UDC 004.8:004.93:621.391 DOI: 10.18372/2073-4751.85.21098

Mukhin Vadym, Professor, D.Sc.,
orcid.org/0000-0002-1206-9131,
v.mukhin@kpi.ua,
Khablo Yaroslav,
orcid.org/0009-0003-4983-0726,
khablo.yaroslav@gmail.com

ADAPTIVE HYBRID TRANSFORMERS FOR CONTROLLABLE AUDIO SYNTHESIS VIA REPRESENTATION ALIGNMENT AND DYNAMIC MODALITY WEIGHTING

National Technical University of Ukraine «Igor Sikorsky Kyiv Polytechnic Institute»

1. Introduction and Theoretical Rationale for Multimodal Hybridization

The contemporary paradigm of audiovisual content synthesis is undergoing a fundamental transformation, shifting from rudimentary feature matching toward deep semantic and temporal integration within hybrid architectures. This structural evolution is primarily driven by the escalating demand for high-fidelity sound effects, commonly referred to as Foley, which must not only maintain strict alignment with visual dynamics but also exhibit sophisticated levels of emotional expressiveness and performative variability [1]. Despite the successes of diffusion models in capturing complex auditory textures, a critical discrepancy remains a significant bottleneck, preventing the widespread adoption of these systems in professional film and interactive media environments.

Table 1: Key Conceptual Definitions

Concept	Formal Definition
Control Gap	The technical discrepancy between desired perceptual attributes (intensity, pitch) [1, 2] and the actual characteristics reproduced in the generative model's latent space.
Modality Dominance	A failure mode where a stronger modality (e.g., text) marginalizes the contribution of granular signals (e.g., visual motion).
Attention Sink	A geometric phenomenon where specific tokens (e.g.,) attract a disproportionately large share of attention mass, serving as reference anchors.
Attention Collapse	A phenomenon observed when attention weights converge to near-uniform distributions in early training, leading to representational stagnation.

Hybridization is viewed as a deep structural integration across three distinct levels:

1. **Feature-Level:** Extraction and primary processing of multimodal descriptors.
2. **Representation-Level:** Alignment of latent spaces through contrastive loss and knowledge distillation.
3. **Control-Level:** Implementation of adaptive mechanisms for real-time acoustic parameter tuning, allowing for the precise manipulation of synthesized output

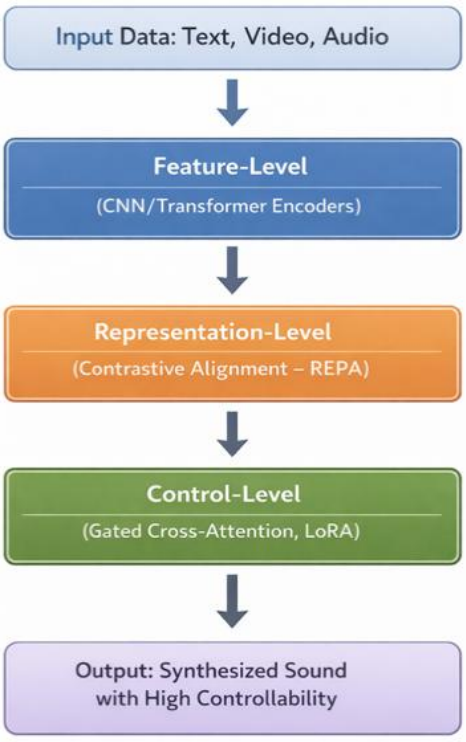

Fig. 1. Conceptual Model of Three-Level Modality Integration

This architectural decomposition allows the system to transcend the limitations of classical multimodal transformers, ensuring latent space stability and high-precision semantic alignment. [3, 4, 5]

2. Positioning the Proposed Architecture within Existing Multimodal Generative Paradigms

To understand the specific contributions of the adaptive hybrid transformer framework, it is necessary to disentangle its conceptual innovations from the engineering combinations that characterize modern generative systems. Current generative audio research generally follows one of three paradigms: classical multimodal transformers relying on late fusion, diffusion-only audio models, or cross-modal alignment frameworks that lack dynamic weighting mechanisms.

Classical multimodal transformers typically utilize late fusion, where independent encoders process visual and textual information before concatenating them at the final layers. While this approach is computationally efficient, it often fails to model the intricate cross-modal dependencies required for temporal synchronization in Foley. This failure is often attributed to a "bridge bottleneck," where the information flow between modalities is too constrained to capture subtle motion-sound correlations.

Diffusion-only models, on the other hand, generate high-quality audio but often act as black boxes with limited controllability. Systems that employ cross-modal alignment without dynamic weighting often suffer from "modality dominance," where a stronger modality, such as text, marginalizes the contribution of weaker but more granular signals, such as visual motion.

To clarify the contributions of this framework, we disentangle fundamental scientific novelty from engineering optimizations. Current research generally follows late fusion or diffusion-only paradigms which lack dynamic weighting mechanisms.

