RETHINKING MIXTURE-OF-EXPERTS FOR AUDIO-VISUAL DIFFUSION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

This work systematically investigates how to design and train Mixture-of-Experts (MoE) architectures for large-scale audio-visual diffusion models. Audio-visual latent tokens form continuous representations with strong spatial, temporal, and cross-modal correlations, motivating a re-evaluation of existing MoE design choices. Through systematic study, we summarize sparse capacity design principles tailored to audio-visual generation. We strengthen text conditioning with richer semantic representations while keeping the text conditioning branch dense and allocating sparse capacity to the visual backbone. Meanwhile, we find that load variation persists under continuous balancing and can arise from either semantic specialization or optimization instability. We therefore combine routing regulation with targeted stabilization to preserve normal expert differentiation while suppressing instability-driven collapse. Finally, we validate these designs using millions of high-quality audio-visual samples in a large-scale distributed training environment, establishing architectural and training principles for large-scale audio-visual diffusion MoE models.

1 INTRODUCTION

Recent advances in generative modeling have increasingly explored Mixture-of-Experts (MoE) for image, video, and audio-visual generation, offering a promising path to scale model capacity under a limited computational budget. However, existing studies (Cheng et al., 2025; Yuan et al., 2025; Ma et al., 2026) have primarily demonstrated the feasibility of sparse expert scaling in particular generative architectures, while the design principles that make MoE effective for audio-visual diffusion models remain insufficiently understood. Unlike autoregressive language modeling, audio-visual diffusion operates on continuous latent representations with strong spatial, temporal, and cross-modal correlations. Visual latent tokens correspond to local patches, while both audio and video representations evolve along the diffusion trajectory. These differences motivate a systematic re-evaluation of MoE design choices established for language models.

This work asks how to turn additional sparse parameters into effective generative capacity under a fixed active-compute budget. We study this question at two levels: where and how to allocate expert capacity, and how to train the resulting experts stably. Although audio-visual sequences contain many latent tokens, their strong correlations mean that token abundance does not necessarily provide equally diverse specialization signals. Increasing nominal sparse capacity therefore does not guarantee a proportional improvement in generation quality.

At the architectural level, we examine activation ratio, expert granularity, and the interaction between shared and routed experts. These choices determine how sparse capacity is distributed and combined within the model. Text conditioning is part of this allocation problem: its sequences are substantially shorter than audio-visual sequences, and its representations are supervised indirectly through the generative objective, which may limit stable expert specialization. We therefore strengthen the text representations while keeping the text conditioning branch dense, and allocate sparse capacity to the visual backbone.

At the training level, we investigate how to maintain balanced expert utilization while allowing meaningful specialization to emerge. Although a routing bias continuously regulates expert loads throughout training, load variation still occurs and can arise from two distinct sources. As noise decreases along the denoising trajectory, expert preferences become more differentiated, producing

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.



Figure 1: Representative frames from videos generated by our model, spanning human interactions, musical performances, natural scenes, and stylized characters. The examples showcase diverse subjects and visual styles, with detailed textures, expressive faces, and varied lighting.

benign variation associated with semantic specialization and changes in the training distribution. In contrast, optimization anomalies such as gradient spikes and expert-output outliers can induce abnormal concentration and destabilize training. We therefore retain continuous routing-based load regulation while introducing additional MoE-specific gradient stabilization only for optimization-induced abnormal dynamics.

Together, these architectural and training choices form a recipe for making sparse capacity effective in audio-visual diffusion models. We study the resulting design through large-scale pretraining on millions of high-quality audio-visual samples in a distributed training environment. Representative generated frames are shown in Fig. 1.

Our main contributions are summarized as follows:

- We investigate sparse capacity allocation in audio-visual diffusion MoE, examining activation ratio, expert granularity, and shared/routed expert interaction under a fixed active-compute budget.
- We analyze load variation under routing regulation, distinguishing benign variation associated with noise-dependent expert specialization from abnormal concentration induced by optimization instability, and develop targeted stabilization strategies for the latter.
- We consolidate these findings into a practical recipe for large-scale audio-visual diffusion MoE and validate it through pretraining on millions of high-quality samples, demonstrating strong generation quality, audio-visual synchronization, and stable training at scale.

2 Related Work

2.1 Audio-Visual Generation

Recent unified audio-visual generation models have begun to explore expert-based architectures for jointly modeling visual and acoustic signals (Polyak et al., 2024; Liu et al., 2026a;b; Wang et al., 2026). These methods investigate modality-specific experts, expert composition, and unified multimodal backbones, demonstrating the feasibility of jointly generating synchronized audio and video. In parallel, advances in semantic conditioning increasingly exploit representations from large pretrained language models to improve prompt understanding and semantic alignment in diffusion- and

flow-based generative models. However, existing audio-visual generation methods generally emphasize model unification and cross-modal synchronization, while the scaling behavior and design principles of sparse expert architectures remain comparatively underexplored.

2.2 Mixture-of-Experts in Large Language Models

Mixture-of-Experts (MoE) is widely adopted for scaling language models through sparse conditional computation (Shazeer et al., 2017; Lepikhin et al., 2021; Fedus et al., 2022; Du et al., 2022). Recent LLMs advance this paradigm with diverse expert architectures and routing mechanisms, including Mixtral (Jiang et al., 2024), the DeepSeek series (Dai et al., 2024; DeepSeek-AI, 2024; 2025), OLMoE (Muennighoff et al., 2025), and Qwen (Yang et al., 2025). Recent work studies how expert number, granularity, top-K routing, activation rate, and active computation affect MoE scaling (Clark et al., 2022; Ludziejewski et al., 2024; Frantar et al., 2024; Abnar et al., 2025). Large-scale studies further show that sparsity is not universally beneficial: optimal configurations depend on active compute, data allocation, architecture (Li et al., 2026; Nakamura et al., 2026), and downstream capability (Zoph et al., 2022). Other work investigates load balancing, router stability, and expert specialization through auxiliary objectives, shared experts, orthogonalization, autonomous routing, and explicit router–expert coupling (Dai et al., 2024; DeepSeek-AI, 2025; Lv et al., 2025; Guo et al., 2025; Lv et al., 2026). These studies show that effective MoE scaling requires sparse activation alongside a balance between expert specialization, collaboration, and routing stability.

However, these principles were mainly developed for autoregressive language models with discrete tokens. Their effectiveness for continuous, highly correlated audio-visual representations remains unclear, especially given the distinct temporal and cross-modal structures of audio and video.

2.3 Mixture-of-Experts for Visual Generation

MoE has recently been extended from language modeling to visual generative models. Diff-MoE (Cheng et al., 2025) introduces time-aware and space-adaptive experts into diffusion Transformers, enabling the model to specialize its computation according to the denoising timestep and spatial content. Expert Race (Yuan et al., 2025) focuses on routing strategies and proposes a flexible mechanism for selecting experts in diffusion Transformers. Dense2MoE (Zheng et al., 2025) restructures a dense diffusion Transformer into an MoE architecture through dense-to-sparse conversion, aiming to improve generation efficiency while preserving the capacity of the pretrained model. ProMoE (Wei et al., 2026) provides explicit routing guidance to stabilize and improve expert selection during diffusion generation.

MoE has also been applied to large-scale video generation. The publicly released Wan2.2 model (Wan-AI, 2025) assigns specialized experts to different denoising stages, while LingBot-Video (Ma et al., 2026) uses sparse feed-forward computation to scale video generation to complex spatiotemporal content. These studies demonstrate the potential of conditional computation for improving the capacity–compute trade-off in image and video generation. However, their designs primarily target visual signals and do not directly address the heterogeneous statistical properties or synchronization requirements of jointly generated audio and video.

Despite this progress, generative MoE studies offer limited systematic guidance on jointly configuring activation ratio, expert granularity, shared and routed capacity, semantic conditioning, and specialization dynamics. These gaps motivate our study of practical design principles and large-scale training recipes for MoE audio-visual diffusion models.

