

AuraLuxMuse: Adaptive Fusion Modeling for Aesthetic Stage Lighting Design with Music and Expert Guidance

JUNYU DENG, Fudan University, China

JIALE CAO, Fudan University, China

MENGTIAN LI*, Fudan University, China and Shanghai University, China

ZHONGXIA JI, Shanghai Theatre Academy, China

RUHUA CHEN, Shanghai Theatre Academy, China

YIYI HE, Nanjing University of the Arts, China

GUANGNAN YE, Fudan University, China

ZUO HU*, Shanghai Theatre Academy, China

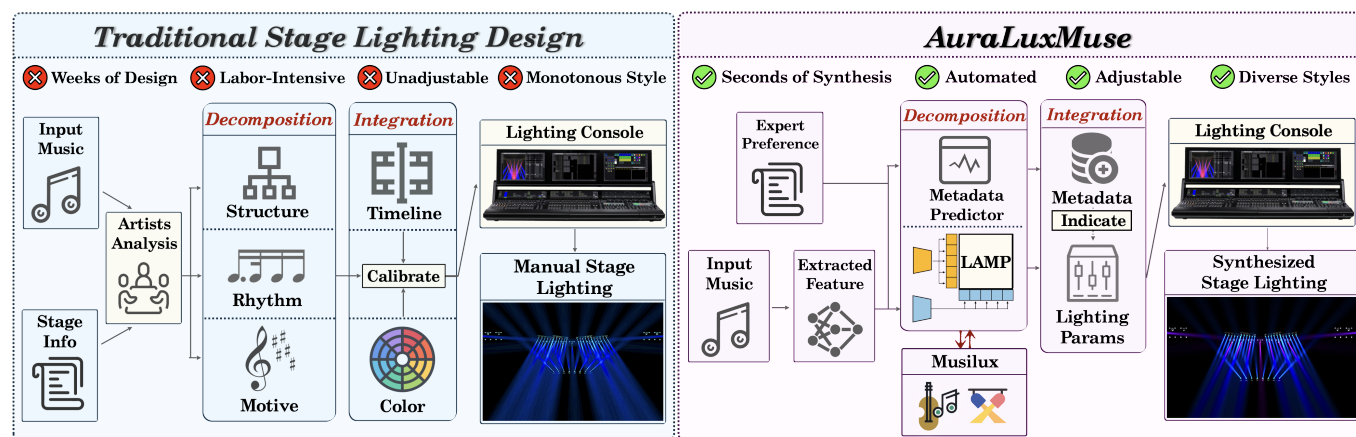


Fig. 1. **Procedure Comparison: Traditional Stage Lighting Design vs. AuraLuxMuse.** Conventional pipelines require designers to manually decompose music into artistic components, then iteratively design cues, align timing, and input parameters into consoles. AuraLuxMuse accelerates this workflow through feature extraction, preference-conditioned retrieval, cue adaptation, and console dispatch.

We present **AuraLuxMuse**, a novel system for automated aesthetic stage lighting design that integrates expert knowledge, representation learning, and preference-adaptive modeling. Lighting design in live performance settings requires the seamless translation of musical features into dynamic lighting behaviors. However, traditional workflows remain time-consuming, labor-intensive, and difficult to transfer. AuraLuxMuse encodes music and professional cue sequences into a shared retrieval space, estimates cue-event

*Mengtian Li and Zuo Hu are the corresponding authors.

Authors' Contact Information: Junyu Deng, Fudan University, Shanghai, China, 21307140005@m.fudan.edu.cn; Jiale Cao, Fudan University, Shanghai, China, 23300240001@m.fudan.edu.cn; Mengtian Li, Fudan University, Shanghai, China, Shanghai Film Academy and Shanghai University, Shanghai, China, mtli@fudan.edu.cn; Zhongxia Ji, Shanghai Theatre Academy, Shanghai, China, zhongxia.ji@sta.edu.cn; Ruhua Chen, Shanghai Theatre Academy, Shanghai, China, janechen2-c@my.cityu.edu.hk; Yiyi He, Nanjing University of the Arts, Nanjing, China, 125751790@qq.com; Guangan Ye, Fudan University, Shanghai, China, yegn@fudan.edu.cn; Zuo Hu, Shanghai Theatre Academy, Shanghai, China, huzuo@sta.edu.cn.



This work is licensed under a Creative Commons Attribution 4.0 International License. SA Conference Papers '26, Kuala Lumpur, Malaysia © 2026 Copyright held by the owner/author(s). ACM ISBN 979-8-4007-2842-6/2026/12 https://doi.org/10.1145/3829340.3842259

density, and retargets selected fixture commands to the destination stage. It assists pre-production authoring by returning editable cues rather than replacing the designer with an unconstrained generator. At the heart of AuraLuxMuse are two key modules: Lighting-Aligned Music Pretraining (LAMP), which performs contrastive learning between audio and lighting cues for alignment, and Preference-Adaptive Mixture of Experts (PAMoE), which conditions preference-aware cue retrieval and adaptation on designers' intent through a gated ensemble of style-specific expert networks. To support training and evaluation, we introduce **Musilux**, the first dataset of paired musical audio and professional lighting cue sequences under diverse performance scenarios. We evaluate AuraLuxMuse across both virtual simulation environments and professional-grade laboratories. Experimental results, including objective and subjective evaluation, demonstrate that AuraLuxMuse retrieves and adapts stage-lighting cues that are visually cohesive, semantically meaningful, and artistically expressive, showing its potential for AI-assisted aesthetic stage design.

CCS Concepts: • **Applied computing** → **Sound and music computing**; **Performing arts**; • **Computing methodologies** → *Computer vision representations*.

Additional Key Words and Phrases: Stage Lighting, Contrastive Learning, Audio Processing, Mixture of Experts

ACM Reference Format:

Junyu Deng, Jiale Cao, Mengtian Li, Zhongxia Ji, Ruhua Chen, Yiyi He, Guangnan Ye, and Zuo Hu. 2026. AuraLuxMuse: Adaptive Fusion Modeling for Aesthetic Stage Lighting Design with Music and Expert Guidance. In *SIGGRAPH Asia 2026 Conference Papers (SA Conference Papers '26)*, December 01–04, 2026, Kuala Lumpur, Malaysia. ACM, New York, NY, USA, 23 pages. <https://doi.org/10.1145/3829340.3842259>

1 Introduction

Stage lighting, as defined by professional lighting designer R. Craig Wolf [Wolf and Block 2013], refers to the art of using lighting techniques to illuminate the stage and shape the appearance of characters and scenery. It fulfills the artistic vision of designers by translating abstract concepts such as atmosphere, rhythm, and emotion into dynamic visual compositions in live performance. Stage lighting plays a vital role in the overall stage experience, engaging audiences through carefully orchestrated lighting behaviors. At the heart of it lies stage lighting design, which encompasses fixture behavior programming and real-time adjustment. An aesthetic lighting design must be technically feasible, artistically expressive, and adaptable to various performance scenarios.

Traditional lighting design [Graham 2018; Palmer 2015] begins with meticulous analysis of musical structure: segmenting phrases, detecting transitions, and extracting rhythmic elements. Crucially, designers must interpret the musical motive into a thematic concept that drives the formation of lighting behaviors, guiding the programming of individual fixtures that are aggregated into control parameters. This process is highly manual and iterative, often requiring multiple revisions informed by visual feedback, as illustrated in Fig. 1. While this handcrafted alignment of music, space, and light can yield emotionally resonant effects, it is inherently time-consuming and labor-intensive. Each lighting cue must be programmed from scratch, leading to low reusability and limited scalability. Moreover, aesthetic quality and structural consistency vary significantly between productions.

Consequently, there is a compelling need for a system capable of modeling and automating the entire stage lighting design process. However, existing machine-learning systems typically restrict lighting control to basic properties such as light position and intensity, overlooking critical expressive dimensions including color, stylistic variation, and temporal rhythm. To the best of our knowledge, this topic has not been systematically explored under modern deep-learning paradigms. With advances in deep audio representation learning, it is now feasible to encode the latent structure of music in high-dimensional embedding spaces, providing a promising foundation for music-driven lighting synthesis. Three key challenges remain. The first is **lack of structured parameterization**: there is no standardized formulation for decomposing lighting attributes into learnable parameters. The second is **multimodal alignment and personalization**: learning effective correspondences between music and lighting remains difficult, particularly when integrating personalized and stylistic preferences. The third is the **absence of comprehensive evaluation metrics**: existing benchmarks fail to assess lighting quality in terms of both objective fidelity and subjective perception.