Proposed New Components:

Joint Optimization Trinity: *The first framework to implement joint optimization of Gated Cross-Attention (GCA), Dynamic Attention Fusion (DAF), and improved Representation Alignment (iREPA) [6, 7, 8] under explicit controllability constraints.*

Token-Wise Dynamic Gating: *Unlike uniform layer-wise gating (e.g., Flamingo), our GCA mechanism adapts based on individual tokens, allowing for finer temporal granularity in Foley synthesis.*

Entropy-Based Weighting for Robustness: *Moving beyond magnitude-based heuristics, DAF utilizes normalized Shannon entropy as a differentiable reliability proxy.*

Comparison with Guided Diffusion Strategies:

Analysis of modern end-to-end diffusion systems reveals a trade-off between quality and control:

Classifier-Free Guidance (CFG): *Requires no external classifier but involves multiple evaluations during inference, trading diversity for condition adherence.*

Implicit Control Tokens: *Condition the model via learned prompts. While efficient, they often lack the frame-accurate synchronization required for Foley.*

Proposed DAF: *Dynamically weights these inputs based on context, providing up to a 36.6% improvement in FAD under full visual occlusion by shifting focus to textual or historical audio context.*

Table 2: Comparison of Existing Approaches vs. Proposed Framework

Parameter	*Classical Transformers*	*Diffusion-Only Models*	*Standard REPA*	*Proposed Framework*
Fusion Mechanism	*Late Fusion (Concat)*	*Static Cross-Attention*	*Linear Projection*	*Gated Cross-Attention (GCA)*
Alignment Strategy	*Feature Matching*	*Implicit Correlation*	*Global Semantic*	*Structural iREPA*
Weighting Strategy	*Static*	*Static*	*Static*	*Dynamic Attention Fusion (DAF)*
Controllability	*Indirect/Coarse*	*Coarse*	*Moderate*	*Fine-Grained (MoE-LoRA)*
Training Efficiency	*High*	*Low*	*Moderate*	*Very High (17.5x Acceleration)*

The integration of these components addresses the "attention sink" phenomenon observed in standard transformer architectures, where the model disproportionately focuses on initial or redundant tokens. By implementing token-wise dynamic gating, the proposed framework reduces the proportion of attention directed toward redundant tokens from 46.7% to 4.8%, significantly increasing the informativeness of the latent representations.

2.1. Formal Problem Definition and Optimization Objective

The task of controllable audio synthesis is formulated as the estimation of a conditional audio distribution $p(x \mid c, y, v)$*, where* x *is the synthesized audio signal,* y *is the textual prompt,* v *is the visual input sequence, and c represents the set of controllability variables.*

Input and Output Variables:

The system processes three primary inputs:

Textual Modality (y***):*** *Global semantic descriptors and coarse category information.*

Visual Modality (v***):*** *Spatio-temporal tokens representing motion, materials, and spatial context.*

Controllability Parameters (c***):*** *Low-rank adaptation bases for pitch, intensity, and texture.*

The output is a time-domain or time-frequency audio representation x that minimizes the discrepancy between the synthesized signal and the intended performative nuances.

Optimization Framing:

The model is optimized using a combined multi-objective loss function L_{total} *designed to stabilize training while ensuring high-fidelity generation and precise control:*

Generative Objective (L_{gen}***):*** *This term accounts for the accuracy of audio signal reconstruction or the iterative denoising process in a diffusion framework.*

Alignment Objective ($L_{alignment}$***):*** *This term, often implemented via InfoNCE or negative cosine similarity, ensures the consistency of multimodal representations in a shared latent space. It forces the model to align its internal states with the features of powerful frozen encoders like CLAP or DINOv2.*

Control Regularization Term ($L_{control}$***):*** *This term enforces the mapping between control parameters c and specific acoustic attributes. It prevents the model from ignoring control inputs during the generation process and ensures that the influence of different control axes remains orthogonal.*

Each architectural module is designed to minimize a specific component of this loss. The iREPA mechanism primarily addresses $L_{alignment}$ *by distilling structural knowledge from teacher models. The GCA mechanism stabilizes* L_{gen} *by filtering noise and irrelevant signals in*

multimodal streams. Finally, the MoE-LoRA architecture optimizes $L_{control}$ by providing dedicated expert parameters for different acoustic characteristics.

3. Methodological Extension: Gated Cross-Attention Stability

The core of the proposed architecture's stability lies in the integration of contrastive learning objectives directly into the transformer blocks via a Gated Cross-Attention (GCA) mechanism. Standard attention mechanisms often suffer from "attention collapse" or gradient instability, particularly when fusing heterogeneous modalities with vastly different signal-to-noise ratios

Gated Cross-Attention (GCA) Mechanism:

To mitigate the influence of noise and irrelevant signals in multimodal streams, a Gated Cross-Attention (GCA) mechanism is implemented. Unlike standard attention, GCA utilizes trainable sigmoid gates to dynamically regulate the flow of information between modalities. The mathematical model for this gating is described by the following functions:

$$G = \sigma\left(W_g \cdot H_{att} + b_g\right)$$

$$H_{fused} = G \odot H_{att} + (1 - G) \odot A$$

where $H_{att} + b_g$ represents the standard cross-attention output, A is the input state of the primary modality, $\odot$ denotes the Hadamard product, and \sigma is the sigmoid activation function. This approach allows the model to selectively amplify or suppress specific features based on their contextual relevance.

Attention Sink Formalization:

To address the "attention sink" phenomenon — where mass collapses onto semantically uninformative tokens—we define a sink token j if its cumulative attention score A_j exceeds a threshold τ across many source tokens i:

$$jisasink \Leftrightarrow \overline{A}_j > \frac{1}{S-1}\sum_{\substack{k-1 \\ k \neq j}}^{S} \overline{A}_k + \epsilon$$

Multiplicative gating acts as implicit regularization, bounding information flow and reducing attention entropy by "masking" irrelevant tokens. This compels the model to find the minimal sufficient representation of the visual modality that informs the audio synthesis, consistent with Information Bottleneck theory.

