert outputs can become negligible relative to the pretrained shared expert output. We therefore introduce a one-sided objective that penalizes a selected expert only when its output norm falls below a prescribed fraction of the corresponding shared-expert output norm.

Let $\overline{\mathcal{S}}_{l,b,i}$ denote the final expert set used for token i of sample b at layer l, after any intervention. For each selected expert $e \in \overline{\mathcal{S}}_{l,b,i}$, we define

$$r_{l,b,i,e} = \frac{\|f_{l,e}(\mathbf{h}_{l,b}^i)\|_2 + \epsilon}{\mathrm{sg}\Big(\|f_{l,\mathrm{shared}}(\mathbf{h}_{l,b}^i)\|_2\Big) + \epsilon}, \qquad \mathcal{L}_{\mathrm{lb},l} = \frac{1}{2BN_l} \sum_{b=1}^{B} \sum_{i=1}^{N_l} \sum_{e \in \overline{\mathcal{S}}_{l,b,i}} \left[\log \frac{\rho}{r_{l,b,i,e}}\right]_+^2,$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient, ϵ ensures numerical stability, ρ is the target lower bound, and $[x]_+ = \max(x, 0)$. The normalization uses $|\overline{\mathcal{S}}_{l,b,i}| = 2$ under top-2 routing. We average the objective over its applied layers $\mathcal{M}_{\mathrm{lb}}$ as $\mathcal{L}_{\mathrm{lb}} = |\mathcal{M}_{\mathrm{lb}}|^{-1} \sum_{l \in \mathcal{M}_{\mathrm{lb}}} \mathcal{L}_{\mathrm{lb},l}$.

It constrains only relative output magnitudes and neither balances experts against one another nor aligns their representations with the shared expert. Finally, we retain the original DiT training objective of MMAudio as the generation objective $\mathcal{L}_{\mathrm{gen}}$. The overall objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{gen}} + \lambda_{\mathrm{bal}} \mathcal{L}_{\mathrm{bal}} + \lambda_{\mathrm{lb}} \mathcal{L}_{\mathrm{lb}},$$

where λ_{bal} and λ_{lb} weight the load-balancing and expert-to-shared lower-bound terms, respectively.

Setting	T2A		V2A			VT2A		
	FD ↓	CLAP ↑	FD ↓	IB ↑	DeSync ↓	FD ↓	CLAP ↑	DeSync ↓
Single-stage training	15.488	0.413	6.938	0.316	0.444	**6.842**	0.324	0.448
w/o Candidate Restriction	15.818	0.414	7.061	0.317	0.445	7.593	0.321	0.447
Full Method	**15.481**	**0.416**	**6.774**	**0.319**	**0.424**	6.910	**0.325**	**0.435**

Table 2: Ablation of the two-stage training and task-conditioned candidate restriction. Single-stage training omits task-specific specialization, while removing candidate restriction allows all tasks to access the unified expert pool. All other settings are kept unchanged. Best results are shown in bold.

Setting	Full Test Set				Polarized Subset			
	FD ↓	IB ↑	CLAP ↑	DeSync ↓	FD ↓	IB ↑	CLAP ↑	DeSync ↓
Ours w/o intervention	6.964	0.325	0.321	0.438	21.701	0.318	0.301	0.422
Ours w/ intervention	**6.910**	0.325	**0.325**	**0.435**	**20.752**	**0.326**	**0.311**	**0.412**

Table 3: Effect of the polarization-aware intervention on the full VT2A test set and a fixed polarized subset identified by the no-intervention model. CLAP denotes MS-CLAP. Best results are bolded.

4 Experiments

4.1 Experimental Settings

4.1.1 Datasets

For T2A training, we use the training splits of AudioCaps Kim et al. (2019), Clotho Drossos et al. (2020), and WavCaps Mei et al. (2024). For video-conditioned training, we use audio–visual samples from VGGSound Chen et al. (2020) and a subset of IF-Caps Tian et al. (2026). Following MMAudio Cheng et al. (2025a), unavailable task conditions are represented by learnable empty tokens; further dataset details are provided in Appendix A.

4.1.2 Objective Evaluation Metrics

Following MMAudio Cheng et al. (2025a), we evaluate generated audio using AV-Benchmark. We report the Fréchet Distance (FD), Kullback–Leibler (KL) divergence, and Inception Score (IS), computed using PANNs Kong et al. (2020). Text–audio and video–audio semantic alignment are evaluated using CLAP Wu et al. (2023) and the ImageBind score (IB-Score) Girdhar et al. (2023), respectively. We measure audio–visual temporal alignment using the desynchronization score (DeSync) predicted by Synchformer Iashin et al. (2024), where lower is better. Detailed definitions are provided in Appendix B.