To address these challenges, we propose **AuraLuxMuse**, a novel system integrating expert knowledge, multimodal representation, and preference-adaptive modeling for stage lighting design. As illustrated in Fig. 1, the pipeline operates directly on music and captures intrinsic properties. It uses these properties and a designer-specified preference to retrieve professional cue sequences, retargets their fixture commands to the destination stage, and dispatches the resulting cues and timecodes to the lighting console.

Specifically, AuraLuxMuse comprises two core modules: Lighting-Aligned Music Pretraining (**LAMP**) and Preference-Adaptive Mixture of Experts (**PAMoE**). At the core of cross-modal alignment, LAMP bridges the semantic gap between music and lighting. It encodes both audio and lighting parameters into a shared embedding space. This enables the system to retrieve the most semantically aligned lighting cues from a corpus, tailored to the corresponding musical content. To support stylistic control, PAMoE combines four learned MLP branches using both cue-dependent and preference-dependent gate weights. The branches are not preassigned to individual style axes; their observed associations with Color, Brightness, Zone, and Rhythm are analyzed after training. The resulting representation steers retrieval toward cues consistent with the designer's selected preference.

A professional lighting program comprises fixture-level *Cue* commands covering attributes such as intensity, color, pan, tilt, and gobo values, together with *Timecodes* that trigger cue sequences on a console. Given an input music track and preference vector $s = [s_c, s_b, s_z, s_r]$, the system encodes both as a retrieval query. The Metadata Predictor estimates the required cue density K , enabling the retriever to select the top- K professional cue sequences. Stage patching converts them into console-ready files.

We further introduce **Musilux**, a temporally aligned dataset containing 187 minutes and 22 seconds of music and more than 80,000 professional lighting instructions across real-world performance scenarios. Our key contributions are:

- We introduce **AuraLuxMuse**, a preference-conditioned retrieval and adaptation system whose outputs are competitive with manual designs in the evaluated scenarios, and **Musilux**, the first dataset to explicitly link professional lighting control parameters with musical audio.
- We develop **LAMP** for cross-modal learning between lighting and music, **PAMoE** for stylistic revision based on designers' intent, and evaluation diagnostic **EGSS** for retrieved and adapted stage-lighting outputs.
- We deploy AuraLuxMuse on virtual and real-world stages, producing editable candidate cues with an average inference time of 7.65 ± 0.98 **seconds** under the reported setup, as supported by qualitative and quantitative evaluations.

2 Related Work

Music-Driven Motion, Visual Generation, and Lighting Design. Music-conditioned synthesis has been explored for dance and articulated motion, including long-horizon 3D choreography from paired music-motion data [Li et al. 2021]. Related audio-to-visual work predicts visual dynamics from sound and context [Chatterjee and

Cherian 2020] or generates human-centric video conditioned on music [Zhao et al. 2024]. These tasks synthesize pixels or body motion and do not produce fixture-addressed, console-executable lighting cues. Lighting research instead spans rendering, design assistance, and illumination control, including simulation-based methods, natural illumination [Pellacini 2010], stage-lighting systems [Dorsey et al. 1995, 1991; Shimizu et al. 2019], virtual entertainment [El-Nasr and Horswill 2004; Kerr and Pellacini 2009; Pellacini et al. 2007], architectural and indoor illumination [Gardner et al. 2017; Jin and Lee 2019; Ren et al. 2023; Walch et al. 2020], and image-based relighting [Galvane et al. 2018; Zeng et al. 2024]. Existing lighting systems mainly formulate the problem as interactive design, physically based simulation, or objective-driven optimization, typically requiring designer-specified objectives, manual control, or predefined programs. To the best of our knowledge, prior work does not retrieve fixture-addressed, console-executable professional cue sequences directly from music while also conditioning the retrieval on designer preference and retargeting it to a new stage. Retrieval-Augmented Generation (RAG) [Lewis et al. 2021] is a powerful paradigm for injecting external knowledge or examples into generative models, with extensions beyond text: WavRAG [Chen et al. 2025] targets spoken dialogue, and TTA-RAG [Yang et al. 2024] improves text-to-audio synthesis in few-shot regimes.

Retrieval-based creative authoring reuses curated assets or examples rather than synthesizing every low-level element from scratch. RAG and its audio extensions [Chen et al. 2025; Lewis et al. 2021; Yang et al. 2024] provide the general retrieval paradigm; AuraLuxMuse applies the idea to a structured repository of expert-authored, console-valid cue sequences. We therefore cast stage-lighting design as preference-conditioned cross-modal retrieval and adaptation: musical context and user-specified expressive intent jointly query the cue repository for cues that are temporally coherent, rhythmically grounded, and stylistically aligned.

Audio Alignment. Convolutional Neural Networks have significantly advanced audio processing [LeCun et al. 1998]. MFCC remains a foundational time-frequency representation [Davis and Mermelstein 1980]. Among CNN-based architectures, VGGish [Hershey et al. 2017] is widely adopted for its transferable features, while transformer-based models such as SpecTNT [Lu et al. 2021] and AST [Gong et al. 2021] have demonstrated strong capacity for modeling spectrotemporal dependencies. Beyond unimodal learning, deep spectrum translation [Baradaran Kashani et al. 2019] and audiovisual localization [Chen et al. 2024] have broadened multimodal artistic applications. Other methods enhance robustness across domains [Wu et al. 2025]. Contrastive learning has emerged as a powerful approach for cross-modal representation, with CLIP [Radford et al. 2021] marking a key milestone. CLaMP [Wu et al. 2023] aligns symbolic music with language to enable robust music retrieval, motivating extensions such as AudioCLIP [Guzhov et al. 2021], CLAP [Wu et al. 2024], SLAP [Guinot et al. 2025] and MusCALL [Manco et al. 2022] to extend this paradigm to audio, supporting the retrieval of multiple modalities. CLMR [Spijkervet and Burgoyne 2021] leverages self-supervised contrastive learning on unlabeled data.

Modality alignment is essential for cross-domain design. We use audio as the guiding modality and extract high-level musical representations to model cross-modal correspondence. Inspired by contrastive multimodal learning, we introduce an alignment module that maps music and lighting into a shared latent space, enabling coherent interactions across modalities.

Designer Preference Adaptation. The Mixture of Experts (MoE) architecture [Jacobs et al. 1991] has gained widespread adoption because it decomposes complex demands into specialized subtasks. Sparse MoE layers have demonstrated scalability in NLP [Fedus et al. 2022; Shazeer et al. 2017] and image synthesis [Park et al. 2018]. Variational MoE autoencoders [Shi et al. 2019] have also been introduced for cross-modal generation. Furthermore, MoE has been adopted for audiovisual learning [Cheng et al. 2024] and style-specific text-to-speech synthesis [Jawaid et al. 2024], underscoring its potential in tasks requiring expressive generation.

As a creative task, stage lighting design inherently reflects the stylistic preferences of artists, necessitating a mechanism for incorporating personal preference and adapting model behavior accordingly. To this end, MoE provides a natural solution, enabling the model to dynamically select specialized expert pathways.

3 Method

3.1 System Overview and Problem Formulation

AuraLuxMuse is designed for offline pre-production, where artists can interact with and fine-tune stage-lighting effects. Its input is a music track together with an optional four-axis preference vector. Its output is a set of editable *Cue Sequences* and *Timecodes*, not rendered pixels. For example, given a concert excerpt and a high-rhythm preference vector, LAMP retrieves temporally compatible professional cues, the Metadata Predictor determines how many spatially complementary sequences are needed, PAMoE changes the preference-conditioned cue representation, and the patching procedure converts fixture identifiers and attributes to the target console layout. The objective is to retrieve and adapt temporally aligned professional lighting events conditioned on music and stylistic preference. Given a musical performance represented as a continuous feature sequence $M = \{M_t\}_{t=1}^T$ and an optional stylistic preference vector $s \in \mathbb{R}^{d_s}$, our objective is to generate a lighting sequence $L = \{(T_i, CS_i)\}_{i=1}^N$, where $T_i \in \mathbb{R}^+$ denotes the timecode at which a lighting cue is triggered, $CS_i \in \mathbb{R}^{d_c}$ represents the corresponding cue parameters retrieved from a collected cue corpus C , and N denotes the estimated total number of lighting events.