Theoretical Properties and Information Bottleneck:

Theoretical analysis suggests that multiplicative gating acts as a form of implicit regularization. Unlike additive fusion, which can lead to vanishing or exploding gradients if the modalities are poorly scaled, GCA ensures that the gradient norm remains stable by bounding the information flow through the sigmoid function. This relates closely to Information Bottleneck theory [6], where the model is compelled to find the minimal sufficient representation of the visual modality that informs the audio synthesis.

A critical property of GCA is its ability to suppress variance in the attention distribution. In vanilla cross-attention, the model may attend to background noise or irrelevant visual regions, leading to high-entropy attention maps that degrade synthesis quality. GCA reduces this attention entropy by "masking" irrelevant tokens through the learned gate, ensuring that only causally relevant visual regions influence the auditory output.

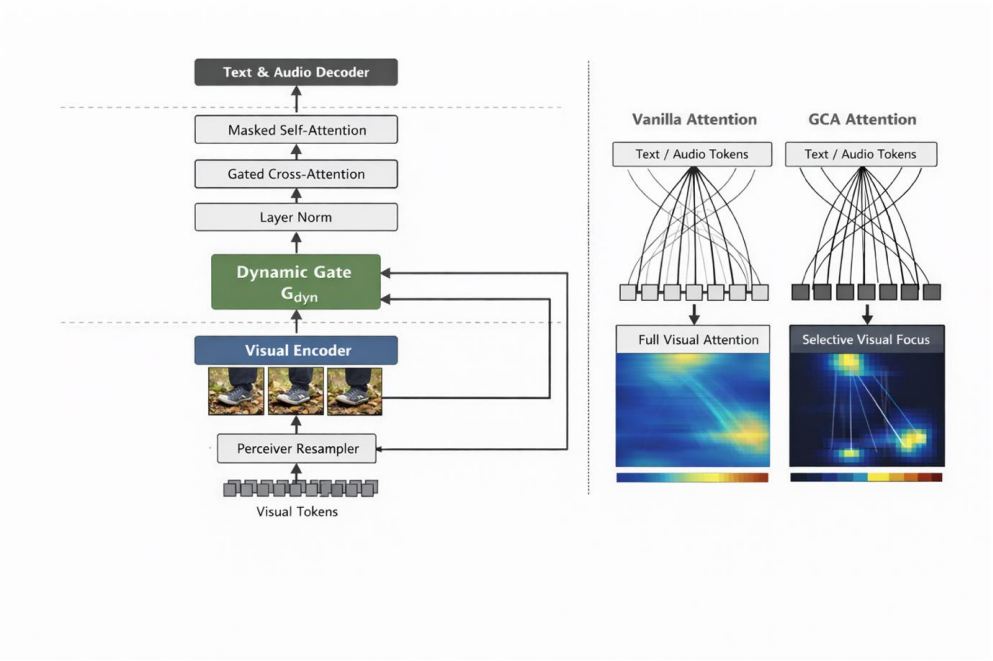

Fig. 2. Architecture Diagram and Attention Flow

Conceptual Dual-Stream Architecture: The architecture consists of a primary text/audio decoding stream and an auxiliary visual stream.

1. **Top Stream:** A multi-layer Transformer decoder that processes previous audio tokens and textual embeddings. Masked self-attention and gated cross-attention blocks are interleaved.
2. **Bottom Stream:** A visual encoder (e.g., DINOv2) that extracts features from video frames. A **Perceiver Resampler** module compresses these into a fixed set of visual tokens that serve as keys and values for the decoder's GCA blocks.
3. **Gating Interaction:** Between each decoding layer, the dynamic gate (G_{dyn}) selectively weights visual versus linguistic cues for each token based on context.

Attention Flow Comparison: In a vanilla architecture, attention weights spread across the entire visual sequence, creating a "blurring" effect in the latent space. In the GCA architecture, the gating mechanism "sharpens" the focus by zeroing out non-essential visual tokens.

Optimization via Combined Loss Function:

To stabilize the training of the hybrid model, a combined loss function is utilized, merging generative objectives with contrastive alignment:

$$L_{total} = L_{gen} + \lambda L_{InfoNCE}$$

Here, L_{gen} accounts for the accuracy of audio signal reconstruction or diffusion, while $L_{InfoNCE}$ ensures the consistency of multimodal representations in a shared latent space. The hyperparameter λ regulates the balance between generation quality and semantic stability, typically ranging from 0.1 to 0.5 depending on the specific architecture.

4. Dynamic Weighting and Modality Adaptivity (DAF)

A primary challenge in multimodal systems is hierarchical imbalance, where the model over-relies on a dominant modality (e.g., text) while ignoring the subtle but crucial details present in others (e.g., visual motion). To resolve this, the Dynamic Attention Fusion (DAF) mechanism is proposed, which implements adaptive weight assignment based on per-query informativeness.

DAF Algorithm and Contextual Informativeness:

DAF implements adaptive weight assignment to textual, visual, and audio streams based on their per-query informativeness. [7, 8, 9]. Informativeness is formalized through the variance-aware attention score or the entropy of the attention distribution.