3 Method: Design Principles for Audio-Visual Diffusion MoE

Architecture Overview. Our architecture supports text-conditioned audio-visual generation. As shown in Fig. 2, we apply MoE only to the video backbone. The audio DiT is warm-started from the pretrained ALIVE (Anonymous, 2026) audio module and is therefore kept dense to preserve its pretrained generative capability during joint audio-visual training. The following sections describe sparse capacity allocation in the visual backbone, text conditioning, and the routing and optimization strategies used to train the model.

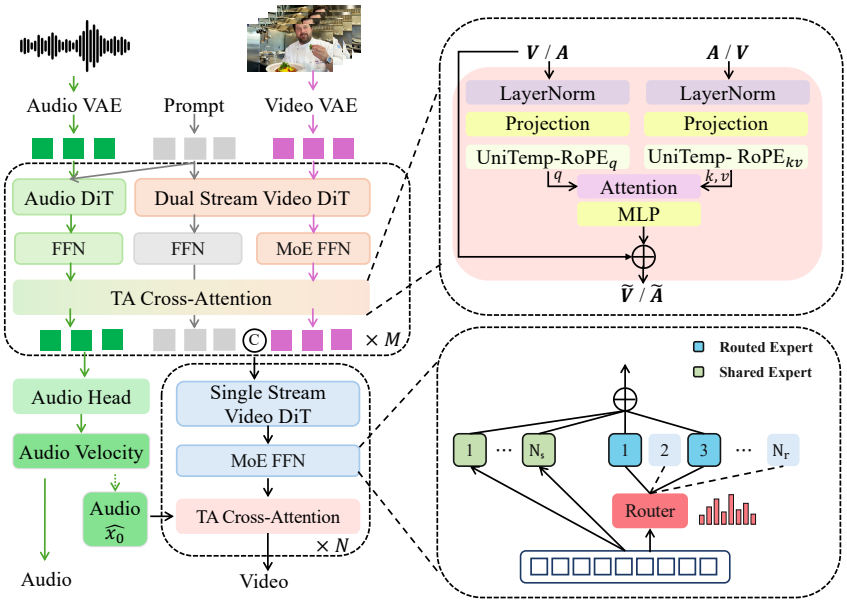

Figure 2: Overview of the audio-visual diffusion MoE architecture. Audio and video branches interact through temporally aligned cross-attention; the video backbone uses shared and routed experts, while dual-stream text FFNs remain dense.

3.1 MoE Formulation

We first give some key definitions of an audio-visual diffusion MoE Transformer. For an input token representation h_i , a routed MoE layer computes router scores

$$s_{i,j} = r_j(h_i), \quad j = 1, \ldots, N,$$

and selects K experts from a total expert pool of size N. The sparse expert output is

$$\mathrm{E}_{\text{routed}}(h_i) = \sum_{j \in \mathcal{T}_K(i)} w_{i,j} E_j(h_i),$$

where $\mathcal{T}_K(i)$ is the selected expert set and $w_{i,j}$ is the normalized routing weight. In practice, we combine routed experts with a shared expert path:

$$\mathrm{FFN}_{\text{MoE}}(h_i) = \lambda_s E_{\text{shared}}(h_i) + \lambda_r E_{\text{routed}}(h_i).$$

The central design question is not merely how to insert MoE layers into a diffusion model, but how to allocate expert capacity, constrain routing, and maintain stable training dynamics under large-scale audio-visual generation.

3.2 Capacity Allocation in Audio-Visual MoE

For **capacity allocation**, we study sparse capacity, expert partitioning, and interactions with shared paths and conditioning branches at a fixed active parameter budget. Visual latent tokens are spatially correlated local patches rather than complete semantic units, motivating a re-evaluation of language-MoE heuristics. We define $\boldsymbol{EL}$ as relative loss-equivalent dense capacity, normalized by the dense anchor used in the controlled sweep (Sec. A.3).

3.2.1 Activation Ratio: Avoiding Over-Sparsification

We define the activation ratio as $A = K/N$, where K is the number of activated experts per token and N is the total number of routed experts. Holding K fixed and increasing N expands the total parameter pool without increasing per-token activated computation. This is the primary mechanism by which sparse expert scaling improves model capacity under a fixed inference budget. Specifically,

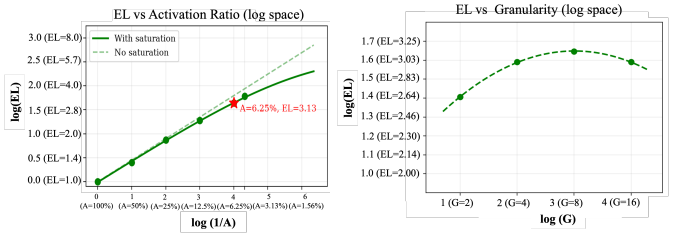

Figure 3: Effects of activation ratio A and expert granularity G on effective capacity. A controls the fraction of routed experts activated per token; G controls sparse capacity partitioning. EL is loss-equivalent dense capacity normalized by the sweep's dense anchor.

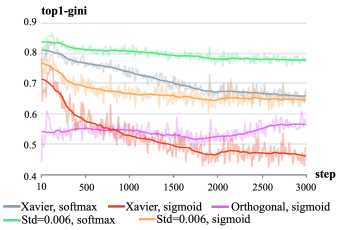

Figure 4: Comparison of initialization and router activation choices. Xavier with sigmoid reaches the lowest Top-1 Gini over the displayed training interval.

we study activation ratios by fixing $K = 8$ and varying $N \in \{8, 16, 32, 64, 128, 160\}$, corresponding to activation ratios from 100% to 5%. The pattern can be fitted as

$$\boldsymbol{EL} = (\mathbf{1}/\widehat{\boldsymbol{A}})^{\mathbf{0.4491}}, \widehat{\mathrm{A}} = \mathrm{A} + \delta, \delta \approx 0.0163$$

and the corresponding log–log plot is shown in Fig. 3. We observe a saturating trend: reducing A initially improves capacity-efficiency, but the gain diminishes when A approaches the low single-digit regime. This suggests that audio-visual MoE benefits from sparsity, but not from blindly pushing toward extremely sparse activation.

3.2.2 Expert Granularity: Avoiding Over-Fragmentation

The second factor is expert granularity. For a standard dense FFN hidden dimension $4 \times d$, we define expert granularity as $G = \frac{4 \times d}{d_{\text{expert}}}$, where d_{expert} is the hidden width of each expert. Larger G means narrower experts and finer capacity partitioning, while smaller G means wider experts and coarser specialization. Since the total activated width is typically fixed, $W_{\text{act}} = K \cdot d_{\text{expert}}$, d_{expert} and K therefore exhibit an inherent trade-off: increasing d_{expert} improves the representational capacity of each expert but reduces how many experts can be activated, whereas decreasing d_{expert} allows more experts to participate but weakens the capacity of each expert. Empirically, as shown in Fig. 3, performance follows an inverted-U trend in log–log space, which can be approximated as

$$\log \boldsymbol{EL} = 0.77 - 0.087(\log G)^2 + 0.36(\log G). \tag{1}$$

The optimal region lies around $G = 8 \sim 10$, suggesting that expert granularity directly controls the balance between per-expert capacity and the number of active experts, and that neither overly wide nor overly narrow experts are optimal.