3.2 Structural Decomposition: Cues & Timecodes

The principled computational representation of stage lighting remains unexplored, with little prior work on decomposing its control parameters. This remains as a challenge. To tackle this, we systematically investigate the decomposition of stage lighting and propose a structured breakdown that formalizes computational representation. In industrial systems, stage lighting is parameterized through two fundamental components: *Cue Sequence* and *Timecodes*. Together, they orchestrate precisely synchronized lighting across performance. A *Cue Sequence* is defined as a structured list of lighting directives,

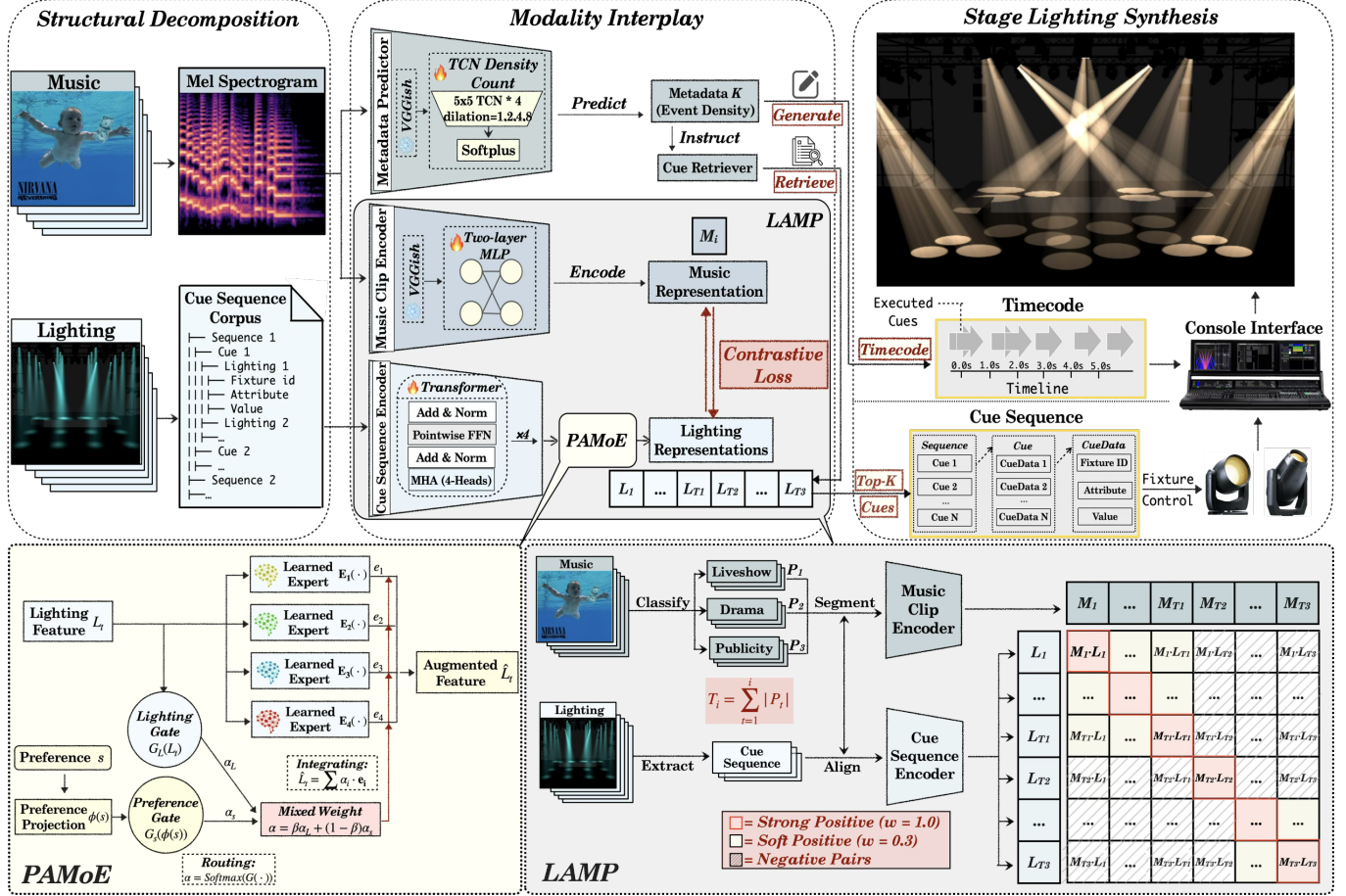


Fig. 2. **Overview of the AuraLuxMuse Pipeline.** Music and lighting data are individually preprocessed before entering LAMP for contrastive learning, and preference representations are injected via PAMoE. $P_1/P_2/P_3$ denote per-style frame counts, and $T_i = \sum_{t=1}^i |P_t|$ is the cumulative sample count used to index the illustrated segments. Cosine similarity and the metadata value K predicted by the non-causal TCN Metadata Predictor are used to query the cue corpus.

containing cues that govern the behavior of each lighting fixture. *Timecodes* serve as the temporal control system for cue execution. Each cue is annotated with a corresponding timestamp, ensuring the lighting transitions are executed with frame-level precision. A Cue Sequence comprises multiple Cues, each containing CueData units that encode control parameters for individual lighting fixtures. Each Timecode entry specifies the exact moment at which a particular Cue Sequence should be triggered. This data structure serves as the foundation underlying our proposed dataset detailed in Section 4.

3.3 Modality Interplay via Alignment: LAMP

The interplay between sonic and visual elements generates a synaesthetic experience in which light can be perceived as having acoustic properties and vice versa.

We formulate this relationship as a cross-modal contrastive learning problem. The interplay between music and lighting is fundamental to artistic expression. We therefore introduce Lighting-Aligned

Music Pretraining (**LAMP**), an architecture that represents lighting features and their corresponding music clips in a shared embedding space, as depicted in Fig. 2. We use only frames where lighting fixtures are active and construct each LAMP batch from paired raw music windows $\tilde{M}_i$ and temporally matched raw cue-sequence segments $\tilde{L}_j$, yielding $M_i = f_M(\tilde{M}_i)$ and $L_j = f_L(\tilde{L}_j)$ from the Music Clip Encoder and Cue Sequence Encoder. Beyond exact same-music/same-moment pairs, we include positive pairs defined by the scenario and style labels described in Section 4. By applying weighted alignment across these multiple positive matches, we optimize the contrastive loss in Equation (1).

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\sum_{p \in \mathcal{P}(i)} w_{i,p} \cdot \exp\left(\frac{\text{sim}(M_i, L_p)}{\tau}\right)}{\sum_{j=1}^N \exp\left(\frac{\text{sim}(M_i, L_j)}{\tau}\right)}. \quad (1)$$

Here, $\text{sim}(M_i, L_j)$ is cosine similarity in the shared embedding space.

$$\text{sim}(M_i, L_j) = \frac{M_i^\top L_j}{\|M_i\|_2 \|L_j\|_2}. \quad (2)$$

The functions f_M and f_L are the music and cue-sequence encoders. The positive set $\mathcal{P}(i)$ contains the original music–lighting pair as a strong positive ($w_{i,p} = 1$) and same-style, same-music segments from different moments as soft positives ($w_{i,p} = 0.3$). Here, N is the batch size and $\tau = 0.07$ is the softmax temperature; Supplementary Section E examines the weights.

3.4 Frame-wise Estimation: Metadata Predictor

The density of lighting events is directly related to the rhythmic pace of performances. Designers must calculate the appropriate frequency of lighting changes to support the narrative without overwhelming the audience.

To address, we propose a Metadata Predictor that estimates the expected number of lighting events in each temporal window, denoted by the metadata value K . This module learns a non-negative frame-level density, inspired by the crowd-counting paradigm [Lempitsky and Zisserman 2010; Zhang et al. 2015], in which lighting events are represented by density values rather than discrete timestamps. Formally, let $M_{s:s+W-1}$ denote a music-feature window of W frames beginning at frame s . The predictor outputs a density vector λ_s , whose sum gives the estimated event count for that window:

$$\lambda_s = \text{Softplus}(\text{TCN}_\phi(M_{s:s+W-1})), \quad (3)$$

$$\hat{c}_s = \sum_{\tau=0}^{W-1} \lambda_{s,\tau}, \quad K_s = \lceil \hat{c}_s \rceil. \quad (4)$$

The estimator runs offline and uses a non-causal TCN with symmetric padding. Consequently, an interior frame can use both preceding and subsequent context within the complete input window before the predicted density values are summed. The TCN uses dilated one-dimensional kernels, and Softplus enforces non-negativity. The window length W , stride, and learning rate are summarized in Section 5.1. This formulation offers three notable advantages aligned with stage-lighting prediction. It mitigates data sparsity by converting extremely sparse frame-wise labels into smooth temporal supervision, improving stability and optimization. It maintains physical validity because enforcing $\lambda_{s,\tau} \geq 0$ ensures non-negative event predictions. It also provides temporal awareness because window-level supervision over W frames captures short-term music–lighting dependencies.