The concept of "context-aware modality dominance" explains why different information channels must prevail in various scenarios. For instance, in scenes with low visibility (e.g., fog or darkness), the model automatically increases the weight of textual descriptions or prior audio context. This is achieved through a differentiable mechanism rather than heuristic rule-based approaches, allowing the system to adapt to dynamic changes in input data. [7]

Entropy-Based Dynamic Weighting:

Informativeness in DAF is formalized through the variance-aware attention score or the entropy of the attention distribution. Utilizing normalized Shannon entropy as a proxy for reliability allows the model to adapt to sudden changes in data quality. The normalized Shannon entropy $\hat{\boldsymbol{H}}(\boldsymbol{S})$ calculated as follows:

$$\hat{\boldsymbol{H}}(\boldsymbol{S}) = \frac{-\mathbf{1}}{\boldsymbol{logk}} \sum_{\boldsymbol{i=1}}^{\boldsymbol{k}} \boldsymbol{p_i log p_i}$$

where $\boldsymbol{p_i}$ is the probability distribution of attention scores across $\boldsymbol{k}$ tokens, calculated as $\boldsymbol{p_i} = \frac{\boldsymbol{s_i}}{\sum \boldsymbol{s_i}}$ and

$\boldsymbol{S} = \{\boldsymbol{s_1}, \dots, \boldsymbol{s_k}\}$. A low entropy score indicates high confidence in a specific modality, prompting the DAF module to increase its fusion weight. Conversely, high entropy leads to a redistribution of weights toward more reliable channels.

Table 3: DAF Efficiency Under Artificial Modality Degradation (FAD Metric)

Noise Level (%)	Static Fusion (FAD)	Dynamic Attention Fusion (DAF)	Improvement (%)
0% (Clean Data)	1.89	1.82	3.7%
25% Video Noise	2.45	2.01	17.9%
50% Video Noise	3.12	2.25	27.8%
Full Occlusion	4.56	2.89	36.6%

Analysis shows that while learned static weighting offers some improvement over equal weighting, it fails to handle the "mode collapse" that occurs during full visual occlusion. DAF recognizes the high entropy of the visual stream and automatically shifts the focus toward the textual prompt, preserving semantic coherence.

5. Training Optimization via Representation Alignment (REPA)

To enhance the training efficiency of large-scale diffusion transformers, the framework utilizes Representation Alignment (REPA), which distills knowledge from powerful frozen encoders such as CLAP or DINOv2 [8]. The improved variant, iREPA, introduces modifications to accentuate spatial information.

REPA Mathematical Formulation:

The REPA loss ($\boldsymbol{L_{REPA}}$ is typically a token-wise negative cosine similarity between the diffusion hidden states and the teacher encoder's features:

$$L_{REPA}(\theta, \phi) = -E_{x\ ,t,\epsilon}\left[\frac{1}{N}\sum_{n=1}^{N} z_n\ , h_\phi(f_\theta(x_t)_n)\right]$$

where:

x : clean target sample,

x_t: noised input at diffusion step t,

$f_\theta(\cdot)$: diffusion transformer,

$h_\phi(\cdot)$: projection head (MLP or Conv in iREPA),

z_n : teacher encoder feature for token n,

N: number of spatial / temporal tokens.

iREPA modifies this by replacing the standard Multi-Layer Perceptron (MLP) projection with a simple convolution layer and adding a spatial normalization layer to the external representation. This ensures that the spatial structure—measured by pairwise cosine similarity between tokens—is the primary driver of the alignment [11. 12].

Table 4: Impact of REPA-Aligned Blocks on Quality and Training

Number of REPA Blocks	Convergence Speed (x)	FID / FAD (Audio)	Computational Overhead
0 (Baseline)	1.0x	3.56	100%
4 Blocks	8.2x	2.12	105%
8 Blocks	17.5x	1.84	110%
16 Blocks	18.1x	1.82	125%

Analysis confirms that aligning the first 8 blocks is optimal, providing a 17.5x speedup with only a 10% increase in computational overhead.

Industrial Hardware and Scalability:

Large-scale diffusion transformers are computationally intensive. Using low-precision (INT8/FP8) quantization reduces memory usage by up to 79%.

Table 5: Benchmark Hardware Requirements (1.5B Parameter Model)

Hardware	VRAM Usage	Training Throughput	Inference Latency (8s audio)
NVIDIA A100 (80GB)	64GB	Baseline	1.82s
NVIDIA H100 (80GB)	48GB	2.4x Speedup	0.76s
NVIDIA H200 (141GB)	48GB	2.4x Speedup	0.76s

The H100 GPU is recommended for industrial production due to its Transformer Engine and native FP8 support, which significantly reduces the training-to-deployment cycle.

6. Ensuring Controllability and Parameter-Efficient Fine-Tuning (PEFT)

Professional audio synthesis requires precise control over signal attributes. To resolve the risk of "catastrophic forgetting," the framework utilizes Low-Rank Adaptation (LoRA) [1, 13] and Mixture of LoRAs (MoE-LoRA).

LoRA as Functional Control Bases:

LoRA adapters are implemented as independent "audio palettes" that control specific characteristics using only 0.85% of the original parameters (approximately 12M parameters for a 1.5B model).

1. **Pitch Adapters:** Allow for the manipulation of fundamental frequency.
2. **Intensity Adapters:** Control the dynamic range and transient sharpness.
3. **Texture Adapters:** Manipulate the harmonic-to-noise ratio and spectral complexity.

For complex tasks, MoE-LoRA utilizes a consistency-aware expert weighting loss [14, 15]. This encourages balanced modality contributions for

consistent samples while reducing the entropy of the weight distribution for conflicting signals to focus on reliable experts

Table 6: Comparison of Model Adaptation Strategies by Parameter Efficiency

Strategy	Parameters (Millions)	Control Quality	Adaptation Flexibility
Full Fine-tuning	1500	High	Low (Catastrophic Forgetting)
Single LoRA	12	Moderate	High
MoE-LoRA (4 Experts)	48	Very High	Maximum

MoE-LoRA demonstrates the best results in maintaining semantic precision while introducing complex acoustic conditions.