3.2.3 Shared/Routed Expert Interaction

We observe a structural norm imbalance between the shared and routed branches in MoE layers. The MoE output is written as

$$\mathbf{y} = \text{gate} \cdot \mathbf{y}^{\text{shared}} + \sum_{i=1}^{K} w_i \, \mathbf{y}_i^{\text{routed}}. \tag{2}$$

where $\mathbf{y}^{\text{shared}}$ is the shared-branch output, $\mathbf{y}_i^{\text{routed}}$ is the output of the i-th selected routed expert, and w_i denotes the normalized routing weight with $\sum_{i=1}^{K} w_i = 1$. Under normalized top-K routing, the routed contribution is a weighted average rather than a direct sum. When the selected expert outputs have comparable norms and weak pairwise correlations, $w_i \approx 1/K$:

$$\left\| \sum_{i=1}^{K} w_i \mathbf{y}_i^{\text{routed}} \right\| \approx \sqrt{\sum_{i=1}^{K} \sum_{j=1}^{K} w_i w_j \left\langle \mathbf{y}_i^{\text{routed}}, \mathbf{y}_j^{\text{routed}} \right\rangle} \approx \sqrt{\sum_{i=1}^{K} w_i^2} \, \|\mathbf{y}^{\text{routed}}\| \approx \frac{1}{\sqrt{K}} \|\mathbf{y}^{\text{routed}}\|, \tag{3}$$

As a result, averaging K routed experts reduces the routed-branch magnitude by roughly $1/\sqrt{K}$, so activating more experts does not maintain the scale of a single expert output. Under a merged-

parameter approximation, the three SwiGLU projections each contribute a $\sqrt{N_s}$ scale factor, yielding $N_s^{3/2}$ for the shared output (Sec. A.1). Combining this estimate with routed averaging gives

$$\left\|\mathbf{y}^{\text{shared}}\right\| / \left\|\sum_{i=1}^{K} w_i \mathbf{y}_i^{\text{routed}}\right\| \;\approx\; \text{gate} \cdot N_s^{3/2} \cdot \sqrt{K}. \tag{4}$$

This imbalance can cause the shared branch to dominate the routed contribution, potentially weakening expert specialization. To compensate for this structural bias, we use

$$\mathbf{y} = \text{gate} \cdot \mathbf{y}^{\text{shared}} + N_s^{3/2} \cdot \sqrt{K} \cdot \sum_{i=1}^{K} w_i \mathbf{y}_i^{\text{routed}}. \tag{5}$$

3.2.4 Text Conditioning Design for Audio-Visual MoE

Increasing sparse model capacity alone does not necessarily improve generation when the conditioning representation provides insufficient semantic information. We therefore strengthen the textual conditioning with semantic representations extracted from Qwen3.6-27B (Qwen Team, 2026). Specifically, we empirically select the hidden representation from its 32nd layer and further adapt it using an adaptor initialized from the pretrained Qwen Layers 28–32. This design aims to provide the generative backbone with richer high-level semantic information while retaining representations suitable for generative conditioning.

Unlike the visual backbone, we keep the text conditioning branch dense rather than introducing MoE. Since it is supervised only indirectly through the flow matching loss, sparse routing may make text experts specialize according to modality-specific optimization dynamics rather than text semantics. We also empirically experimented with sparse text experts under different MoE configurations, but consistently observed expert domination with limited expert differentiation. We attribute this behavior to the short text sequence and indirect supervision through the generative objective, which provide weaker routing statistics than the visual backbone. We therefore keep the text conditioning branch dense.

3.3 Load Balancing as Constrained Routing Optimization

Capacity allocation defines the potential expert space, but load-balanced routing determines whether this space can be trained effectively. Without routing constraints, MoE training can easily collapse into a small subset of experts, causing low utilization, overloaded experts, unstable gradients, and eventually degraded generation quality.

3.3.1 Constrained Routing

We formulate routing as a constrained assignment problem. Given router score $s_{i,j}$ between token i and expert j, the goal is to maximize semantic matching while satisfying load constraints:

$$\max_{\mathbf{x}} \sum_{i,j} s_{i,j} x_{i,j} \quad \text{s.t.} \quad \sum_j x_{i,j} = K,\ \forall i, \qquad \sum_i x_{i,j} \le mK/N,\ \forall j, \tag{6}$$

where m is the number of tokens in the routing batch and $x_{i,j} \in \{0,1\}$ indicates whether token i is assigned to expert j. The first constraint ensures each token activates exactly K experts, and the second tries to prevent each expert from exceeding its expected load quota.

Following the LP-dual view of constrained routing, we introduce dual variables α_i and β_j as the Lagrange multipliers associated with the per-token top-K constraint and the per-expert capacity constraint, respectively. Beyond their optimization role, they also admit clear semantic interpretations: α_i acts as a token-specific threshold, while β_j acts as an expert-specific load penalty. Token i selects experts according to corrected scores $s_{i,j} - \beta_j$, with α_i given by

$$\alpha_i = \text{Quantile}(\{s_{i,j} - \beta_j\}_{j=1}^{N},\ 1 - K/N), \tag{7}$$

which ensures exactly K assignments. The expert-level multiplier is updated as

$$\beta_j^{t+1} = \beta_j^t - \gamma\, \text{sign}(mK/N - \text{load}_j), \tag{8}$$

thereby increasing the penalty for overloaded experts and decreasing it for under-utilized ones. The full derivation is deferred to Sec. A.1

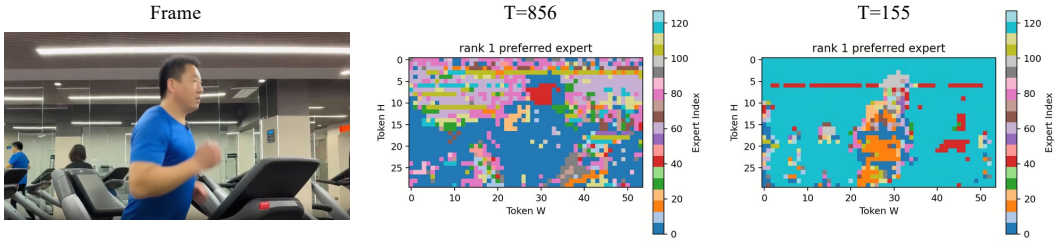

Figure 5: The top-1 expert selection pattern under different timesteps. During high-noise stages, expert preferences are not obvious, but as the noise decreases and semantics gradually form, expert choices naturally become more focused.

3.3.2 Benign Load Variation versus Harmful Collapse

Although the load penalty Eq. (8) is maintained throughout training, expert loads can still fluctuate. We observe two types of such variation: benign fluctuations associated with semantic specialization, which can evolve under normal routing dynamics, and abnormal fluctuations triggered by optimization instability, which may lead to expert collapse and require explicit stabilization.

Harmful collapse is typically triggered by optimization instability: a gradient spike produces an expert output outlier, which polarizes the router and repeatedly overloads a small subset of experts. This feedback loop further amplifies gradient instability and should be actively suppressed. In contrast, load variation can also result from stage-dependent timestep sampling during flow-matching training (Esser et al., 2024). We use a lower timestep shift during low-resolution pretraining, increase it when scaling to higher resolution to emphasize higher-noise samples and adapt coarse structural knowledge to the new token layout, and later reduce it to focus more on fine-grained details. This schedule shifts the frequency of different noise levels during training. As shown in Fig. 5, expert preferences become more concentrated at lower noise levels, so changing the timestep distribution can alter aggregate expert loads even under stable optimization. Such variation should therefore be interpreted jointly with the training stage and noise-dependent expert preferences.

Load concentration alone does not indicate routing collapse. Load penalty continuously regulates expert utilization, while additional stabilization is only applied when concentration is coupled with optimization instability. Variations driven by timestep sampling or specialization are otherwise left to evolve under normal routing dynamics.

3.4 Stabilizing Large-Scale Audio-Visual MoE Training

For audio-visual MoE training, stability is not only an optimizer issue; it is a joint property of initialization, routing control, gradient management, parallelism, kernels, data packing, and online monitoring. We therefore address stability through their combined design.