3.5 Mechanism for Artists’ Specific Needs: PAMoE

Each lighting designer develops a recognizable visual vocabulary, so stage lighting embodies personal style and preference that must be explicitly modeled in any computational formulation.

To address this requirement, we incorporate personalized stylistic control into the system. In collaboration with professional lighting designers, we identify and annotate four ordinal stylistic dimensions: Color c , Brightness b , Zone z , and Rhythm r , with label definitions detailed in Supplementary Section C. The preference vector $s = [s_c, s_b, s_z, s_r]$ is used during joint LAMP–PAMoE training and can also be adjusted by a designer at inference to steer cue retrieval. To integrate this control signal, we propose Preference-Adaptive Mixture of Experts (**PAMoE**). As illustrated in Fig. 2, PAMoE is a dense soft mixture of four architecturally symmetric MLP branches, denoted as Learned Experts E_1 – E_4 . All four branches are evaluated

for each lighting sequence. The branches are not hard-assigned to exact style dimensions. Specialization is learned jointly and reported through the gate-sensitivity correlations in Supplementary Section C. Given a lighting-feature sequence $L = \{L_t\}_{t=1}^n \in \mathbb{R}^{n \times d}$ and preference vector $s = [s_c, s_b, s_z, s_r]^\top$, each ordinal value is represented by a 32-D embedding and the four vectors are concatenated into the 128-D preference projection $\phi(s)$. The Lighting Gate first performs sequence mean pooling and produces the data-dependent distribution $\alpha_L = \text{softmax}(G_L(L))$. Independently, the Preference Gate produces $\alpha_s = \text{softmax}(G_s(\phi(s)))$. A learned mixing coefficient β combines the two distributions as $\alpha = \beta\alpha_L + (1 - \beta)\alpha_s$. Each branch transforms every lighting token as $e_{k,t} = E_k(L_t)$, and PAMoE returns the augmented feature $\hat{L}_t = \sum_{k=1}^4 \alpha_k e_{k,t}$.

3.6 Adaptive Stage Lighting Retrieval

Now that we have addressed the core aesthetic challenges through our proposed modules, we integrate them into the unified system as illustrated in Fig. 2. During training, temporally matched music windows and cue-sequence segments are embedded jointly, and professional preference/style annotations supervise the PAMoE conditioning path. The aligned pairs are fed to LAMP for joint representation learning, while PAMoE injects the dense preference condition into expert-weighted lighting representations. At inference, music and the selected preference query the learned cue space rather than an unconstrained free-form generator. During inference, the LAMP encoder extracts high-level audio representations from an input music clip. The system ranks the Cue Sequence Corpus using the cosine similarity $\text{sim}(M_i, L_j)$ defined above and selects the top- K sequences. The cardinality K is dynamically determined by the Metadata Predictor. All K retrieved cue sequences are conflict-checked, retargeted to the destination stage, and dispatched to the lighting console for execution. The sequences are mutually distinct, each monitoring a specific spatial component of the stage illumination. The Lighting Encoder adopts a transformer-based [Vaswani et al. 2017] architecture with four MHA heads, and the Music Encoder is composed of a VGGish network [Hershey et al. 2017] with multilayer perceptrons, all selected based on ablation studies in Section 5.4.

4 Dataset: Musilux

We construct a multimodal dataset, **Musilux**, comprising music and lighting data. We collect Musilux collaboratively with leading institutions in stage lighting design, following the comprehensive construction procedure illustrated in Fig. 3. Musilux comprises 187 minutes and 22 seconds of musical audio material, accompanied by more than 80,000 associated lighting instructions.

Data Collection. The dataset captures a diverse range of real-world performance scenarios, categorized into the following three primary domains:

- **Concert.** Live performances in which lighting dynamically adapts to musical structure. This domain totals 4,666 seconds and includes Pop, Rock, Dance, Rap, and Electronic subtypes.
- **Drama.** Theatrical productions that employ lighting to enhance narrative and emotional engagement, totaling 3,307 seconds.

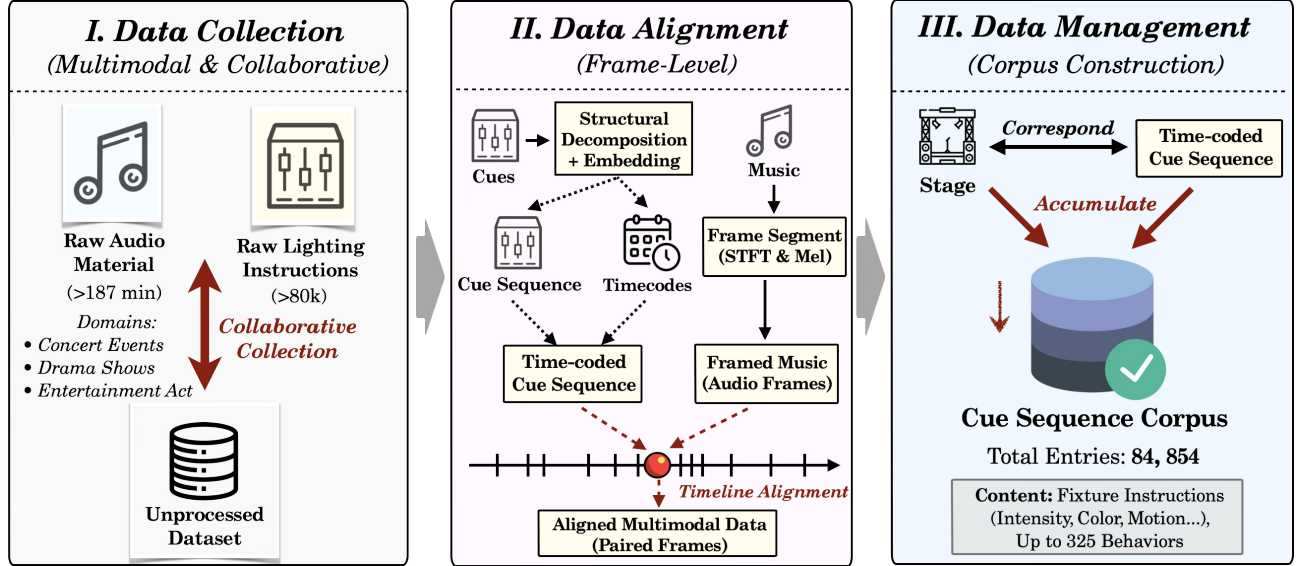


Fig. 3. **Musilux Construction Procedure.** Music and lighting cues are collected, aligned both within cue structures and across modalities, and then aggregated into the Cue Sequence Corpus that constitutes the dataset.

- **Entertainment.** Broadcasts, promotional events, and multimedia presentations in which lighting creates a specific atmosphere, totaling 3,269 seconds.

Musilux includes 13 distinct stages, each with multiple lighting shows (46.1% with one program, 30.7% with two, and 23.2% with more). This process provides a rich foundation for learning the underlying relationships between modalities.

Data Alignment. We unify Musilux at frame level by aligning each audio frame with its corresponding lighting instructions. Cue Sequences are first matched to annotated Timecodes for temporal information. We then perform frame-wise alignment between the music and time-stamped cues. Audio frames are transformed via Short-Time Fourier Transform (STFT) and projected onto the Mel scale, approximating human auditory perception. Concurrently, we parse each Cue Sequence into timestamps and attributes. Timestamps are mapped to the frame grid, while lighting attributes are embedded into continuous tensors via a dedicated mapping network.

Data Management. As lighting design is formulated as a retrieval-and-adaptation task, we accordingly construct *Cue Sequence Corpus* of 84,854 entries, each holding multiple CueData units that specify fixture-level instructions (intensity, color, motion). It covers 325 distinct fixture behaviors—color transitions, spatial movements, and speed modulations—forming a behavior-rich foundation.