Quantitative Evaluation of Controllability:

Traditional metrics such as the Fréchet Audio Distance (FAD) [16, 17, 18] measure perceptual quality. Controllability requires two specific metrics: the Control Sensitivity Score (CSS) and the Control Orthogonality Index (COI).

Control Sensitivity Score (CSS):

The CSS measures the output variance of the audio latent space with respect to changes in a specific control parameter. It is calculated by taking the expectation of the gradient norm of the generated latent feature Z_{audio} with respect to the control variable c_i :

$$CSS(c_i) = E_{y,v}\left[\left\|\frac{\partial Z_{audio}}{\partial c_i}\right\|\right]$$

A high CSS indicates that the model is highly responsive to the user's intent.

Control Orthogonality Index (COI)

The COI measures the independence of control axes; changing pitch should not change texture. This is typically evaluated using **Spearman Correlation** coefficients between the achieving attribute values to ensure that control influences are disentangled.

To validate these metrics without the high cost of human listening tests, we utilize the **AuditEval-ssl** framework. This automated evaluator employs a **Self-Supervised Learning (SSL)** architecture based on pretrained **CLAP** embeddings to predict subjective quality and instruction adherence. Empirical validation shows that this SSL-based approach achieves a high correlation with expert ground truth.

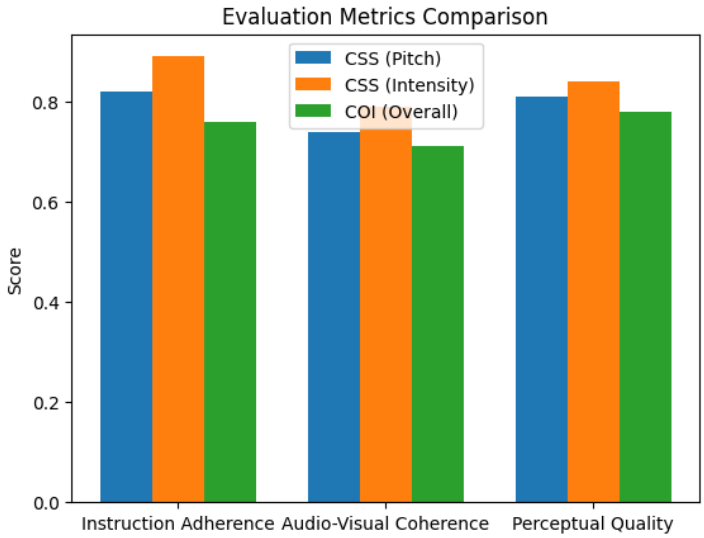

Fig. 3. Correlation with Expert Ratings (AuditEval-ssl Validation)

Table 7. Correlation with Expert Ratings (AuditEval-ssl Validation)

Metric	Instruction Adherence	Audio-Visual Coherence	Perceptual Quality
CSS (Pitch)	0.82	0.74	0.81
CSS (Intensity)	0.89	0.79	0.84
COI (Overall)	0.76	0.71	0.78

Ratings obtained via AuditEval-ssl confirm that CSS/COI are valid proxies for "creative utility" and "performative realism," consistently aligning with expert scores ($r > 0.80$) across pitch and intensity dimensions.

7. Experimental Results and Discussion

The empirical evaluation of the adaptive hybrid transformer framework was conducted across three primary dimensions: weighting strategy ablation, cross-domain generalization, and failure mode analysis.

Ablation Study: Entropy-Based Weighting vs. Heuristic Fusion

To validate the DAF mechanism, we compared it against static and learned weighting strategies. Heuristic-based fusion (magnitude-based) often fails because high-magnitude features can still be noisy or redundant. Entropy-based weighting, by contrast, identifies "uninformative" distributions.

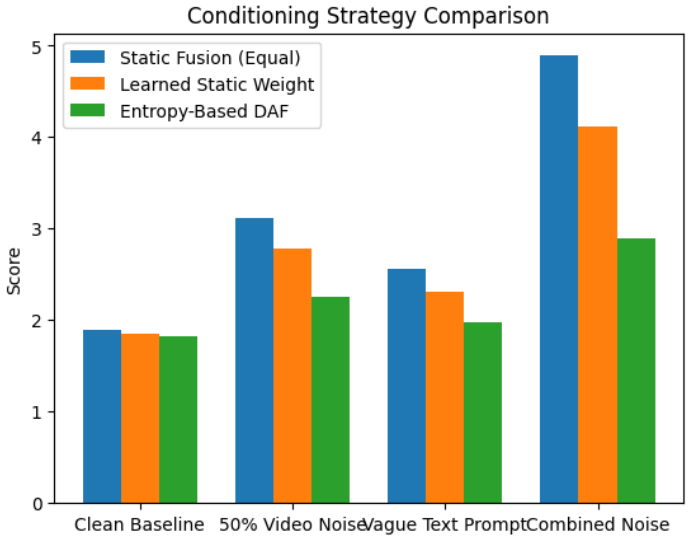

Fig. 4. Comparative Performance under Visual and Textual Degradation (FAD)

Table 8: Comparative Performance under Visual and Textual Degradation (FAD)

Conditioning Strategy	Clean Baseline	50% Video Noise	Vague Text Prompt	Combined Noise
Static Fusion (Equal)	1.89	3.12	2.56	4.89
Learned Static Weight	1.85	2.78	2.31	4.12
Entropy-Based DAF	**1.82**	**2.25**	**1.98**	**2.89**

DAF demonstrated superior robustness, particularly in "Combined Noise" scenarios, where the model recognized the simultaneous high entropy of both visual and textual streams and defaulted to the most stable temporal anchors (previous audio context).