4.1.3 Subjective Evaluation Metrics

We conduct a human evaluation across T2A, V2A, and VT2A for acoustic quality (MOS-Q), semantic consistency (MOS-S), and audio–visual temporal synchronization (MOS-T). The complete protocol is provided in Appendix C.

4.1.4 Implementation Details

We replace the transformer blocks at zero-based indices $\{0, 4, 8, 12\}$ with MoE blocks and apply the polarization-aware intervention only at index 8, with $\tau_{\text{dom}} = 0.75$ and $q_{\text{target}} = 0.25$. We set the expert-to-shared lower-bound target to $\rho = 0.15$. Additional implementation details are provided in Appendix D. Further ablations and analyses of MoE block-placement, intervention sensitivity, lower-bound objective, and computational costs are reported in Appendices E, F, G, and H, respectively.

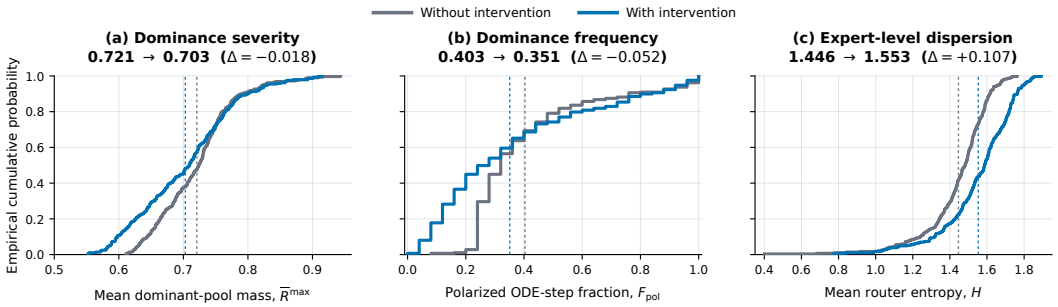

Figure 2: Inference-time routing distributions on the fixed polarized subset: (a) mean expert-pool dominance, (b) dominance frequency across ODE steps, and (c) expert-level routing entropy. Both models use unrestricted top-2 routing at inference.

Method	MOS-Q ↑	MOS-S ↑	MOS-T ↑	Overall ↑
MMAudio	3.47	3.33	3.49	3.43
AudioX	2.68	2.60	2.09	2.46
Omni2Sound	3.50	3.59	3.65	3.58
Ours	**3.79**	**3.86**	**3.79**	**3.82**

(a) Subjective evaluation

Method	Int.	FD ↓	IB ↑	CLAP ↑
Ours	×	10.463	0.336	0.462
Ours	✓	**10.075**	**0.344**	**0.467**

(b) Off-screen VT2A ablation

Table 4: Subjective and off-screen VT2A evaluation. (a) Mean opinion scores for acoustic quality (MOS-Q), semantic consistency (MOS-S), and temporal synchronization (MOS-T). (b) Performance on the off-screen VT2A subset with and without the training-time polarization intervention. CLAP denotes MS-CLAP. Best results are bolded.

4.2 Main Results

4.2.1 Objective Evaluation

Table 1 compares all methods on AudioCaps and VGGSound under the same evaluation protocol and task-specific conditioning inputs, using samples reproduced from their official implementations. On AudioCaps, our model achieves competitive T2A performance against task-specific models and outperforms existing unified models in KL and CLAP, indicating strong distributional correspondence and text–audio alignment. For V2A, it achieves the best KL, FD, IS, and DeSync, reflecting strong distribution-level performance and audio–visual temporal alignment. It also outperforms the MMAudio backbone on most T2A and V2A metrics. Together with the structural ablations in Table 2, these results support the role of task-associated expert pools in reducing direct cross-task competition over routed experts. For VT2A, our model achieves the best results in five of the six reported metrics, including the highest CLAP and lowest DeSync, indicating strong text–audio semantic and audio–visual temporal alignment under joint conditioning. These gains are obtained with only a 7.7% increase in parameters over the MMAudio backbone (Appendix H).

4.2.2 Subjective Evaluation

We conducted a user study with 12 evaluators on 30 samples. As shown in Table 4(a), our model achieves the highest MOS-Q, MOS-S, MOS-T, and overall scores, demonstrating strong perceptual quality, conditioning consistency, and temporal synchronization. Detailed task-wise results are provided in Table 5.

4.3 Ablation Study

4.3.1 Polarization-Aware Routing

We evaluate the polarization-aware routing intervention on the full VT2A test set and a fixed polarized subset. For a paired comparison, we define the subset once using the inference-time routing

distributions of the model trained without the intervention and evaluate both variants on the same samples.