5 Experiments

5.1 Experiment Setups

Configurations. All experiments were conducted on an NVIDIA GeForce RTX 4090 GPU. The Metadata Predictor was trained with Adam [Kingma and Ba 2017] at a learning rate of 0.001, using a total

window length of $W = 64$ frames and stride $S = 4$, achieving convergence within 100 epochs with an early-stopping patience of 8. LAMP was optimized for 500 epochs with learning rate 5×10^{-4} and weight decay 10^{-4} ; PAMoE was trained jointly on the aligned batches and professional preference labels. Supplementary Section E reports structural and hyperparameter ablations. Joint LAMP+PAMoE training takes about 1 hour; the Metadata Predictor takes 10–20 minutes; and the complete training pipeline takes approximately 1.0–1.5 hours. To assess performance across diverse scenarios, we use the disjoint validation/test protocol: validation provides manual references for Table 1, whereas the test set uses unseen music on an unseen stage. Evaluation covers virtual and physical environments. We render virtual lighting with GrandMA2 and MA3D, then validate on a physical console and stage.

Implementation. We incorporate a flexible patching mechanism that maps stage lighting between setups. It begins by accepting the retrieved and adapted cue $C_{out} = \langle p_{norm}, c_{req}, Attribute, val \rangle$ and a digital twin of the physical stage. The digital-physical address mapping defines a patch table for DMX fixture control.¹ The assignment cost combines normalized spatial distance with hardware and attribute mismatch; the Kuhn–Munkres solution maps commands to compatible fixtures before console dispatch. For different stage setups, the mechanism formulates lighting retargeting as a Linear Sum Assignment Problem, projects fixture coordinates into a normalized space, and builds a cost matrix from spatial distances and hardware mismatches. Solving the matrix with the Kuhn–Munkres algorithm yields the optimal permutation that aligns commands with the most appropriate devices, thereby aggregating fixture-level commands into stage-level transitions. The procedure also resolves conflicts

¹ $p_{norm} = (x, y, z)$ is the spatial fixture address; $c_{req} = [c_{pan}, c_{tilt}, c_{rgb}, c_{cmu}, c_{gobo}, \dots]$ encodes required hardware capabilities.

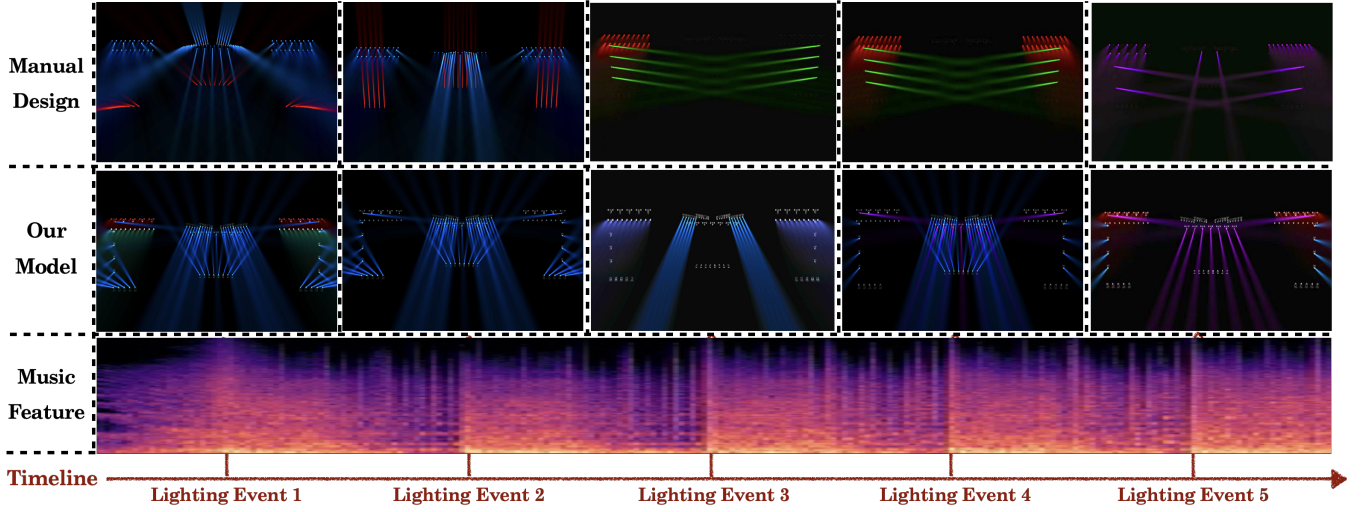


Fig. 4. **Qualitative Results.** AuraLuxMuse and the manual design exhibit different spatial and temporal choices in the same Drama scenario; these examples illustrate scenario-dependent competitiveness rather than universal superiority.

in which two or more retrieved commands assign different values to the same fixture–attribute pair within one execution window. Supplementary Sections D, F, and G provide the patching analysis, implementation procedure, console, and fixture details.

5.2 Evaluation Metrics

Objective: FID. We use Fréchet Inception Distance (FID) [Heusel et al. 2018] as a feature-space diagnostic of the distributional discrepancy between retrieved/adapted and reference cue-sequence representations. Because multiple plausible lighting designs can match the same music, FID is not treated as a definitive perceptual or aesthetic metric; lower values indicate closer distributions only in the selected encoder space, not necessarily better perceived quality.

$$\text{FID} = \|\mu_r - \mu_g\|^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}). \quad (5)$$

μ_r and Σ_r are the empirical mean and covariance of reference cue representations in the lighting-encoder feature space, while μ_g and Σ_g denote AuraLuxMuse outputs in that space.

Subjective: EGSS. To assess perceptual quality, we interviewed professional lighting designers and identified five expert-rated criteria, each scored on a 5-point Likert scale. We also collected general-audience feedback because readily observable impressions complement expert assessments.

We propose the Expert-Guided Subjective Score (EGSS), which combines expert and non-expert perspectives through the convex combination in Equation (6). Supplementary Section B defines the evaluation dimensions and standards.

$$\text{EGSS} = \lambda \cdot S_{\text{exp}} + (1 - \lambda) \cdot S_{\text{non}}, \quad \lambda \in [0, 1]. \quad (6)$$

Here, S_{exp} and S_{non} denote the average scores from experts and non-experts, respectively. We set λ to 0.7 to emphasize expertise while retaining audience perception, based on the ablation studies.

5.3 Human Evaluation

We conduct a user study comparing AuraLuxMuse with manually designed stage lighting across three scenarios. Forty expert participants and 40 general-audience participants rate full-length performance videos presented in randomized, anonymized order. Expert participants are selected based on specialized education and experience in large-scale productions. The expert group comprises the following two subgroups:

- **Student group (15%).** Participants are 21–24 years old, undergraduates and graduates with practical stage-lighting experience.
- **Teacher group (85%).** Participants are 35–45 years old, serve as stage-lighting design faculty, and hold advanced academic degrees related to lighting design and production.

The two groups exhibit different preferences. Defining the mean difference as AuraLuxMuse minus manual design, experts favor AuraLuxMuse in Concert (+1.075, $p = 0.00071$) and Entertainment (+0.748, $p = 0.025$), while the Drama difference is not significant (−0.366, $p = 0.260$). The expert-only cross-scenario average is a positive but non-significant trend (+0.486, $p = 0.079$). Non-expert ratings favor manual design on average (−0.431, $p = 0.0049$). Supplementary Section B provides the per-scenario breakdown.