3.4.1 Router and Expert Initialization

Initialization plays a critical role in early MoE routing dynamics. We compare three plausible strategies: fixed small-variance, Xavier, and orthogonal initialization. Fixed small-variance aims to keep early router logits close and maintain balanced expert loads; orthogonal initialization encourages experts to start from distinct directions and may promote early specialization; Xavier preserves activation scale and provides a moderate level of routing diversity.

Empirically, fixed small-variance initialization produces router scores with a narrow dynamic range and delays expert differentiation, while orthogonal initialization strengthens separation but increases sensitivity to sparse-token gradient noise. Xavier initialization best balances load stability and expert specialization. We use sigmoid rather than softmax to score experts independently, reducing direct competition during early training before meaningful specialization emerges. We compute Top-1 Gini by tallying each token's top-1 expert index and calculating the Gini coefficient of the resulting per-expert counts; lower values indicate more balanced usage. In Fig. 4, sigmoid yields lower Top-1 Gini than softmax under both Xavier and fixed small-variance initialization. Our final design uses Xavier initialization for both router and expert modules, with sigmoid router activation. A detailed variance analysis is provided in Sec. A.2.

Table 1: Quantitative comparison with audio-visual generation models. Dense/MoE models appear above/below the separator. A denotes active parameters. Overall is the unweighted mean of six higher-is-better metrics. Bold indicates the column best.

Model	Parameters	Vis	Aud	AV	Face	Speech	Adherence	Overall
MiniMax H3	33B	**91.55**	**69.57**	75.85	56.58	72.60	64.77	**71.82**
LTX-2.5	22B	84.99	69.20	**82.56**	36.99	**87.37**	60.44	70.26
OVI	5B	81.59	64.72	63.01	40.88	41.69	52.75	57.44
MOVA	32B-A18B	70.80	63.47	52.02	46.70	44.62	43.25	53.48
Ours	32B-A4.7B	90.89	65.06	61.36	**66.57**	72.71	**70.87**	71.24

3.4.2 Maintaining Load Stability under Gradient-Norm Spikes

During large-scale audio-visual MoE training, gradient-norm spikes often coincide with expert-output outliers and abrupt routing shifts. Once a few experts receive excessive tokens, their overloaded updates can further enlarge expert-output outliers, which in turn make subsequent routing decisions even more concentrated. This amplification loop can turn a transient imbalance into persistent load collapse.

We monitor routed-expert gradient norms and apply clipping when spikes occur while retaining continuous bias-based load regulation. In the recorded large-scale continuation, enabling gradient-norm clipping stabilized the load dynamics; details are reported in Sec. A.4.2.

4 Experiments

4.1 Quantitative Comparison

Evaluation Set. Our evaluation set comprises 87 prompts covering intimate conversations and multilingual speech, musical performances, environmental ambience, as well as character interactions, rapid scene changes, and dramatic narratives. Together, these cases probe visual fidelity, speech and sound generation.

Evaluation Metrics. We assess visual quality (Vis) with Q-Align (Wu et al., 2024), which combines predictions across ordered quality levels for the generated frames. Audio quality (Aud) uses the Production Quality output of AudioBox Aesthetics (Tjandra et al., 2025) to reflect the technical quality of the generated sound. Audio-visual synchronization (AV) is based on the temporal offset estimated by Synchformer (Iashin et al., 2024), with smaller offsets receiving higher scores. The Face metric evaluates both agreement with the prompt's expected number of main subjects and the consistency of their face identities across the video, using DBSCAN (Ester et al., 1996) to cluster face embeddings. Speech correctness is assessed by transcribing generated dialogue with Whisper (Radford et al., 2022) and using Gemini (Gemini Team, 2023) to compare it with the dialogue specified in the prompt. Finally, Adherence uses Gemini to assess how well the video and audio follow the prompt's narrative, visual, shot, and sound requirements.

Tab. 1 compares against MiniMax H3 (MiniMax, n.d.), LTX-2.5 (HaCohen et al., 2026), OVI (Low et al., 2025), and MOVA (OpenMOSS Team et al., 2026). Our model is comparable to the 22B LTX-2.5 in Overall score and approaches the 33B MiniMax H3 with 4.7B active parameters, primarily in the video tower. It leads in Face and Adherence but trails the strongest dense baselines in Aud and AV.

4.2 Ablation Studies

Shift Scheduling for Resolution Transitions. As discussed in our training strategy Sec. A.4.5, low-resolution pretraining uses a low timestep shift to learn denoising across the trajectory. At the resolution transition, however, retaining this shift does not adequately adapt the pretrained model to the new token layout and spatial relationships. We therefore raise the shift to emphasize higher-noise states, allowing training to continue from the low-resolution checkpoint while relearning structure at



Figure 6: Effect of timestep shift at the resolution transition. Retaining a low shift produces head and hand distortions (red boxes), while increasing the shift yields more coherent structure.



Figure 7: Qualitative comparison with and without the text adaptor. Red boxes highlight the man's initial camera-facing pose, the girl's collar, and her upward-to-downward gaze transition. *Prompt summary:* A man in a beige cardigan and a girl in a light pink blouse with a **ruffled collar** sit on a white sofa in a sunlit living room. He **initially faces the camera**, then **turns toward her**. She first watches his right hand, then looks at him and lowers her gaze, repeating the **upward glance and downward look**. The camera remains static.

the higher resolution. In Fig. 6, retaining a low shift leads to distorted head and hand geometry (red boxes), whereas the increased shift preserves more coherent structure. This comparison supports the shift increase as the resolution-transition schedule.

Qwen Adaptor for Text Conditioning. We compare generation with and without the Qwen-initialized text adaptor in Fig. 7. With the adaptor, the model better reproduces the girl's light pink blouse with a ruffled collar, the man's initial camera-facing pose and subsequent turn, and the girl's glance toward him followed by a downward gaze. Without it, the girl remains oriented away from the man in the displayed frames, missing the intended interaction. This qualitative comparison illustrates improved adherence to appearance details and temporally ordered actions.

5 Conclusion

We studied how to make sparse expert capacity effective in large-scale audio-visual diffusion models under a fixed active-compute budget. Our analysis highlights the importance of activation ratio, expert granularity, and the relative contributions of shared and routed experts. The resulting design combines a sparse visual backbone with enhanced dense text conditioning. We further distinguish noise-dependent expert preferences from routing concentration associated with optimization instability, motivating coordinated initialization, routing control, and expert-gradient management. Qualitative ablations illustrate the role of timestep-shift adjustment in preserving structural coherence after resolution increases and the benefit of a Qwen-initialized text adaptor for following ordered actions. Together, these architectural and training choices provide a practical recipe for audio-visual diffusion MoE, emphasizing that effective scaling depends on how expert capacity is allocated, conditioned, and optimized.

AI USE STATEMENT

We used generative AI tools solely to improve the language and clarity of the manuscript. All AI-assisted revisions were reviewed by the authors, who take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

To support reproducibility, we document the complete multi-stage training schedule (tasks, resolution, FPS, timestep shift, learning rate, epochs, and data volume), the data construction and filtering pipeline, the constrained-routing derivation, the router/expert initialization analysis, and the routing and gradient-stabilization hyperparameters in Sections A.1, A.2 and A.4. The evaluation prompt categories, automatic metrics, and aggregate scoring protocol are described in Section 4. Owing to internal infrastructure and data-distribution restrictions, we are unable to release the codebase, training data, or model checkpoints with this submission. We instead provide detailed architectural and training descriptions so that the experimental setup and the reported analyses can be understood and, where feasible, independently replicated.

ETHICS STATEMENT

This work uses curated web videos containing speech and human subjects, and generates audio-visual content that may depict recognizable people. The training pipeline applies visual-quality, audio-quality, and audio-visual coherence filtering as well as speaker-identity correction; it does not intentionally collect sensitive personal attributes. Generated identity likenesses can be misused, so the released system should be accompanied by content provenance and appropriate access controls, and we do not support impersonation or deceptive use. A proprietary multimodal model is used only as an offline automatic judge for Speech and Adherence; only transcripts and generated content are sent for evaluation, and the judgments are treated as approximate signals rather than personal annotations. We recommend that downstream users obtain the necessary rights for source data and clearly label synthetic outputs.