As shown in Table 3, the intervention preserves or slightly improves performance across all metrics on the full test set. The gains are more pronounced on the fixed polarized subset, where the intervention consistently improves FD, IB-Score, CLAP, and DeSync. In particular, FD decreases from 21.701 to 20.752, while CLAP increases from 0.301 to 0.311. The concurrent improvements in IB-Score and DeSync further reflect stronger video–audio semantic correspondence and temporal alignment, respectively. The larger and consistent gains on the fixed polarized subset, compared with the modest changes on the full test set, indicate that the intervention primarily benefits the cases identified by the polarization criterion, supporting more effective joint conditioning on text and video while preserving full-set performance.

Inference-time routing behavior. We next examine whether the performance gains are accompanied by changes in the learned inference-time routing distributions. Figure 2 compares both variants on the same fixed polarized subset. Both models use unrestricted top-2 routing at inference; therefore, the observed differences reflect routing behavior learned during training rather than an inference-time constraint.

Training with the intervention reduces the mean dominant-pool mass from 0.721 to 0.703 and the polarized ODE-step fraction from 0.403 to 0.351. These changes indicate reductions in both the average severity of expert-pool dominance and its occurrence frequency over the denoising trajectory. Meanwhile, the mean router entropy increases from 1.446 to 1.553, reflecting a less concentrated expert-level routing distribution. Combined with the larger and consistent performance gains on the fixed polarized subset in Table 3, these routing changes support excessive polarization as a meaningful criterion for identifying cases that benefit from cross-pool intervention, while retaining unrestricted routing at inference.

Off-screen evaluation. To evaluate the intervention when text provides complementary information beyond the visual input, we use the off-screen VT2A test set Dai et al. (2026), where the target sound source is not directly visible in the video. As shown in Table 4(b), the intervention consistently improves all reported metrics. Specifically, FD decreases from 10.463 to 10.075, while IB-Score and CLAP increase from 0.336 to 0.344 and from 0.462 to 0.467, respectively. The concurrent improvements in IB-Score and CLAP show stronger alignment of the generated audio with both the video context and the text-specified off-screen sound. These results complement the fixed-subset analysis, showing that the intervention benefits off-screen cases susceptible to intra-task competition, where text provides cues absent from the video.

4.3.2 Two-Stage Specialization and Candidate Restriction

We compare the full method with single-stage training without prior T2A/V2A specialization and a global-candidate variant without task-conditioned candidate restriction, keeping all other settings unchanged.

As shown in Table 2, single-stage training improves VT2A FD, whereas the full method performs better on most other metrics. This suggests that Stage 1 provides a task-associated initialization rather than a prerequisite for specialization. The consistent degradation without candidate restriction supports its role in separating routed computation paths and thereby mitigating direct cross-task competition over experts.

5 Conclusion

We presented a unified MoE framework that independently specializes T2A- and V2A-associated expert pools before composing them for unified multi-task generation. The resulting task-associated pathways reduce direct cross-task competition over routed experts and enable the identification of expert-pool polarization in VT2A. To address this polarization, we introduced a training-only routing intervention that selectively promotes the under-selected pool while retaining unrestricted inference-time routing. Experiments demonstrate strong performance across all three tasks, while the intervention improves targeted VT2A cases and mitigates inference-time expert-pool polarization.

AI USE STATEMENT

Generative AI tools were used to assist with language editing, clarity improvement, and manuscript compression. All AI-assisted text was reviewed and verified by the authors. The authors remain responsible for the final manuscript.

REPRODUCIBILITY STATEMENT

We provide the model architecture, expert configuration, two-stage training procedure, optimization hyperparameters, routing intervention schedule, auxiliary-objective settings, and random seed in Appendix D. The training datasets, splits, and task-specific construction of T2A, V2A, and VT2A examples are described in Appendix A, while the evaluation metrics and protocols are specified in Appendix B. Additional ablations and analyses of MoE block placement, intervention sensitivity, the lower-bound objective, and computational costs are provided in Appendices E, F, G, and H, respectively.

REFERENCES

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)*, pp. 721–725, 2020. doi: 10.1109/ICASSP40776.2020.9053174.

Ho Kei Cheng, Masato Ishii, Akio Hayakawa, Takashi Shibuya, Alexander Schwing, and Yuki Mitsufuji. MMAudio: Taming Multimodal Joint Training for High-Quality Video-to-Audio Synthesis . In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28901–28911, Los Alamitos, CA, USA, June 2025a. IEEE Computer Society. doi: 10.1109/CVPR52734.2025.02691. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52734.2025.02691.

Kun Cheng, Xiao He, Lei Yu, Zhijun Tu, Mingrui Zhu, Nannan Wang, Xinbo Gao, and Jie Hu. Diff-MoE: Diffusion transformer with time-aware and space-adaptive experts. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 10010–10024. PMLR, 13–19 Jul 2025b. URL https://proceedings.mlr.press/v267/cheng25d.html.