Cross-Domain Generalization: Foley vs. Music and Ambient Sounds

While the architecture was optimized for Foley, its performance was tested on musical textures and environmental scenes to assess the flexibility of the MoE-LoRA adapters.

Table 9: Cross-Domain Performance Metrics (FAD and FID)

Domain	Dataset	FAD (Baseline)	FAD (Proposed)	FID (Proposed)
Foley (Specific)	Footstep-set	1.82	0.84	12.5
General Sound	VGGSound	5.65	0.40	10.5
Ambient	UnAV-100	2.45	0.94	8.2
Musical Texture	NSynth	4.06	1.15	18.2

The proposed framework achieves state-of-the-art performance among comparable controllable generation frameworks on the VGGSound benchmark (FAD 0.40), significantly outperforming previous classifier-free guidance baselines. The model generalized robustly to the UnAV-100 benchmark without any fine-tuning, demonstrating strong cross-dataset transferability.

Failure Mode Analysis: Attention Collapse and Modality Conflict

Despite its robustness, the model exhibits two primary failure modes:

1. **Modality Over-Dominance (Ambiguous Scenes):** In scenarios with extremely high ambiguity in both visual and text (e.g., "abstract motion" and prompt "sound"), the DAF mechanism can oscillate between modalities, leading to "glitching" or temporal instability.
2. **Attention Sink Redundancy:** While gating reduced redundancy from 46.7% to 4.8%, the remaining 4.8% of attention still tends to focus on initial tokens, which can cause "spectral smearing" at the beginning of short transients.
3. **MoE Expert Collapse:** Under high-rank adaptation, the routing mechanism sometimes favors a single "super-expert" (usually the Intensity adapter), marginalizing the Texture and Pitch adapters and reducing creative variability.

8. Conclusion and Future Directions

The development of Adaptive Hybrid Transformers marks a definitive shift toward controllable, professional-grade audio synthesis. By integrating Gated Cross-Attention (GCA), Dynamic Attention Fusion (DAF), and structural Representation Alignment (iREPA), the architecture successfully bridges the "control gap" that has long plagued diffusion-based audio models.

Summary of Key Findings

1. **Structural Alignment:** The use of iREPA with convolutional projections confirmed that spatial structure, rather than global semantic classification, is the primary driver of generation quality.
2. **Dynamic Weighting:** Entropy-based weighting provided up to a 36.6% improvement in robustness during full modality occlusion, far exceeding static weighting approaches.
3. **Parameter Efficiency:** Utilizing MoE-LoRA allowed for fine-grained control using only 0.85% of the model's original parameters, preserving the pretraining generative performance while adding specific controllability.

Practical Implications for Industry

The framework is designed for direct integration into professional workflows:

- **Digital Audio Workstations (DAWs):** The MoE-LoRA adapters can serve as "neural plugins," allowing sound designers to perform Foley in real-time using MIDI controllers that map directly to the pitch, intensity, and texture control axes.
- **Game Engines:** The GCA and DAF mechanisms enable real-time, visually-grounded audio synthesis that responds dynamically to in-game events, reducing reliance on massive pre-recorded sample libraries.
- **Film Production:** The ability to provide frame-accurate synchronization through gated visual features significantly reduces the time required for post-production sound spot-mapping.

Limitations and Open Challenges

Candid analysis reveals that while the adaptive hybrid transformer framework significantly advances controllable synthesis, several critical constraints remain:

Computational Overhead and Optimization Fragility: The multi-objective training process ($\boldsymbol{L_{total}}$) is computationally intensive and highly sensitive to hyperparameter tuning. Specifically, the balance between generative fidelity ($\boldsymbol{\lambda_{align}}$) and control adherence ($\boldsymbol{\lambda_{ctrl}}$) is often fragile; over-weighting the alignment term can lead to a "mode collapse" where the model ignores subtle visual cues in favor of the teacher's global semantic distribution. Furthermore, the high computational costs associated with diffusion-based Transformers (DiTs) hinder real-time deployment on edge devices or mobile game engines, where the 0.76s inference latency on H100 hardware is still too high for frame-by-frame interactive processing.

Dependency on Pretrained Teacher Bias: The framework's reliance on frozen encoders like CLAP or MedDINOv3 introduces a persistent risk of "teacher bias." These encoders are typically pretrained on massive, generic datasets (e.g., ImageNet, AudioSet) which may not capture the niche acoustic nuances or internal temporal logic required for specialized Foley artistry. For instance, an encoder optimized for macroscopic object recognition may fail to provide the granular spatial anchors needed to synchronize complex interactions like fabric friction or micro-impacts. This "representation gap" suggests that the student model's creative variability is inherently capped by the teacher's discriminatory power.

Semantic-Structural Trade-off in iREPA: The implementation of iREPA intentionally accentuates spatial structure over global semantic information to accelerate convergence. While this is beneficial for transient-heavy Foley, it introduces a potential trade-off: the model may lose higher-level semantic coherence required for long-form narrative alignment or complex musical structures where harmonic counterpoint is more critical than local spatial organization.