REFERENCES

Samira Abnar, Harshay Shah, Dan Busbridge, Alaaeldin Mohamed Elnouby Ali, Josh Susskind, and Vimal Thilak. Parameters vs flops: Scaling laws for optimal sparsity for mixture-of-experts language models, 2025. URL https://arxiv.org/abs/2501.12370.

Anonymous. ALIVE: Animate your world with lifelike audio-video generation. *arXiv preprint arXiv:2602.08682*, 2026. URL https://arxiv.org/abs/2602.08682.

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 *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=JCUsWrwkKw.

Aidan Clark, Diego de las Casas, Aurelia Guy, Arthur Mensch, Michela Paganini, Jordan Hoffmann, Bogdan Damoc, Blake Hechtman, Trevor Cai, Sebastian Borgeaud, George van den Driessche, Eliza Rutherford, Tom Hennigan, Matthew Johnson, Katie Millican, Albin Cassirer, Chris Jones, Elena Buchatskaya, David Budden, Laurent Sifre, Simon Osindero, Oriol Vinyals, Jack Rae, Erich Elsen, Koray Kavukcuoglu, and Karen Simonyan. Unified scaling laws for routed language models. In *Proceedings of the 39th International Conference on Machine Learning*. PMLR, 2022.

Damai Dai, Chengqi Deng, Chenggang Zhao, R. X. Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y. Wu, Zhenda Xie, Y. K. Li, Panpan Huang, Fuli Luo, Chong Ruan, Zhifang Sui, and Wenfeng Liang. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models, 2024. URL https://arxiv.org/abs/2401.06066.

Dao-AILab. Quack: A quirky assortment of cute kernels. https://github.com/Dao-AILab/quack/tree/main, 2026.

DeepSeek-AI. Deepseek-v2: A strong, economical, and efficient mixture-of-experts language model, 2024. URL https://arxiv.org/abs/2405.04434.

DeepSeek-AI. Deepseek-v3 technical report, 2025. URL https://arxiv.org/abs/2412.19437.

Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. Glam: Efficient scaling of language models with mixture-of-experts. In *International conference on machine learning*, pp. 5547–5569. PMLR, 2022.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL https://arxiv.org/abs/2403.03206.

Martin Ester, Hans-Peter Kriegel, J"org Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, pp. 226–231, 1996.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Elias Frantar, Carlos Riquelme Ruiz, Neil Houlsby, Dan Alistarh, and Utku Evci. Scaling laws for sparsely-connected foundation models. In *International Conference on Learning Representations*, 2024.

Gemini Team. Gemini: A family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Hongcan Guo, Haolang Lu, Guoshun Nan, Bolun Chu, Jialin Zhuang, Yuan Yang, Wenhao Che, Xinye Cao, Sicong Leng, Qimei Cui, and Xudong Jiang. Advancing expert specialization for better moe. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=iydmH9boLb.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, Eitan Richardson, Guy Shiran, Itay Chachy, Jonathan Chetboun, Michael Finkelson, Michael Kupchick, Nir Zabari, Nitzan Guetta, Noa Kotler, Ofir Bibi, Ori Gordon, Poriya Panet, Roi Benita, Shahar Armon, Victor Kulikov, Yaron Inger, Yonatan Shiftan, Zeev Melumian, and Zeev Farbman. LTX-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. Synchformer: Efficient synchronization from sparse cues. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5325–5329, 2024.

Sam Ade Jacobs, Masahiro Tanaka, Chengming Zhang, Minjia Zhang, Shuaiwen Leon Song, Samyam Rajbhandari, and Yuxiong He. Deepspeed ulysses: System optimizations for enabling training of extreme long sequence transformer models. *arXiv preprint arXiv:2309.14509*, 2023.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. {GS}hard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=qrwe7XHTmYb.

Houyi Li, Ka Man Lo, Shijie Xuyang, Ziqi Wang, Wenzhen Zheng, Haocheng Zhang, Zhao Li, Shuigeng Zhou, Xiangyu Zhang, and Daxin Jiang. Mixture-of-experts can surpass dense LLMs under strictly equal resource. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=oIdzliJAeA.

Kai Liu, Wei Li, Lai Chen, Shengqiong Wu, Yanhao Zheng, Jiayi Ji, Fan Zhou, Jiebo Luo, Ziwei Liu, Hao Fei, and Tat-Seng Chua. Javisdit: Joint audio-video diffusion transformer with hierarchical spatio-temporal prior synchronization. In *The Fourteenth International Conference on Learning Representations*, 2026a. URL https://openreview.net/forum?id=y7HV7KT3Bd.

Kai Liu, Yanhao Zheng, Kai Wang, Shengqiong Wu, Rongjunchen Zhang, Jiebo Luo, Dimitrios Hatzinakos, Ziwei Liu, Hao Fei, and Tat-Seng Chua. Javisdit++: Unified modeling and optimization for joint audio-video generation. In *The Fourteenth International Conference on Learning Representations*, 2026b. URL https://openreview.net/forum?id=hRRWfFpKRp.

Chetwin Low, Weimin Wang, and Calder Katyal. Ovi: Twin backbone cross-modal fusion for audio-video generation, 2025. URL https://arxiv.org/abs/2510.01284.

Jan Ludziejewski, Jakub Krajewski, Kamil Adamczewski, Maciej Pióro, Michał Krutul, Szymon Antoniak, Kamil Ciebiera, Krystian Król, Tomasz Odrzygóźdź, Piotr Sankowski, Marek Cygan, and Sebastian Jaszczur. Scaling laws for fine-grained mixture of experts. In *ICLR 2024 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2024. URL https://openreview.net/forum?id=Iizr8qwH7J.

Ang Lv, Ruobing Xie, Yining Qian, Songhao Wu, Xingwu Sun, Zhanhui Kang, Di Wang, and Rui Yan. Autonomy-of-experts models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=8BIDrYWCeg.

Ang Lv, Jin Ma, Yiyuan Ma, and Siyuan Qiao. Coupling experts and routers in mixture-of-experts via an auxiliary loss. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=MpeyjgWbKt.

Qianli Ma, Yaowei Zheng, Zhelun Shi, Zhongkai Zhao, Bin Jia, Ziyue Huang, Zhiqi Lin, Youjie Li, Jiacheng Yang, Yanghua Peng, et al. Veomni: Scaling any modality model training with model-centric distributed recipe zoo. *arXiv preprint arXiv:2508.02317*, 2025.

Shuailei Ma, Jiaqi Liao, Xinyang Wang, Jingjing Wang, Chaoran Feng, Zijing Hu, Chong Bao, Zichen Xi, Yuqi Gan, Weisen Wang, Yanhong Zeng, Qin Zhao, Zifan Shi, Wei Wu, Hao Ouyang, Qiuyu Wang, Shangzhan Zhang, Jiahao Shao, Yipengjing Sun, Liangxiao Hu, Lunke Pan, Nan Xue, Kecheng Zheng, Yinghao Xu, Xing Zhu, Yujun Shen, and Ka Leong Cheng. Scaling mixture-of-experts video pretraining for embodied intelligence. *arXiv preprint arXiv:2607.07675*, 2026.

MiniMax. MiniMax H3. GitHub repository, n.d. URL https://github.com/MiniMax-AI/MiniMax-H3/tree/main. Accessed September 25, 2026.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Evan Pete Walsh, Oyvind Tafjord, Nathan Lambert, Yuling Gu, Shane Arora, Akshita Bhagia, Dustin Schwenk, David Wadden, Alexander Wettig, Binyuan Hui, Tim Dettmers, Douwe Kiela, Ali Farhadi, Noah A. Smith, Pang Wei Koh, Amanpreet Singh, and Hannaneh Hajishirzi. OLMoe: Open mixture-of-experts language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=xXTkbTBmqq.