Yusheng Dai, Zehua Chen, Yuxuan Jiang, Qiuhong Ke, Jianfei Cai, and Jun Zhu. Omni2sound: Towards unified video-text-to-audio generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1661–1671, 2026.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: an audio captioning dataset. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 736–740, 2020. doi: 10.1109/ICASSP40776.2020.9052990.

Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10095889.

Zach Evans, Julian D. Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10888461.

Zhengcong Fei, Mingyuan Fan, Changqian Yu, Debang Li, and Junshi Huang. Scaling diffusion transformers to 16 billion parameters, 2024. URL https://arxiv.org/abs/2407.11633.

Deepanway Ghosal, Navonil Majumder, Ambuj Mehrish, and Soujanya Poria. Text-to-audio generation using instruction guided latent diffusion model. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, pp. 3590–3598, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701085. doi: 10.1145/3581783.3612348. URL https://doi.org/10.1145/3581783.3612348.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15180–15190, 2023. doi: 10.1109/CVPR52729.2023.01457.

Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 13916–13932. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/huang23i.html.

Chia-Yu Hung, Navonil Majumder, Zhifeng Kong, Ambuj Mehrish, Amir Zadeh, Chuan Li, Rafael Valle, Bryan Catanzaro, and Soujanya Poria. Tangoflux: Super fast and faithful text to audio generation with flow matching and clap-ranked preference optimization. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=qgNs5NmQB7.

Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5325–5329, 2024. doi: 10.1109/ICASSP48485.2024.10448489.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating captions for audios in the wild. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 119–132, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1011. URL https://aclanthology.org/N19-1011/.

Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D. Plumbley. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 28:2880–2894, November 2020. ISSN 2329-9290. doi: 10.1109/TASLP.2020.3030497. URL https://doi.org/10.1109/TASLP.2020.3030497.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. Audiogen: Textually guided audio generation, 2023. URL https://arxiv.org/abs/2209.15352.

Saksham Singh Kushwaha and Yapeng Tian. Vintage: Joint video and text conditioning for holistic audio generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13529–13539, 2025.

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. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 21450–21474. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/liu23f.html.

Huadai Liu, Kaicheng Luo, Jialei Wang, Wen Wang, Qian Chen, Zhou Zhao, and Wei Xue. Thinksound: Chain-of-thought reasoning in multimodal LLMs for audio generation and editing. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=mj8VN4MyrO.

Simian Luo, Chuanhao Yan, Chenxu Hu, and Hang Zhao. Diff-foley: Synchronized video-to-audio synthesis with latent diffusion models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 48855–48876. Curran Associates, Inc., 2023. doi: 10.52202/075280-2121. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/98c50f47a37f63477c01558600dd225a-Paper-Conference.pdf.

Navonil Majumder, Chia-Yu Hung, Deepanway Ghosal, Wei-Ning Hsu, Rada Mihalcea, and Soujanya Poria. Tango 2: Aligning diffusion-based text-to-audio generations through direct preference optimization. In *Proceedings of the 32nd ACM International Conference on Multimedia*, MM '24, pp. 564–572, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706868. doi: 10.1145/3664647.3681688. URL https://doi.org/10.1145/3664647.3681688.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D. Plumbley, Yuexian Zou, and Wenwu Wang. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3339–3354, 2024. doi: 10.1109/TASLP.2024.3419446.

William Peebles and Saining Xie. Scalable Diffusion Models with Transformers . In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182, Los Alamitos, CA, USA, October 2023. IEEE Computer Society. doi: 10.1109/ICCV51070.2023.00387. URL https://doi.ieeecomputersociety.org/10.1109/ICCV51070.2023.00387.

Haotian Sun, Tao Lei, Bowen Zhang, Yanghao Li, Haoshuo Huang, Ruoming Pang, Bo Dai, and Nan Du. EC-DIT: Scaling diffusion transformers with adaptive expert-choice routing. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=PxlfzEePC0.

Zeyue Tian, Zhaoyang Liu, Yizhu Jin, Ruibin Yuan, Liumeng Xue, Xu Tan, Qifeng Chen, Wei Xue, and Yike Guo. Audiox: A unified framework for anything-to-audio generation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=qjJWxK3yWo.

Heng Wang, Jianbo Ma, Santiago Pascual, Richard Cartwright, and Weidong Cai. V2a-mapper: a lightweight solution for vision-to-audio generation by connecting foundation models. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024a. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i14.29475. URL https://doi.org/10.1609/aaai.v38i14.29475.