5.4 Experiment Results

Qualitative Analysis. For qualitative assessment, we fix representative music frames and generate the corresponding stage lighting. As shown in Fig. 4, AuraLuxMuse closely matches manual cues in spatial composition and color dynamics. At certain timestamps, it demonstrates **greater visual complexity and temporal continuity**. We further present three model variations in Supplementary Section A. Representative stages across different stylistic dimensions under various music styles are presented in Fig. 5. The visualization

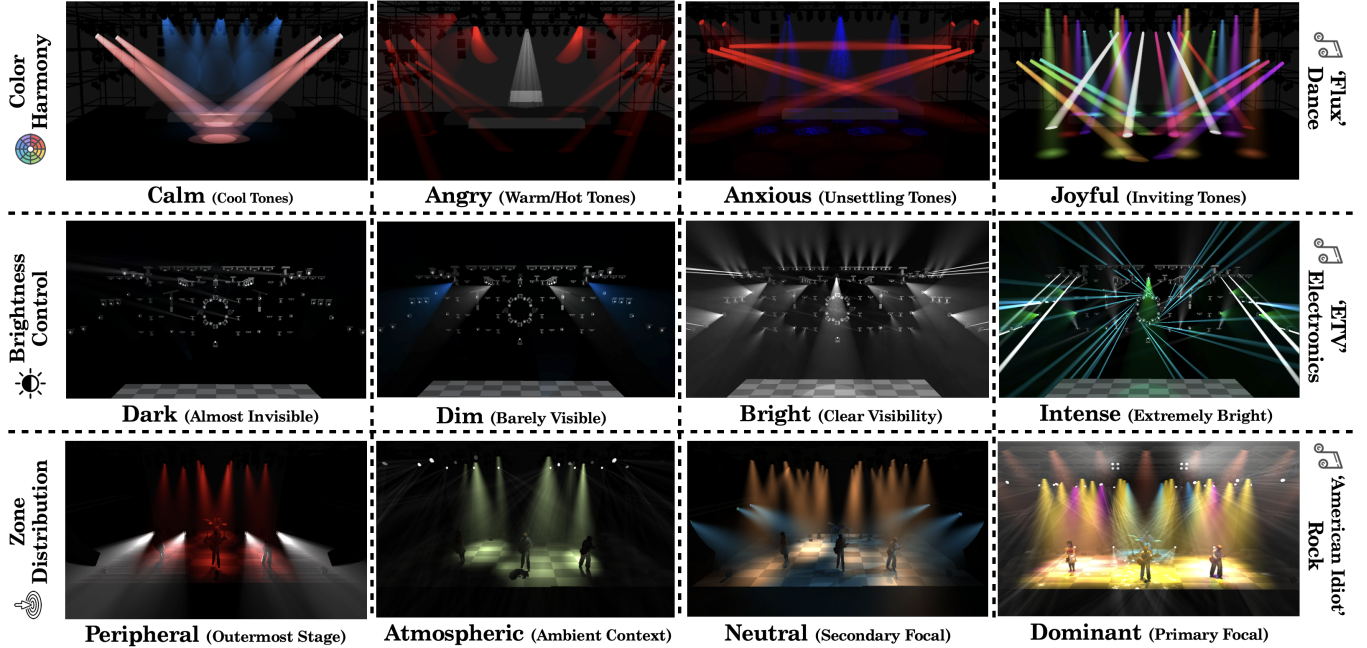


Fig. 5. **Stage Demonstrations across Lighting Styles.** Lighting configurations are visualized across varying levels of Color harmony, Brightness control, and Zone distribution. Definitions of stage-lighting terms are provided in parentheses.

illustrates preference-conditioned variation across artistic dimensions and music styles, making the abstract lighting dimensions visually interpretable.

Table 1. **Quantitative Results.** Average expert and non-expert ratings are reported for all scenarios. AuraLuxMuse achieves scenario-dependent competitiveness with the manual counterparts. Values in parentheses are standard deviations. FID measures the representational discrepancy between AuraLuxMuse and the corresponding reference.

Scenario	Source	FID	$S_{\text{exp}} \uparrow$	$S_{\text{non}} \uparrow$	EGSS $\uparrow$
Concert	Manual	8.21	2.58 (1.33)	4.31 (0.73)	3.09 (1.15)
	AuraLuxMuse		3.65 (1.00)	3.69 (0.88)	3.66 (0.75)
Drama	Manual	4.47	3.08 (1.10)	4.03 (0.93)	3.36 (1.04)
	AuraLuxMuse		2.71 (1.24)	3.27 (1.11)	2.88 (0.93)
Entertainment	Manual	2.00	2.46 (1.25)	3.31 (0.98)	2.71 (1.16)
	AuraLuxMuse		3.21 (1.25)	3.39 (1.13)	3.26 (0.94)
Average	Manual	4.89	2.71 (1.23)	3.88 (0.88)	3.05 (1.12)
	AuraLuxMuse		3.19 (1.17)	3.45 (1.05)	3.27 (0.88)

Quantitative Analysis. The model achieves an average inference time of 7.65 ± 0.98 seconds. Manual designs and AuraLuxMuse outputs are compared across all scenarios in Table 1. Across the evaluated scenarios, the expert-only comparison is scenario dependent: AuraLuxMuse is favored in Concert and Entertainment, while the Drama difference and the cross-scenario expert average are not statistically significant. These results indicate that the system is

competitive with manual designs in the evaluated scenarios while operating with substantially greater efficiency. In addition to comparisons with manual designs, we compare the similarity-based retrieval strategy with several alternative retrieval methods using both standard retrieval metrics and the real-world performance metric EGSS. As shown in Table 2, similarity retrieval performs best among the tested retrieval strategies. The rule-based baseline partitions the cue corpus using coarse audio heuristics before sampling valid professional cues. Random retrieval is a nontrivial corpus-prior lower bound because it also inherits metadata-predicted timing and samples console-valid cues. However, it lacks music-conditioned semantic alignment. Furthermore, we assess subjective performance across various music styles. We evaluate the model on styles present in Musilux to demonstrate its ability to reproduce known musical genres and test generalization on styles not included in the training set. Detailed results of these evaluations are presented in Table 3.

Table 2. **Comparison of Retrieval Strategies.** We compare similarity retrieval with random, position-based, frequency-based, nearest-neighbor, and rule-based strategies. Values in parentheses are standard deviations.

Strategy	FID $\downarrow$	Precision@5 $\uparrow$	mAP $\uparrow$	EGSS $\uparrow$
Random	6.32	0.07	0.03	2.71 (0.72)
Position-based	25.14	0.07	0.04	1.98 (0.73)
Frequency-based	16.32	0.07	0.03	2.44 (0.76)
Nearest-neighbor	21.65	0.08	0.04	2.10 (0.73)
Rule-based	6.61	0.06	0.02	2.81 (0.81)
Similarity-based	4.89	0.17	0.10	3.27 (0.88)

Table 3. **Subjective Results across Music Styles.** EGSS is reported for six validation styles and six test styles. Validation uses held-out examples with manual references; test tracks and the test stage are excluded from training.

(a) Validation Music Styles						
Music Styles	Pop	Rock	Opera	Dance	Rap	Electronic
EGSS	3.02 (1.23)	3.66 (0.75)	2.88 (0.93)	3.26 (0.94)	2.83 (1.04)	2.84 (1.31)
(b) Test Music Styles						
Music Styles	Funk	Disco	Jazz	Country	Folk	Classical
EGSS	3.25 (1.17)	2.94 (1.20)	3.51 (1.38)	2.98 (1.36)	2.70 (0.97)	3.05 (1.31)

Ablation Studies. We ablate the LAMP backbones [He et al. 2015; Vaswani et al. 2017], Metadata Predictor, and PAMoE; compare contrastive [Guinot et al. 2025; Guzhov et al. 2021; Wu et al. 2024] and generative [Lewis et al. 2019; Lu et al. 2021; Raffel et al. 2020; Wu et al. 2022] baselines; analyze K , soft/hard-alignment weights $w_{i,p}$, and lighting embeddings; and test metric sensitivity and the FID-EGSS correlation. Supplementary Section E provides the full designs, quantitative results, training dynamics, and sensitivity analyses.

5.5 Discussion

Across the evaluated scenarios, AuraLuxMuse is competitive with manual designs while producing candidate cues with an average inference time of 7.65 ± 0.98 seconds under the reported setup. It retrieves and adapts professional cues rather than replacing unconstrained human creativity. This approach outperforms the tested sequence-to-sequence generative baselines because small lighting-parameter perturbations can yield divergent effects and end-to-end models often fail to capture the constraints required for high-fidelity synthesis.

More broadly, our results show that an underexplored modality can be represented and integrated into a computational system, suggesting a path toward multimodal modeling of other unconventional modalities. AuraLuxMuse is, to the best of our knowledge, the first retrieval-and-adaptation system to combine **high quality, efficiency, and editable flexibility** in automated stage lighting design.