Taishi Nakamura, Satoki Ishikawa, Masaki Kawamura, Takumi Okamoto, Daisuke Nohara, Jun Suzuki, and Rio Yokota. Optimal sparsity of mixture-of-experts language models for reasoning tasks. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=XFw2EPRUUR.

OpenMOSS Team, Donghua Yu, Mingshu Chen, Qi Chen, Qi Luo, Qianyi Wu, Qinyuan Cheng, Ruixiao Li, Tianyi Liang, Wenbo Zhang, Wenming Tu, Xiangyu Peng, Yang Gao, Yanru Huo, Ying Zhu, Yinze Luo, Yiyang Zhang, Yuerong Song, Zhe Xu, Zhiyu Zhang, Chenchen Yang, Cheng Chang, Chushu Zhou, Hanfu Chen, Hongnan Ma, Jiaxi Li, Jingqi Tong, Junxi Liu,

Ke Chen, Shimin Li, Shiqi Jiang, Songlin Wang, Wei Jiang, Zhaoye Fei, Zhiyuan Ning, Chunguo Li, Chenhui Li, Ziwei He, Zengfeng Huang, Xie Chen, and Xipeng Qiu. MOVA: Towards scalable and synchronized video-audio generation, 2026. URL https://arxiv.org/abs/2602.08794.

Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27b dense model, April 2026. URL https://qwen.ai/blog?id=qwen3.6-27b.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2022.

Jay Shah, Ganesh Bikshandi, Ying Zhang, Vijay Thakkar, Pradeep Ramani, and Tri Dao. Flashattention-3: Fast and accurate attention with asynchrony and low-precision. *arXiv preprint arXiv:2407.08608*, 2024. URL https://arxiv.org/abs/2407.08608.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=B1ckMDqlg.

Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, Carleigh Wood, Ann Lee, and Wei-Ning Hsu. Meta Audiobox Aesthetics: Unified automatic quality assessment for speech, music, and sound. *arXiv preprint arXiv:2502.05139*, 2025.

Wan-AI. Wan2.2: Open and advanced large-scale video generative models. https://github.com/Wan-Video/Wan2.2, 2025. Accessed September 26, 2026.

Duomin Wang, Wei Zuo, Aojie Li, Ling-Hao Chen, Xinyao Liao, Deyu Zhou, Zixin Yin, Xili Dai, Daxin Jiang, and Gang YU. Universe-1: Unified audio-video generation via stitching of experts, 2026. URL https://openreview.net/forum?id=8aFYx2mDyE.

Yujie Wei, Shiwei Zhang, Hangjie Yuan, Yujin Han, Zhekai Chen, Jiayu Wang, Difan Zou, Xihui Liu, Yingya Zhang, Yu Liu, and Hongming Shan. Routing matters in moe: Scaling diffusion transformers with explicit routing guidance. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=1w1jCfYM8P.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-Align: Teaching LMMs for visual scoring via discrete text-defined levels. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54015–54029, 2024. URL https://proceedings.mlr.press/v235/wu24ah.html.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

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.

Youwei Zheng, Yuxi Ren, Xin Xia, Xuefeng Xiao, and Xiaohua Xie. Dense2moe: Restructuring diffusion transformer to moe for efficient text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18661–18670, 2025.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam M. Shazeer, and William Fedus. St-moe: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*, 2022. URL https://arxiv.org/abs/2202.08906.

A Appendix

A.1 Derivation of Constrained Routing

We formulate routing as the following constrained assignment problem:

$$\max_{\mathbf{x}} \sum_{i,j} s_{i,j} x_{i,j} \quad \text{s.t.} \quad \sum_{j} x_{i,j} = K,\ \forall i, \qquad \sum_{i} x_{i,j} \le \frac{mK}{N},\ \forall j, \tag{9}$$

where m is the number of tokens in the routing batch, N is the number of experts, and $x_{i,j} \in \{0,1\}$ indicates whether token i is assigned to expert j. The first constraint enforces that each token is routed to exactly K experts, while the second constrains the load of each expert to the balanced capacity mK/N.

Relaxed objective. We introduce auxiliary variables α_i for the per-token cardinality constraint and β_j for the per-expert load-balancing term. The resulting relaxed objective is

$$\mathcal{L}(x, \alpha, \beta) = \sum_{i,j} s_{i,j} x_{i,j} - \sum_{i} \alpha_i \left(\sum_{j} x_{i,j} - K \right) - \sum_{j} \beta_j \left(\sum_{i} x_{i,j} - \frac{mK}{N} \right). \tag{10}$$

Rearranging terms gives

$$\mathcal{L}(x, \alpha, \beta) = \sum_{i,j} (s_{i,j} - \alpha_i - \beta_j) x_{i,j} + K \sum_{i} \alpha_i + \frac{mK}{N} \sum_{j} \beta_j. \tag{11}$$

Objective after maximizing over x. For fixed (α, β), the objective is separable over the binary variables $x_{i,j}$. Hence, for each pair (i, j),

$$\max_{x_{i,j} \in \{0,1\}} (s_{i,j} - \alpha_i - \beta_j) x_{i,j} = \max(s_{i,j} - \alpha_i - \beta_j,\ 0).$$

Substituting this into Eq. (11) yields

$$g(\alpha, \beta) = \sum_{i,j} \max(s_{i,j} - \alpha_i - \beta_j, 0) + K \sum_{i} \alpha_i + \frac{mK}{N} \sum_{j} \beta_j. \tag{12}$$

Exact conditional update for α_i. Given the current β, the optimization over α decouples across tokens. For each token i, we have

$$g_i(\alpha_i) = \sum_{j} \max(s_{i,j} - \alpha_i - \beta_j, 0) + K\alpha_i.$$

Thus, α_i acts as a scalar threshold on the adjusted scores $\{s_{i,j} - \beta_j\}_{j=1}^{N}$. Its subgradient is

$$\frac{\partial g}{\partial \alpha_i} = K - \sum_{j} \mathbf{1}(s_{i,j} - \alpha_i - \beta_j > 0). \tag{13}$$

At the optimum of this conditional subproblem, α_i is chosen so that exactly K adjusted scores exceed the threshold. Equivalently, α_i is the threshold induced by the K-th largest value of $\{s_{i,j} - \beta_j\}_{j=1}^{N}$, i.e.,

$$\alpha_i = \text{Quantile}\left(\{s_{i,j} - \beta_j\}_{j=1}^{N},\ 1 - \frac{K}{N} \right). \tag{14}$$

Therefore, once β is fixed, each α_i can be recomputed exactly as a per-token top-K threshold. This is an exact solution of the conditional subproblem in α_i, rather than of the full joint optimization over (α, β). In the measure-zero tie case where multiple adjusted scores coincide at the threshold, ties can be broken arbitrarily so that exactly K experts are selected.

Iterative update for β_j. The variable β_j adjusts the expert-specific offset according to the mismatch between the target capacity and the realized expert load. Its derivative is

$$\frac{\partial g}{\partial \beta_j} = \frac{mK}{N} - \sum_i \mathbf{1}(s_{i,j} - \alpha_i - \beta_j > 0) = \frac{mK}{N} - \mathrm{load}_j, \tag{15}$$

where $\mathrm{load}_j = \sum_i \mathbf{1}(s_{i,j} - \alpha_i - \beta_j > 0)$ denotes the current load of expert j. Unlike α_i, which is recomputed exactly from the current adjusted scores, β_j is updated iteratively so that load-balancing information accumulates over training steps. Specifically, we use the sign-based update

$$\beta_j^{t+1} = \beta_j^t - \gamma \cdot \mathrm{sign}\left(\frac{mK}{N} - \mathrm{load}_j^t\right). \tag{16}$$

This update increases the offset for overloaded experts and decreases it for under-utilized ones.