Yongqi Wang, Wenxiang Guo, Rongjie Huang, Jiawei Huang, Zehan Wang, Fuming You, Ruiqi Li, and Zhou Zhao. Frieren: Efficient video-to-audio generation network with rectified flow matching. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 128118–128138. Curran Associates, Inc., 2024b. doi: 10.52202/079017-4068. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/e7384de302bef52319b5067c3115bbfb-Paper-Conference.pdf.

Yake Wei, Ruoxuan Feng, Zihe Wang, and Di Hu. Enhancing multimodal cooperation via sample-level modality valuation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27338–27347, 2024.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10095969.

Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. Seeing and hearing: Open-domain visual-audio generation with diffusion latent aligners. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7151–7161, 2024. doi: 10.1109/CVPR52733.2024.00683.

Yike Yuan, Ziyu Wang, Zihao Huang, Defa Zhu, Xun Zhou, Jingyi Yu, and Qiyang Min. Expert race: A flexible routing strategy for scaling diffusion transformer with mixture of experts. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=9W3CSuHoFA.

Yiming Zhang, Yicheng Gu, Yanhong Zeng, Zhening Xing, Yuancheng Wang, Zhizheng Wu, and Kai Chen. FOLEYCRAFTER: BRING SILENT VIDEOS TO LIFE WITH LIFELIKE AND SYNCHRONIZED SOUNDS, 2025. URL https://openreview.net/forum?id=4cQVUNpPkt.

Appendix

A Details of Datasets

A.1 Text-to-Audio Data

For T2A training, we combine the training splits of AudioCaps Kim et al. (2019), Clotho Drossos et al. (2020), and WavCaps Mei et al. (2024). AudioCaps consists of in-the-wild audio clips derived from AudioSet and paired with human-written descriptions. Clotho contains general environmental recordings collected from Freesound, with five human-written captions describing each recording. WavCaps provides approximately 400K weakly labeled audio–text pairs collected from FreeSound, BBC Sound Effects, SoundBible, and the strongly labeled subset of AudioSet. Its captions are obtained by filtering and rewriting the original web-sourced descriptions. Each audio–caption pair is treated as an independent T2A training example.

A.2 Video-Conditioned Data

For V2A and VT2A training, we use VGGSound Chen et al. (2020) and a subset of IF-Caps Tian et al. (2026). VGGSound contains in-the-wild YouTube videos spanning more than 300 sound-event categories, with each example provided as a 10-second audio–visual segment. The dataset was constructed to provide audio–visual correspondence, such that the labeled sound source is expected to be visually evident in the associated video.

IF-Caps augments large-scale video–audio datasets with fine-grained text descriptions of their acoustic content. Its annotations include global audio captions and structured information about sound-event categories, counts, and temporal organization. We use approximately 60K general-audio examples after excluding samples that overlap with VGGSound.

The video–audio samples from both datasets are used to construct V2A and VT2A training examples. For V2A, the text condition is replaced with a learnable empty token. For VT2A, the video condition is paired with the original text description provided by VGGSound or the corresponding fine-grained description from IF-Caps. Thus, the V2A and VT2A examples share the same video–audio data and differ only in whether the associated text condition is provided.

B Details of Objective Evaluation Metrics

Fréchet Distance (FD). We assess distributional fidelity using FD, computed as the Fréchet distance between the PANNs embedding distributions of generated and reference audio Kong et al. (2020). Lower values indicate that the generated samples more closely match the reference distribution in the PANNs feature space.

Kullback–Leibler Divergence (KL). To quantify sample-level semantic agreement, we compare the multi-label audio-event posterior distributions predicted by PANNs Kong et al. (2020) for each generated audio sample and its paired reference sample. The resulting KL divergences are averaged over the evaluation set, with lower values indicating greater agreement in audio-event content.

Inception Score (IS). IS evaluates generated audio using the audio-event posterior distributions predicted by PANNs Kong et al. (2020). It rewards confident predictions for individual samples and a diverse marginal prediction distribution across the generated set. Higher values indicate greater prediction confidence and diversity.

ImageBind Score (IB-Score). Video–audio semantic correspondence is measured by the cosine similarity between the conditioning video and generated-audio embeddings from the corresponding ImageBind encoders Girdhar et al. (2023). A higher IB-Score reflects stronger semantic alignment between the generated audio and video condition.