5.6 Limitations

AuraLuxMuse is an auxiliary-design system. It retrieves from a fixed expert-authored corpus and therefore cannot create fixture behaviors or artistic structures absent from that corpus. Music without a clear groove, script-only material without an audio timing signal, and rapidly changing styles remain difficult. Results do not establish population-scale generalization to all genres due to the size of validation and test set. The visual outputs can be over-bright or washed out. Person-following spotlights, narrative focus changes, and other high-level show decisions still require manual authoring and final approval. FID measures distance only in the selected encoder space and is an imperfect objective metric. The system supports efficient prototyping and inspection of candidate lighting productions; it does not replace professional judgment or creative work.

6 Conclusion

We introduce **AuraLuxMuse**, a multimodal system for intelligent stage lighting design via preference-conditioned retrieval and retargeting, together with the first professional lighting-music dataset Musilux. Our approach advances the field through a dedicated representation module that correlates music with lighting, a previously underexplored modality. PAMoE enables the injection of personalized preferences and the capture of underlying style characteristics. AuraLuxMuse marks a significant step toward AI-assisted stagecraft, offering both theoretical contributions to multimodal learning and practical value in real-world applications.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (Grant No. 62402306), the Natural Science Foundation of Shanghai (Grant No. 24ZR1422400), and the Shuguang Program of Shanghai Education Development Foundation and Shanghai Municipal Education Commission (Project No. 24SG50). This work is also supported by the Shanghai Universities Industry-Education-Research Collaborative Education Team, specifically the “AI + Performing Arts” Education Team.

References

- Hamidreza Baradaran Kashani, Ata Jodeiri, Mohammad Mohsen Goodarzi, and Iman Sarraf Rezaei. 2019. Speech Enhancement via Deep Spectrum Image Translation Network. arXiv:1911.01902 [eess.AS] <https://arxiv.org/abs/1911.01902>
- Maitreya Chatterjee and Anoop Cherian. 2020. Sound2Sight: Generating Visual Dynamics from Sound and Context. In *Computer Vision – ECCV 2020*. Springer International Publishing, Cham, Switzerland, 701–719. doi:10.1007/978-3-030-58583-9_42
- Liangyu Chen, Zihao Yue, Boshen Xu, and Qin Jin. 2024. Unveiling Visual Biases in Audio-Visual Localization Benchmarks. arXiv:2409.06709 [cs.MM] <https://arxiv.org/abs/2409.06709>
- Yifu Chen, Shengpeng Ji, Haoxiao Wang, Ziqing Wang, Siyu Chen, Jinzheng He, Jin Xu, and Zhou Zhao. 2025. WavRAG: Audio-Integrated Retrieval-Augmented Generation for Spoken Dialogue Models. arXiv:2502.14727 [cs.SD] <https://arxiv.org/abs/2502.14727>
- Ying Cheng, Yang Li, Junjie He, and Rui Feng. 2024. Mixtures of Experts for Audio-Visual Learning. In *Advances in Neural Information Processing Systems*, A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (Eds.), Vol. 37. Curran Associates, Inc., Red Hook, NY, USA, 219–243. https://proceedings.neurips.cc/paper_files/paper/2024/file/009729d26288b9a8826023692a876107-Paper-Conference.pdf
- S. Davis and P. Mermelstein. 1980. Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing* 28, 4 (1980), 357–366. doi:10.1109/TASSP.1980.1163420
- J. Dorsey, J. Arvo, and D. Greenberg. 1995. Interactive Design of Complex Time-Dependent Lighting. *IEEE Computer Graphics and Applications* 15, 2 (1995), 26–36. doi:10.1109/38.365003
- Julie O’B. Dorsey, François X. Sillion, and Donald P. Greenberg. 1991. Design and Simulation of Opera Lighting and Projection Effects. In *Proceedings of the 18th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH ’91) (SIGGRAPH ’91)*. Association for Computing Machinery, New York, NY, USA, 41–50. doi:10.1145/122718.122723
- Magy Seif El-Nasr and Ian Horswill. 2004. Automating Lighting Design for Interactive Entertainment. *Computers in Entertainment* 2, 2 (2004), 15. doi:10.1145/1008213.1008238
- William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity. arXiv:2101.03961 [cs.LG] <https://arxiv.org/abs/2101.03961>
- Quentin Galvane, Christophe Lino, Marc Christie, and Rémi Cozot. 2018. Directing the Photography: Combining Cinematic Rules, Indirect Light Controls and Lighting-by-Example. *Computer Graphics Forum* 37, 7 (2018), 45–53. doi:10.1111/cgf.13546
- Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gamberetto, Christian Gagné, and Jean-François Lalonde. 2017. Learning to Predict Indoor Illumination from a Single Image. arXiv:1704.00090 [cs.CV] <https://arxiv.org/abs/1704.00090>