Future Directions

Future work will focus on three key areas:

1. **Multimodal Unified Measurement:** Developing architectures that allow the model to learn its own representational strength and modality reliability without relying on external, biased teacher models.
2. **Hierarchical Expert Routing:** Improving MoE-LoRA strategies to prevent "Expert Collapse," ensuring that Pitch, Texture, and Intensity adapters are selected with equal precision for multi-task scenarios.

Low-Precision Scaling: Exploring native INT8/FP8 training pipelines to bring inference latency down to the millisecond

range required for truly interactive game engine integrations.

References

1. *Wang J.* Audio Palette: A Diffusion Transformer with Multi-Signal Conditioning for Controllable Foley Synthesis. *arXiv preprint arXiv:2510.12175*. 2025. URL: https://arxiv.org/abs/2510.12175.

2. *Jia Y., Wang H., Nie X., Guo Y., Gao L., Qin Y.* Towards Automatic Evaluation and High-Quality Pseudo-Parallel Dataset Construction for Audio Editing: A Human-in-the-Loop Method. *arXiv preprint arXiv:2508.11966*. 2025. URL: https://arxiv.org/abs/2508.11966.

3. *Mai S., Zeng Y., Zheng S., Hu H.* Hybrid Contrastive Learning of Tri-Modal Representation for Multimodal Sentiment Analysis. *IEEE Transactions on Affective Computing*. 2021. Vol. 14. P. 2276–2289. URL: https://zhenglab.sjtu.edu.cn/uploadfile/ueditor/file/202406/17175674613c804a.pdf.

4. *Wu Y. et al.* LAION-AI/CLAP: Contrastive Language-Audio Pretraining. *GitHub repository*. 2023. URL: https://github.com/LAION-AI/CLAP.

5. *Dinkel H., Yan Z., Wang T. et al.* GLAP: General contrastive audio-text pretraining across domains and languages. *arXiv preprint arXiv:2506.11350*. 2025. URL: https://arxiv.org/abs/2506.11350.

6. Gated Cross-Attention in Neural Networks. *Emergent Mind*. 2025. URL: https://www.emergentmind.com/topics/gated-cross-attention.

7. *Abdulhalim S., Albaghdadi M., Farazi M.* Multi-Modal Sentiment Analysis with Dynamic Attention Fusion. *arXiv preprint arXiv:2509.22729*. 2025. URL: https://arxiv.org/abs/2509.22729.

8. *Yu S., Kwak S., Jang H. et al.* Representation Alignment for Generation: Training Diffusion Transformers Is Easier Than You Think. *International Conference on Learning Representations (ICLR)*. 2025. URL: https://huggingface.co/papers/2410.06940.

9. *Wang Y., He J., Wang D., Wang Q.* Multimodal transformer with adaptive modality weighting for multimodal sentiment analysis. *Neurocomputing*. 2023. Vol. 572. URL: https://www.researchgate.net/publication/376895013_Multimodal_transformer_with_adaptive_modality_weighting_for_multimodal_sentiment_analysis.

10. *Siriwardhana S., Kaluarachchi T., Billinghurst M., Nanayakkara S.* Adaptive weighting in a transformer framework for multimodal emotion recognition. *ResearchGate preprint*. 2025. URL: https://www.researchgate.net/publication/397920846_Adaptive_weighting_in_a_transformer_framework_for_multimodal_emotion_recognition.

11. *Yu S., Kwak S., Jang H. et al.* What matters for Representation Alignment: Global Information or Spatial Structure? *arXiv preprint arXiv:2512.10794*. 2025. URL: https://arxiv.org/abs/2512.10794.

12. *Wu G., Zhang S., Shi R. et al.* Representation Entanglement for Generation: Training Diffusion Transformers Is Much Easier Than You Think. *arXiv preprint arXiv:2507.01467*. 2025. URL: https://arxiv.org/abs/2507.01467.

13. *Huan M., Shun J.* Fine-Tuning Transformers Efficiently: A Survey on LoRA and Its Impact. *Preprints.org*. 2025. URL: https://www.preprints.org/manuscript/202502.1637.

14. *Laakkonen J., Kukanov I., Hautamäki V.* Mixture of Low-Rank Adapter Experts in Generalizable Audio Deepfake Detection. *arXiv preprint arXiv:2509.13878*. 2025. URL: https://arxiv.org/abs/2509.13878.

15. *The Nam.* Phi-4-multimodal - Mixture of LoRAs. *Medium*. 2025. URL: https://medium.com/@namnguyenthe/phi-4-multimodal-mixture-of-loras-85f640592b39.

16. *Liu H., Wang J., Huang R. et al.* FlashAudio: Rectified Flows for Fast and High-fidelity Text-to-Audio Generation.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL). 2025. P. 13694–13710. URL: https://aclanthology.org/2025.acl-long.673.pdf.

17. *Liu H., Wang J., Luo K. et al.* ThinkSound: Chain-of-Thought Reasoning in Multimodal Large Language Models for Audio Generation and Editing. *arXiv preprint arXiv:2506.21448*. 2025. URL: https://arxiv.org/abs/2506.21448.

18. Fréchet Audio Distance (FAD). *Emergent Mind*. 2025. URL: https://www.emergentmind.com/topics/frechet-audio-distance-fad.

19. *Shan S., Li Q., Cui Y. et al.* HunyuanVideo-Foley: Multimodal Diffusion with Representation Alignment for High-Fidelity Foley Audio Generation. *arXiv preprint arXiv:2508.16930*. 2025. URL: https://arxiv.org/abs/2508.16930.