Method	T2A		V2A			VT2A			Overall		
	MOS-Q	MOS-S	MOS-Q	MOS-S	MOS-T	MOS-Q	MOS-S	MOS-T	MOS-Q	MOS-S	MOS-T
MMAudio Cheng et al. (2025a)	3.17	3.08	3.64	3.52	3.48	3.59	3.38	3.49	3.47	3.33	3.49
AudioX Tian et al. (2026)	2.98	3.21	2.23	2.02	1.77	2.84	2.57	2.41	2.68	2.60	2.09
Omni2Sound Dai et al. (2026)	3.23	3.35	**3.77**	**3.92**	3.58	3.49	3.51	3.73	3.50	3.59	3.65
Ours	**3.84**	**3.97**	3.73	3.83	**3.74**	**3.80**	**3.80**	**3.84**	**3.79**	**3.86**	**3.79**

Table 5: Task-wise subjective evaluation results. Each score is averaged over 12 evaluators and 10 conditioning instances. MOS-T is not applicable to T2A because no video condition is provided. Higher values indicate better performance. The best result in each column is shown in bold.

CLAP Score. We measure text–audio semantic alignment as the cosine similarity between the text condition and generated-audio embeddings produced by a pretrained Contrastive Language–Audio Pretraining model. Unless otherwise specified, we use LAION-CLAP Wu et al. (2023) for the main experiments in Table 1 of the main paper and MS-CLAP Elizalde et al. (2023) for all ablation and supplementary evaluations. A higher CLAP score indicates that the generated audio more closely matches the semantic content of its text condition.

DeSync Score. Temporal synchronization is evaluated using DeSync, which reports the average audio–visual temporal misalignment, in seconds, predicted by a pretrained Synchformer Iashin et al. (2024). Lower values correspond to better temporal alignment between the generated audio and its conditioning video.

C Details of Subjective Evaluation

C.1 Evaluation setup

We conducted a subjective evaluation covering text-to-audio (T2A), video-to-audio (V2A), and video-text-to-audio (VT2A) generation. The study involved 12 evaluators and 30 test samples, with 10 samples selected for each task. For each sample, evaluators assessed outputs from our model and three unified baselines, all of which support T2A, V2A, and VT2A, resulting in 120 generated outputs. Each evaluator assessed all 120 generated outputs. The system identities were hidden from the evaluators, and the presentation order was randomized independently for each evaluator and conditioning instance to mitigate method-identity and ordering biases.

C.2 Evaluation criteria

We evaluated the generated samples in terms of acoustic quality, semantic consistency with the conditioning inputs, and, for video-conditioned tasks, audio–visual temporal synchronization. Each applicable criterion was rated on a five-point scale from 1 to 5, with higher scores indicating better performance. The evaluation criteria are defined as follows:

- **Acoustic quality (MOS-Q).** Evaluators rated the overall perceptual quality of the generated audio, considering its naturalness, clarity, and freedom from audible artifacts.
- **Semantic consistency (MOS-S).** Evaluators rated how well the semantic content of the generated audio matches the provided conditioning input. For T2A, MOS-S assessed consistency with the text prompt. For V2A, it assessed semantic correspondence with the visual content. For VT2A, it assessed consistency with both the video and text conditions.
- **Temporal synchronization (MOS-T).** For V2A and VT2A, evaluators rated how accurately the generated sound events are temporally aligned with the corresponding visible events in the video. MOS-T was not evaluated for T2A because no video condition is provided.

C.3 Task-wise Results

Table 5 presents the task-wise subjective evaluation results. Our model achieves the highest MOS-Q and MOS-S scores on T2A, indicating strong perceptual quality and semantic consistency with the text condition. On V2A, Omni2Sound obtains slightly higher MOS-Q and MOS-S scores, while our

Setting	KL ↓	FD ↓	IS ↑	IB ↑	CLAP ↑	DeSync ↓
All	1.869	7.098	5.573	0.322	0.322	0.436
Every 2nd	1.874	7.110	5.591	0.321	0.323	0.442
Every 4th†	**1.853**	**6.910**	**5.600**	**0.325**	**0.325**	**0.435**

Table 6: Effect of MoE block placement on VT2A performance. The selected configuration is marked by †, and the best results are shown in bold.

model achieves the best MOS-T score, indicating stronger audio–visual temporal synchronization. For the more challenging VT2A task, our model performs best across all three criteria, suggesting that it effectively integrates the joint video and text conditions while maintaining high acoustic quality and temporal alignment. Our model achieves the highest overall MOS-Q and MOS-S scores averaged across all three tasks, as well as the highest MOS-T score averaged across V2A and VT2A.

D Implementation Details

We build our model on the pretrained MMAudio-L-44.1kHz backbone Cheng et al. (2025a), a flow-matching-based Diffusion Transformer supporting text, video, and joint video–text conditioning. The backbone comprises 7 multimodal Transformer blocks followed by 14 fused blocks, with a hidden dimension of 896 and 14 attention heads. We replace the FFNs at zero-based indices $\{0, 4, 8, 12\}$ within the fused-block stack with the proposed MoE blocks. Each MoE block retains the original pretrained FFN as a shared expert and contains two task-associated pools of four routed experts each. Each routed expert is implemented as a rank-16 low-rank residual adaptation of the shared expert. Its up-projection is zero-initialized so that the residual contribution is zero at initialization.