Per-step procedure. At training step t, constrained routing proceeds as follows:

1. Read the current offsets $\{\beta_j^t\}_{j=1}^N$.
2. For each token i, compute $\alpha_i^t = \mathrm{Quantile}(\{s_{i,j} - \beta_j^t\}_{j=1}^N, 1 - K/N)$.
3. Route token i to expert j whenever $s_{i,j} - \alpha_i^t - \beta_j^t > 0$; in the tie case, break ties arbitrarily so that exactly K experts are selected for each token.
4. Compute the expert loads $\mathrm{load}_j^t = \sum_i \mathbf{1}(s_{i,j} - \alpha_i^t - \beta_j^t > 0)$.
5. Update each offset as $\beta_j^{t+1} = \beta_j^t - \gamma \cdot \mathrm{sign}(mK/N - \mathrm{load}_j^t)$.

In summary, α_i is recomputed exactly at each step as the threshold enforcing the per-token top-K constraint, while β_j is updated iteratively as an expert-specific balancing offset according to the discrepancy between the target and actual loads.

Shared-path scale approximation. Let each shared SwiGLU projection be represented as $W_\ell^{(s)} = \sum_{r=1}^{N_s} W_{\ell,r}$, where $\ell \in \{g, u, d\}$ denotes the gate, up, and down projections. If the component matrices are independent, zero-mean, and similarly scaled, each aggregate has standard deviation proportional to $\sqrt{N_s}$. For $W_d^{(s)}(\mathrm{SiLU}(W_g^{(s)}x) \odot W_u^{(s)}x)$, treating SiLU as locally scale-preserving gives a factor N_s from the multiplicative gate–up path and another $\sqrt{N_s}$ from the down projection, hence $N_s^{3/2}$. This is an initialization-scale approximation rather than an exact identity; correlations and nonlinear saturation can alter the observed ratio.

A.2 Variance Analysis of Router and Expert Initialization

We provide a simple variance analysis for the initialization choices discussed in Sec. 3.4.1. Consider a router or expert linear projection

$$y = Wx, \tag{17}$$

where $x \in \mathbb{R}^{d_{\mathrm{in}}}$ is the input token representation, $W \in \mathbb{R}^{d_{\mathrm{out}} \times d_{\mathrm{in}}}$ is the projection matrix, and $y \in \mathbb{R}^{d_{\mathrm{out}}}$ denotes the projected router logit or expert hidden activation. Assuming independent zero-mean inputs and weights, the output variance can be approximated as

$$\mathrm{Var}(y) \approx d_{\mathrm{in}} \,\mathrm{Var}(W)\,\mathrm{Var}(x). \tag{18}$$

This approximation allows us to compare how different initialization strategies affect the early scale of router scores and expert activations.

The fixed small-variance initialization is motivated by early load stability. Since the variance of W is manually set to a small value, the resulting router scores and expert activations have a narrow dynamic range at the beginning of training. This makes different experts appear relatively similar to the router, which can reduce early routing imbalance. However, the same property also weakens routing discrimination: token–expert scores become too close, and the model may require a longer period to form meaningful expert specialization.

Xavier initialization instead ties the weight variance to the input and output dimensions. As shown in Tab. 2, the resulting output variance remains on the same order as the input variance, up to the

Table 2: Comparison of initialization strategies for router and expert projections.

Initialization	Formula	Intuition	Observed effect
Small variance	$\mathrm{Var}(W) = \sigma^2$ $\mathrm{Var}(y) \approx d_{\mathrm{in}}\sigma^2\mathrm{Var}(x)$	Keep experts similar initially.	Balanced load, but weak specialization.
Xavier	$\mathrm{Var}(W) = \frac{2}{d_{\mathrm{in}}+d_{\mathrm{out}}}$ $\mathrm{Var}(y) \approx \frac{2d_{\mathrm{in}}}{d_{\mathrm{in}}+d_{\mathrm{out}}}\mathrm{Var}(x)$	Preserve a moderate signal scale.	Stable routing with sufficient diversity.
Orthogonal	$w_a^\top w_b = 0, \quad a \neq b$ w_a, w_b: rows of W	Separate expert directions early.	Faster differentiation, but more sensitive.

factor $\frac{2d_{\mathrm{in}}}{d_{\mathrm{in}}+d_{\mathrm{out}}}$. This provides a moderate signal scale: router scores are not collapsed into an overly narrow range, but they are also not separated too aggressively at initialization. In practice, this gives a better trade-off between early routing diversity and training stability.

Orthogonal initialization is motivated by encouraging different projection directions at the start of training. Let w_a and w_b denote two different rows of the projection matrix W. Orthogonal initialization encourages $w_a^\top w_b = 0, \quad a \neq b$, so different output channels start from less redundant directions. This can help early expert differentiation. However, in sparse MoE training, each expert only receives a subset of tokens. Thus, making expert projections separated too early can also make the model more sensitive to noisy token assignments and expert-specific gradient fluctuations.

Overall, Xavier initialization provides the most stable compromise in our audio-visual MoE setting. It avoids the under-discrimination of fixed small-variance initialization while being less sensitive to sparse expert-specific noise than orthogonal initialization. We therefore use Xavier initialization for both router and expert projections, together with sigmoid router activation as described in Sec. 3.4.1.

Softmax couples all expert scores through a shared normalization denominator, so increasing one expert's probability reduces those of the others. We use sigmoid to score experts independently before selection, reducing this direct competition before meaningful specialization emerges. Top-K selection and subsequent weight normalization still couple expert contributions.

A.3 Scope of the Empirical Capacity Fits

The curves in Figs. 3 and 4 are controlled development-scale ablations used to select the final recipe. Repeating every candidate through full-scale pretraining and the complete final-quality evaluation is computationally prohibitive; we therefore reserve full-scale training for the selected configuration. The figures consequently support architecture selection through development flow-matching loss and diagnostics, rather than claiming uncertainty-qualified final-quality differences for every candidate.

The activation-ratio and granularity fits reported in Sec. 3.2.1–3.2.2 use flow-matching loss as the matching criterion. We define $\mathrm{EL} = P_{\mathrm{dense}}^{\mathrm{match}}/P_{\mathrm{dense}}^{\mathrm{anchor}}$, where $P_{\mathrm{dense}}^{\mathrm{match}}$ is the dense capacity matching a given MoE configuration in flow-matching loss and $P_{\mathrm{dense}}^{\mathrm{anchor}}$ is the dense normalization anchor used throughout the controlled sweep. Comparisons use the same dataset, accelerator count, batch size, and random seed. They support loss-equivalent capacity at the development stage, not identical final perceptual quality.

The fits cover a finite configuration sweep ($K = 8$, $N \in \{8, 16, 32, 64, 128, 160\}$ and a matched granularity grid at fixed active-compute budget), not asymptotic scaling laws. The reported exponents are therefore descriptive of this training regime, and the optimum near $G \approx 8$–10 is consistent with the vertex of the reported quadratic fit.

A.4 Implementation Details

The audio DiT architecture follows ALIVE (Anonymous, 2026).

A.4.1 Model Specifications

Tab. 3 summarizes the configuration of the evaluated model.

Table 3: Model specifications. A denotes the number of activated parameters.

Item	Configuration
Model scale	32B total parameters, 4.7B activated parameters (32B-A4.7B); the active count is dominated by the video tower
Visual backbone	36 DiT blocks: 10 dual-stream blocks followed by 26 single-stream blocks
MoE placement	Visual FFNs; the first visual block remains dense
Routed experts	$N = 128$, top-K routing with $K = 8$ ($A = 6.25\%$)
Shared experts	$N_s = 1$
Expert FFN	SwiGLU with hidden width $d_{\text{expert}} = 1024$
Active expert width	$Kd_{\text{expert}} = 8192 = 4d$, with model width $d = 2048$ and granularity $G = 8$
Router	Sigmoid scoring, normalized top-K weights, and expert-specific load-balancing offsets
Audio branch	Dense DiT warm-started from the pretrained ALIVE audio module
Text conditioning	Dense FFNs using Qwen3.6-27B layer-32 features and an adaptor initialized from layers 28–32

A.4.2 Stability Hyperparameters

In the recorded continuation after a gradient-norm spike, gradient clipping with a threshold of 0.3 stabilized the maximum load violation while bias-based load regulation remained active. The dual-update step size for the expert load-balancing bias is $\gamma = 10^{-3}$.