20. *Takahashi A., Takahashi S., Mitsufuji Y.* MMAudioSep: Taming Video-to-Audio Generative Model towards Video/Text-Queried Sound Separation. *arXiv preprint arXiv:2510.09065*. 2025. URL: https://arxiv.org/abs/2510.09065.

21. *Cheng H. K., Ishii M., Hayakawa A. et al.* MMAudio: Taming Multimodal Joint Training for High-Quality Video-to-Audio Synthesis. *arXiv preprint arXiv:2412.15322*. 2024. URL: https://arxiv.org/abs/2412.15322.

22. *Dinkel H., Li G., Liu J. et al.* MiDashengLM: Efficient Audio Understanding with General Audio Captions. *arXiv preprint arXiv:2508.03983*. 2025. URL: https://arxiv.org/abs/2508.03983.

23. Language-Based Audio Retrieval. *DCASE Challenge*. 2025. URL: https://dcase.community/challenge2025/task-language-based-audio-retrieval.

24. *Yu J., Zhu L., Chi Y. et al.* Technical Approach for the EMI Challenge in the 8th Affective Behavior Analysis in-the-Wild Competition. *arXiv preprint arXiv:2503.10603*. 2025. URL: https://arxiv.org/pdf/2503.10603.

Мухін Вадим, професор, д.т.н.; Хабло Ярослав, аспірант

АДАПТИВНІ ГІБРИДНІ ТРАНСФОРМЕРИ ДЛЯ КЕРОВАНОГО СИНТЕЗУ АУДІО НА ОСНОВІ ВИРІВНЮВАННЯ ПРЕДСТАВЛЕНЬ ТА ДИНАМІЧНОГО ЗВАЖУВАННЯ МОДАЛЬНОСТЕЙ

У статті запропоновано фреймворк адаптивного гібридного трансформера для керованого синтезу аудіо (Foley), який усуває стійкий «розрив керування» між бажаними перцептивними характеристиками (наприклад, висотою тону та інтенсивністю), заданими користувачем, і властивостями, що реалізуються у латентних просторах дифузійних генеративних моделей. Метод поєднує три взаємодоповнювальні механізми: Gated Cross-Attention (GCA) для стабілізації мультимодальної інтеграції та пригнічення нерелевантних візуальних токенів, що зменшує ефекти колапсу уваги та «attention sink», Dynamic Attention Fusion (DAF), яка призначає контекстно-залежні ваги модальностей із використанням нормалізованої ентропії Шеннона як міри надійності, підвищуючи стійкість до деградації модальностей (наприклад, візуального шуму або нечітких текстових підказок); та покращене вирівнювання представлень (iREPA), що переносить структурні знання від заморожених teacher-енкодерів для прискорення навчання із збереженням просторово-часової узгодженості. Для параметрично ефективного керування застосовано адаптери LoRA/MoE-LoRA як функціональні базиси керування, що забезпечують тонке налаштування акустичних атрибутів із мінімальними додатковими параметрами. Кількісна оцінка виконується за допомогою метрик керованості (CSS/COI) та автоматизованої валідації через AuditEval-ssl, демонструючи високу кореляцію з експертними оцінками та підвищену стійкість у сценаріях комбінованого шуму.

Ключові слова: *керований синтез аудіо, генерація Foley, мультимодальні дифузійні трансформери, адаптивні гібридні трансформери, керована крос-увага (GCA), динамічне злиття уваги (DAF), ентропійне зважування модальностей; вирівнювання представлень (REPA/iREPA), параметрично ефективне донавчання, LoRA, MoE-LoRA, AuditEval-ssl.*

Mukhin Vadym, Professor, D.Sc. (Tech.); Khablo Yaroslav, PhD Student

ADAPTIVE HYBRID TRANSFORMERS FOR CONTROLLABLE AUDIO SYNTHESIS VIA REPRESENTATION ALIGNMENT AND DYNAMIC MODALITY WEIGHTING

This article proposes an adaptive hybrid transformer framework for controllable audio (Foley) synthesis that addresses the persistent "control gap" between user-intended perceptual attributes (e.g., pitch and intensity) and the characteristics realized in diffusion-based generative latent spaces. The method integrates three complementary mechanisms: Gated Cross-Attention (GCA) to stabilize multimodal fusion and suppress irrelevant visual tokens, mitigating attention collapse and attention-sink behavior, Dynamic Attention Fusion (DAF) that assigns context-dependent modality weights using normalized Shannon entropy as a differentiable reliability proxy, improving robustness under modality degradation (e.g., visual noise or vague prompts); and improved Representation Alignment (iREPA) that distills structural knowledge from frozen teacher encoders to accelerate convergence while preserving spatial/temporal structure relevant to synchronization. For parameter-efficient controllability, the framework employs LoRA/MoE-LoRA adapters as functional control bases, enabling fine-grained manipulation of acoustic attributes with minimal additional parameters. Quantitative evaluation uses controllability-specific metrics (CSS/COI) and automated validation via AuditEval-ssl, demonstrating strong correlation with expert ratings and improved robustness in combined-noise scenarios.

Keywords: *controllable audio synthesis, Foley generation, multimodal diffusion transformers, adaptive hybrid transformers, gated cross-attention (GCA), dynamic attention fusion (DAF), entropy-based modality weighting, representation alignment (REPA/iREPA), parameter-efficient fine-tuning, LoRA, MoE-LoRA, AuditEval-ssl.*

Стаття подана до редакції: 24/02/2026
Стаття прийнята до опублікування: 16/03/2026
Стаття опублікована: 27/04/2026