We retain the pretrained condition encoders and processing pipeline of MMAudio without modification. Text descriptions are represented by 77 token-level, $1{,}024$-dimensional features from the DFN5B-CLIP ViT-H/14-384 text encoder. Video frames sampled at 8 fps are encoded into $1{,}024$-dimensional semantic features by the corresponding visual encoder. A pretrained Synchformer additionally processes frames sampled at 25 fps and produces 768-dimensional temporally localized features at an effective rate of 24 features per second. The original conditional synchronization module aligns these features with the audio latent sequence. Following MMAudio, we generate 8-second audio at 44.1 kHz using its pretrained audio VAE, which represents each mel spectrogram as a sequence of 345 latent tokens.

We use a random seed of $14{,}159{,}265$. In Stage 1, the T2A- and V2A-associated experts and their corresponding routers are independently trained for $10{,}000$ optimization steps, while the remaining pretrained parameters are frozen. In Stage 2, the composed model is jointly trained on T2A, V2A, and VT2A for $65{,}000$ steps. All parameters except the newly initialized unified routers use a learning rate of 1×10^{-5}, whereas the unified routers use 3×10^{-5}. All learning rates are reduced by a factor of 10 at Stage 2 step $40{,}000$.

We apply the polarization-aware routing intervention only at fused block 8, using a dominance threshold of $\tau_{\mathrm{dom}} = 0.75$ and a target intervention probability of $q_{\mathrm{target}} = 0.25$. The intervention is disabled during the first 10% of Stage 2 training. From 10% to 20% of the training schedule, its activation probability increases linearly from 0 to q_{target} and remains fixed thereafter. With $65{,}000$ Stage 2 steps, this corresponds to starting the probability ramp at step $6{,}500$, reaching q_{target} at step $13{,}000$, and maintaining it for the remaining steps. We set the task-conditioned load-balancing weight to $\lambda_{\mathrm{bal}} = 0.001$. For the expert-to-shared lower-bound objective, we use $\lambda_{\mathrm{lb}} = 0.005$ and $\rho = 0.15$. All remaining optimization, EMA, classifier-free guidance, sampling, audio decoding, and waveform reconstruction settings follow MMAudio-L-44.1kHz Cheng et al. (2025a).

E MoE Block Placement Analysis

We examine the effect of replacing every fused block, every second block, or every fourth block with an MoE block. As shown in Table 6, placing MoE blocks every fourth block achieves the best performance across all reported metrics. In particular, compared with replacing every fused

Setting	q_{target}	τ_{dom}	Block	KL↓	FD↓	IS↑	IB↑	CLAP↑	DeSync↓
A. Intervention probability									
$q_{\text{target}} = 0.25^{\dagger}$	0.25	0.75	8	**1.853**	**6.910**	**5.600**	**0.325**	**0.325**	0.435
$q_{\text{target}} = 0.50$	0.50	0.75	8	1.864	6.985	5.551	**0.325**	0.324	0.444
$q_{\text{target}} = 1.00$	1.00	0.75	8	1.870	6.987	5.585	0.323	0.324	**0.434**
B. Detection threshold									
$\tau_{\text{dom}} = 0.60$	0.25	0.60	8	1.869	7.007	5.603	0.323	0.324	0.438
$\tau_{\text{dom}} = 0.75^{\dagger}$	0.25	0.75	8	**1.853**	**6.910**	5.600	**0.325**	**0.325**	**0.435**
$\tau_{\text{dom}} = 0.90$	0.25	0.90	8	1.870	7.010	**5.604**	0.323	0.324	0.440
C. Intervention location									
Fused block 0	0.25	0.75	0	1.855	7.034	5.508	**0.325**	0.323	0.441
Fused block 4	0.25	0.75	4	1.870	7.016	**5.608**	0.323	**0.325**	0.442
Fused block $8^{\dagger}$	0.25	0.75	8	**1.853**	**6.910**	5.600	**0.325**	**0.325**	**0.435**
Fused block 12	0.25	0.75	12	1.856	7.012	5.526	**0.325**	0.324	0.442

Table 7: Sensitivity of VT2A performance to the target intervention probability q_{target}, detection threshold τ_{dom}, and intervention location. Each panel varies one factor while keeping the others fixed at $q_{\text{target}} = 0.25$, $\tau_{\text{dom}} = 0.75$, and fused block 8, as applicable. The selected configuration is marked by †, and the best results within each panel are shown in bold.