Table 4: Development-stage evidence for routing and training stability. `avg_vio` is the training monitor's average load-violation statistic, utilization is the fraction of routed experts receiving traffic, and `max_vio` tracks the largest load-to-balanced-target ratio.

Comparison	Observed diagnostic	Interpretation and scope
Bias-based load regulation	`avg_vio`: $1.70 \to 0.09$; expert utilization: $35\% \to 100\%$	Direct evidence that the controller improves load coverage in the development run; it is not a final-quality metric.
Timestep shift $2.5 \to 9$	`avg_vio` rose transiently and then recovered; top-1 share decreased and the effective-expert count recovered	Supports distinguishing stage-induced expert redistribution from persistent routing collapse.
Post-spike continuation	Gradient clipping at 0.3 kept the maximum load violation stable	Supports gradient-norm clipping as the effective recovery intervention in the recorded run.

A.4.3 Training Data

Our audio-visual data follow the data construction procedure of ALIVE (Anonymous, 2026). We curate videos with audio from a large-scale video pool, covering speech, human interactions, and general sound-producing events. Processing includes visual-quality screening, structured captioning, audio-quality and audio-visual coherence filtering, speaker-identity correction, clarity assessment, and category balancing. Later stages use curated data for continued training and SFT, including real footage and high-aesthetic synthetic videos. The data volume used in each stage is reported in Tab. 5.

A.4.4 Structured Captions

We adopt the three-part caption schema of ALIVE (Anonymous, 2026): **Subjects** ($\mathcal{S}$) records entity identities, appearance, and voice attributes; **Visual Context** ($\mathcal{V}$) describes the scene; and **Interleaved Narration** ($\mathcal{N}$) orders actions, dialogue, and sound events. These components are serialized as

$$\mathcal{C} = [\texttt{Subjects:}\ \mathcal{S};\quad \texttt{Visual:}\ \mathcal{V};\quad \texttt{Narration:}\ \mathcal{N}]. \tag{19}$$

Table 5: Multi-stage training schedule. Rows are executed in order. Data volume is the number of samples used in a stage, with M denoting million; it is not a cumulative count across stages. Shift denotes the timestep shift (Esser et al., 2024) parameter.

Task	Stage	Resolution / FPS	Data volume	Epochs	Learning rate	Shift
T2V	Pre-train	192p / 16	200M	3.0	6×10^{-5}	1.5
T2V	Pre-train	480p / 16	44M	1.5	3×10^{-5}	9.0
T2V	Pre-train	480p / 16	44M	0.5	3×10^{-5}	5.0
T2V+TI2V	Pre-train	480p / 16	44M	0.5	3×10^{-5}	5.0
T2VA	Joint-train	480p / 24	11M	1.2	1×10^{-4}	3.0
T2VA+TI2VA	Joint-train	480p / 24	11M	0.3	5×10^{-5}	3.0
T2VA+TI2VA	Continue-train	480p / 24	4.3M	3.0	5×10^{-5}	3.0
T2VA+TI2VA	SFT	480p / 24	5M	1.0	1×10^{-5}	3.0

Within the narration, `<W>...</W>` marks spoken words and `<I>...</I>` marks non-speech sounds. Subject identifiers associate actions and speech with the corresponding entities.

A.4.5 Multi-stage Training Schedule

Tab. 5 summarizes the eight training stages. T2V and TI2V denote text-to-video and text-and-image-to-video generation, respectively; T2VA and TI2VA denote their audio-visual counterparts. The schedule first establishes video generation capability, then extends training to audio-visual tasks, followed by continued training and supervised fine-tuning (SFT).

Video Pretraining. We begin with T2V pretraining at 192p and 16 FPS on 200M samples for three epochs. The initial shift of 1.5 provides broad timestep coverage for learning denoising across the trajectory. On moving to 480p, we raise the shift to 9.0 and train on 44M samples for 1.5 epochs, emphasizing structural adaptation to the new token arrangement. We then reduce the shift to 5.0 for a further 0.5 epoch to increase the relative emphasis on detail refinement. A final 0.5-epoch pretraining stage combines T2V and TI2V at the same resolution and shift. The learning rate is 6×10^{-5} at 192p and 3×10^{-5} throughout 480p video pretraining.

Audio-Visual Training. We next use 480p videos at 24 FPS with a shift of 3.0. Training starts with T2VA on 11M samples for 1.2 epochs at a learning rate of 1×10^{-4}. We then introduce TI2VA alongside T2VA for another 0.3 epoch, reducing the learning rate to 5×10^{-5}. This stage extends the task coverage to both text-conditioned and text-and-image-conditioned audio-visual generation.

Continued Training and SFT. We continue training on 4.3M samples for three epochs at a learning rate of 5×10^{-5}, followed by SFT on 5M samples for one epoch at 1×10^{-5}. Both stages retain the T2VA and TI2VA task mixture, 480p resolution, 24 FPS, and a shift of 3.0.

A.4.6 Large-Scale Audio-Visual MoE Efficient Training

Parallelism Strategy. We combine FSDP2, Ulysses sequence parallelism Jacobs et al. (2023), and expert parallelism. FSDP2 shards parameters, gradients, and optimizer states across data-parallel ranks. DiT blocks use Ulysses SP to partition long sequences across GPUs. Each MoE module is independently FSDP2-wrapped with experts partitioned along the EP dimension, keeping all-to-all communication local and EP degree independently tunable. SP is selectively enabled only for long-sequence workloads. This balances memory efficiency, expert capacity, and throughput.

Sequence packing. Video samples in the training dataset can vary significantly in both duration and resolution, resulting in large variance in sequence length across iterations. To improve utilization, we enable a strategy called sequence packing. Given a budget of max sequence length, each iteration can adjust the batch size dynamically according to practical training dataset received from

dataloader. Longer sequences use a smaller batch size, while shorter sequences are grouped into larger batches, keeping the total workload of each iteration approximately balanced.

Memory Optimization. Memory consumption is a major bottleneck in large-scale MoE training. We employ two optimizations to reduce memory usage. Large intermediate tensors with long lifetimes can substantially increase peak activation memory. We therefore use fine-grained and lazy materialization. It can reduce both the size and the lifetime of live intermediate tensors. Our training extensively uses activation checkpointing to reduce memory footprint, but its effectiveness depends critically on checkpointing granularity. A checkpoint region that is too coarse can accumulate large intermediate activations before they can be released, whereas excessively fine-grained checkpointing incurs additional recomputation and checkpointing overhead. We therefore choose an intermediate granularity that balances peak memory and recomputation cost.

High-Performance Operators. We integrate FlashAttention-3 (Shah et al., 2024) for memory-efficient long-sequence attention, optimized RMSNorm kernels from QuACK (Dao-AILab, 2026) to reduce normalization overhead, and GroupedGEMM kernels from VeOmni (Ma et al., 2025) to accelerate MoE expert computation. These collectively reduce end-to-end training overhead for DiT-based MoE models.

MoE Monitoring. We attach lightweight forward hooks to each MoE router to collect expert assignments, routing probabilities, and top-k weights without retaining the computation graph. Metrics are synchronized across ranks to produce global utilization statistics, including load heatmaps, routing entropy, expert usage ratio, top-1 Gini, and top-k weight distribution. These diagnostics detect expert imbalance, router collapse, and degenerate routing during training.