Setting	KL ↓	FD ↓	IS ↑	CLAP ↑	DeSync ↓	$\tilde{r}_{\text{sel}}$
w/o $\mathcal{L}_{\text{lb}}$	1.854	6.991	5.561	0.322	0.448	0.117
$\rho = 0.15^{\dagger}$	**1.853**	**6.910**	**5.600**	**0.325**	**0.435**	0.401
$\rho = 0.50$	1.858	7.075	5.529	0.324	0.440	1.140
$\rho = 1.00$	1.854	7.053	5.529	0.323	0.444	2.513

Table 8: Ablation and sensitivity analysis of the expert-to-shared lower-bound objective on VT2A. $\widetilde{r}_{\text{sel}}$ denotes the selection-count-weighted mean selected-expert output norm normalized by the mean shared-expert output norm. The selected value is marked by †, and the best performance results are shown in bold.

block, this configuration reduces FD from 7.098 to 6.910 and improves IB from 0.322 to 0.325. These results indicate that selectively distributing MoE blocks within the fused-block stack is more effective than densely replacing all FFNs in our configuration. We therefore apply MoE blocks at zero-based fused-block indices $\{0, 4, 8, 12\}$ in all main experiments.

F Polarization Intervention Sensitivity Analysis

We examine the sensitivity of the polarization-aware routing intervention to its activation probability q_{target}, detection threshold τ_{dom}, and intervention location. Table 7 summarizes the VT2A results while varying one component at a time.

F.1 Intervention probability

A moderate intervention probability of $q_{\text{target}} = 0.25$ provides the best overall trade-off, achieving the lowest KL and FD and the highest IS and CLAP. Increasing q_{target} to 0.50 or 1.00 generally degrades these metrics, although $q_{\text{target}} = 1.00$ yields a slightly lower DeSync score. These results indicate that selectively intervening in a subset of detected polarization cases is preferable to applying the intervention deterministically.

F.2 Detection threshold

The default threshold $\tau_{\text{dom}} = 0.75$ achieves the best KL, FD, IB, CLAP, and DeSync results. A lower threshold of 0.60 may activate the intervention for less severe routing asymmetries, whereas a higher threshold of 0.90 restricts it to more strongly polarized cases. Although $\tau_{\text{dom}} = 0.90$ produces a

marginally higher IS, its remaining metrics are weaker than those obtained with $\tau_{\mathrm{dom}} = 0.75$. This result supports the use of an intermediate threshold that targets excessive pool dominance without reacting to moderate, sample-dependent routing preferences.

F.3 Intervention location

Applying the intervention at fused block 8 results in the strongest overall performance, with the lowest KL, FD, and DeSync scores. Block 4 achieves a slightly higher IS and matches block 8 in CLAP, while the remaining locations produce weaker results on most metrics. These results empirically support fused block 8 as the intervention location for our model configuration.

Overall, the selected configuration uses $q_{\mathrm{target}} = 0.25$, $\tau_{\mathrm{dom}} = 0.75$, and fused block 8. It achieves the best performance on most metrics and remains stable across nearby settings.

G Expert-to-Shared Lower-Bound Loss Analysis

Table 8 evaluates the contribution of the expert-to-shared lower-bound objective and its sensitivity to the target ratio ρ. Without $\mathcal{L}_{\mathrm{lb}}$, $\widetilde{r}_{\mathrm{sel}}$ is 0.117, indicating that the selected expert outputs remain small relative to the shared-expert output. Using $\rho = 0.15$ increases this ratio to 0.401 and improves all reported performance metrics. In particular, FD decreases from 6.991 to 6.910, DeSync decreases from 0.448 to 0.435, and MS-CLAP increases from 0.322 to 0.325.

Increasing ρ to 0.50 and 1.00 further raises $\widetilde{r}_{\mathrm{sel}}$ to 1.140 and 2.513, respectively, but does not yield additional performance gains. Instead, these stronger constraints generally weaken the reported metrics. This pattern shows that the objective is most effective as a moderate lower bound that prevents negligible selected-expert outputs, rather than as a mechanism for maximizing their magnitude. We therefore use $\rho = 0.15$ in all main experiments.

H Model Size and Training Cost

Excluding the pretrained condition encoders, the MMAudio-L-44.1kHz backbone contains approximately 1.03B parameters. Introducing the task-associated routed experts and unified routers increases the total parameter count to approximately 1.11B, corresponding to an increase of approximately 80M parameters, or 7.7% relative to the backbone.

Under top-2 routing, approximately 1.05B parameters lie on the nonzero routing paths for each token. This count includes the backbone, all unified routers, and the two selected routed experts in each MoE block.

Training is performed with a per-GPU batch size of 96 on an NVIDIA RTX PRO 6000 GPU and requires approximately 18 hours for the complete training procedure. The pretrained condition encoders are excluded from the parameter counts reported above.