- Yuan Gong, Yu-An Chung, and James Glass. 2021. AST: Audio Spectrogram Transformer. arXiv:2104.01778 [cs.SD] <https://arxiv.org/abs/2104.01778>
- Katherine Joy Graham. 2018. *Scenographic Light: Towards an Understanding of Expressive Light in Performance*. Ph. D. Dissertation. University of Leeds. <https://etheses.whiterose.ac.uk/id/eprint/20414/>
- Julien Guinot, Alain Riou, Elio Quinton, and György Fazekas. 2025. SLAP: Siamese Language-Audio Pretraining without Negative Samples for Music Understanding. arXiv:2506.17815 [cs.SD] <https://arxiv.org/abs/2506.17815>
- Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. 2021. AudioCLIP: Extending CLIP to Image, Text and Audio. arXiv:2106.13043 [cs.SD] <https://arxiv.org/abs/2106.13043>
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs.CV] <https://arxiv.org/abs/1512.03385>
- Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. 2017. CNN Architectures for Large-Scale Audio Classification. arXiv:1609.09430 [cs.SD] <https://arxiv.org/abs/1609.09430>
- Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2018. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. arXiv:1706.08500 [cs.LG] <https://arxiv.org/abs/1706.08500>
- Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. 1991. Adaptive Mixtures of Local Experts. *Neural Computation* 3, 1 (1991), 79–87. doi:10.1162/neco.1991.3.1.79
- Ahad Jawaid, Shreeram Suresh Chandra, Junchen Lu, and Berrak Sisman. 2024. Style Mixture of Experts for Expressive Text-to-Speech Synthesis. arXiv:2406.03637 [eess.AS] <https://arxiv.org/abs/2406.03637>
- Sam Jin and Sung-Hee Lee. 2019. Lighting Layout Optimization for 3D Indoor Scenes. *Computer Graphics Forum* 38, 7 (2019), 733–743. doi:10.1111/cgf.13875
- William B. Kerr and Fabio Pellacini. 2009. Toward Evaluating Lighting Design Interface Paradigms for Novice Users. In *ACM SIGGRAPH 2009 Papers* (New Orleans, Louisiana) (SIGGRAPH '09). Association for Computing Machinery, New York, NY, USA, Article 26, 9 pages. doi:10.1145/1576246.1531332
- Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs.LG] <https://arxiv.org/abs/1412.6980>
- Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-Based Learning Applied to Document Recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324. doi:10.1109/5.726791
- Victor Lempitsky and Andrew Zisserman. 2010. Learning to Count Objects in Images. In *Advances in Neural Information Processing Systems (NIPS)*, Vol. 23. Curran Associates, Inc., Red Hook, NY, USA, 1324–1332.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. arXiv:1910.13461 [cs.CL] <https://arxiv.org/abs/1910.13461>
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401 [cs.CL] <https://arxiv.org/abs/2005.11401>
- Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. 2021. AI Choreographer: Music Conditioned 3D Dance Generation with AIST++. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, Piscataway, NJ, USA, 13381–13392. doi:10.1109/ICCV48922.2021.01315
- Wei-Tsung Lu, Ju-Chiang Wang, Minz Won, Keunwoo Choi, and Xuchen Song. 2021. SpecTNT: A Time-Frequency Transformer for Music Audio. arXiv:2110.09127 [cs.SD] <https://arxiv.org/abs/2110.09127>
- Ilaria Manco, Emmanouil Benetos, Elio Quinton, and György Fazekas. 2022. Contrastive Audio-Language Learning for Music. arXiv:2208.12208 [cs.SD] <https://arxiv.org/abs/2208.12208>
- David Scott Palmer. 2015. *Light, Scenography and the Choreography of Space*. Ph. D. Dissertation. University of Leeds. <https://etheses.whiterose.ac.uk/id/eprint/11793/>
- David Keetae Park, Seungjoo Yoo, Hyojin Bahng, Jaegul Choo, and Noseong Park. 2018. MEGAN: Mixture of Experts of Generative Adversarial Networks for Multimodal Image Generation. arXiv:1805.02481 [cs.CV] <https://arxiv.org/abs/1805.02481>
- Fabio Pellacini. 2010. *envyLight: An Interface for Editing Natural Illumination*. *ACM Transactions on Graphics* 29, 4, Article 34 (2010), 8 pages. doi:10.1145/1778765.1778771
- Fabio Pellacini, Frank Battaglia, R. Keith Morley, and Adam Finkelstein. 2007. Lighting with paint. *ACM Trans. Graph.* 26, 2 (June 2007), 9–es. doi:10.1145/1243980.1243983
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models from Natural Language Supervision. arXiv:2103.00020 [cs.CV] <https://arxiv.org/abs/2103.00020>
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv:1910.10683 [cs.LG] <https://arxiv.org/abs/1910.10683>
- Haocheng Ren, Hangming Fan, Rui Wang, Yuchi Huo, Rui Tang, Lei Wang, and Hujun Bao. 2023. Data-Driven Digital Lighting Design for Residential Indoor Spaces. *ACM Transactions on Graphics* 42, 3, Article 28 (2023), 18 pages. doi:10.1145/3582001
- Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. arXiv:1701.06538 [cs.LG] <https://arxiv.org/abs/1701.06538>
- Yuge Shi, N. Siddharth, Brooks Paige, and Philip H. S. Torr. 2019. Variational Mixture-of-Experts Autoencoders for Multi-Modal Deep Generative Models. arXiv:1911.03393 [stat.ML] <https://arxiv.org/abs/1911.03393>
- Evan Shimizu, Sylvain Paris, Matt Fisher, Ersin Yumer, and Kayvon Fatahalian. 2019. Exploratory Stage Lighting Design using Visual Objectives. *Computer Graphics Forum* 38, 2 (2019), 417–429. doi:10.1111/cgf.13648
- Janne Spijkervet and John Ashley Burgoyne. 2021. Contrastive Learning of Musical Representations. arXiv:2103.09410 [cs.SD] <https://arxiv.org/abs/2103.09410>
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. arXiv:1706.03762 [cs.CL] <https://arxiv.org/abs/1706.03762>
- Andreas Walch, Michael Schwarzler, Christian Luksch, Elmar Eisemann, and Theresia Gschwandtner. 2020. LightGuider: Guiding Interactive Lighting Design Using Suggestions, Provenance, and Quality Visualization. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2020), 569–578. doi:10.1109/TVCG.2019.2934658
- R. C. Wolf and D. Block. 2013. *Scene Design and Stage Lighting*. Cengage Learning, Boston, MA, USA. <https://books.google.com/books?id=STIXAAAAQBAJ>
- Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. 2022. Autoformer: Decomposition Transformers with Auto-Correlation for Long-Term Series Forecasting. arXiv:2106.13008 [cs.LG] <https://arxiv.org/abs/2106.13008>
- Shangda Wu, Dingyao Yu, Xu Tan, and Maosong Sun. 2023. CLAMP: Contrastive Language-Music Pre-training for Cross-Modal Symbolic Music Information Retrieval. arXiv:2304.11029 [cs.SD] <https://arxiv.org/abs/2304.11029>
- Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Marianna Nezhurina, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2024. Large-Scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation. arXiv:2211.06687 [cs.SD] <https://arxiv.org/abs/2211.06687>
- Yi-Chiao Wu, Dejan Marković, Steven Krenn, Israel D. Gebru, and Alexander Richard. 2025. ComplexDec: A Domain-Robust High-Fidelity Neural Audio Codec with Complex Spectrum Modeling. arXiv:2502.02019 [eess.AS] <https://arxiv.org/abs/2502.02019>
- Mu Yang, Bowen Shi, Matthew Le, Wei-Ning Hsu, and Andros Tjandra. 2024. Audiobox TTA-RAG: Improving Zero-Shot and Few-Shot Text-to-Audio with Retrieval-Augmented Generation. arXiv:2411.05141 [eess.AS] <https://arxiv.org/abs/2411.05141>
- Chong Zeng, Yue Dong, Pieter Peers, Youkang Kong, Hongzhi Wu, and Xin Tong. 2024. DiLightNet: Fine-Grained Lighting Control for Diffusion-Based Image Generation. In *ACM SIGGRAPH 2024 Conference Papers (SIGGRAPH '24)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3641519.3657396
- Cong Zhang, Hongsheng Li, Xiaogang Wang, and Xiaokang Yang. 2015. Cross-Scene Crowd Counting via Deep Convolutional Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 833–841. doi:10.1109/CVPR.2015.7298684
- Zimeng Zhao, Binghui Zuo, and Yangang Wang. 2024. Music Conditioned Generation for Human-Centric Video. *IEEE Signal Processing Letters* 31 (2024), 506–510. doi:10.1109/LSP.2024.3358978

Supplementary Material

for AuraLuxMuse: Adaptive Fusion Modeling for Aesthetic Stage Lighting Design with Music and Expert Guidance

- Section A: Cross-Scenario Visualization of Lighting Effects
- Section B: Evaluation Dimensions and Standards
- Section C: Preference and Style Details
- Section D: Patching Mechanism Analysis
- Section E: Ablation Studies
- Section F: Stage Lighting Environments
- Section G: Stage Lighting Implementation Procedure
- Section H: Dataset Split and Generalization Analysis

A Cross-Scenario Visualization of Lighting Effects

To further illustrate the system, we present qualitative visualizations across multiple performance scenarios. Figures S1, S2, and S3 compare manual designs with AuraLuxMuse outputs for the three scenarios. These examples show music-responsive cue choices and scenario-dependent strengths, but they do not establish universal superiority over professional designers. Instead, AuraLuxMuse provides an editable retrieval-and-adaptation workflow that substantially reduces operating time.

In the illustrated Concert example, AuraLuxMuse selects cues that increase on-stage visibility for instrumentalists and differ stylistically from the manual design. In the Entertainment example, the retrieved cues emphasize vibrant and complex behavior. These are qualitative observations for the shown cases, not general claims about audience engagement or superiority.

B Evaluation Dimensions and Standards

B.1 Experts

Tables S1–S5 define the five expert evaluation dimensions—color harmony, brightness control, zone distribution, rhythmic tempo, and narrativity—and their rating standards.

Before rating, all professionals reviewed the definitions and standards. Each rated 12 scenes in five dimensions: one manual design and three AuraLuxMuse variants per scenario. Section C.1 details the dimensions used in evaluation.

B.2 Non-Experts

For non-expert participants, we designed a clear and accessible protocol covering rhythmic accuracy, thematic coherence, and narrative visual integrity. Table S6 details the criteria, which general-audience participants perceived as straightforward. For example, lighting actions synchronized with musical downbeats indicated strong rhythmic alignment. Audience members also assessed whether the lighting conveyed the music’s theme and remained stylistically consistent. All participants reviewed the definitions and standards before rating every scene in the overall study.

B.3 Human-Study Statistical Details

Values in parentheses in Tables 1–3 of the main paper are standard deviations. Defining each difference as AuraLuxMuse minus manual design, the expert results are: Concert $+1.075$ (95% CI [0.480, 1.671],

Table S1. **Lighting Evaluation for Experts: Color Harmony.** This dimension rates the aesthetic balance and blending of stage-lighting colors across the complete visual composition of each scene.

Ratings	Color Harmony
1	Harsh, unbalanced colors; poor harmony
2	Colors clash; mildly unbalanced
3	Colors blend acceptably; moderate harmony
4	Colors blend smoothly; strong harmony
5	Colors blend flawlessly; outstanding harmony

Table S2. **Lighting Evaluation for Experts: Brightness Control.** This dimension rates the consistency and adjustment of brightness across an entire performance sequence under review.

Ratings	Brightness Control
1	Uneven brightness; excessively dim/bright
2	Uneven brightness; slightly harsh
3	Reasonable brightness control; minor issues
4	Consistent brightness; well-adjusted
5	Perfect brightness; optimal balance

Table S3. **Lighting Evaluation for Experts: Zone Distribution.** This dimension rates evenness and coverage across the visible stage area.

Ratings	Zone Distribution
1	Uneven zones; inadequate coverage
2	Uneven zones; noticeable gaps
3	Fair zoning; slight inconsistencies
4	Even zoning; effective coverage
5	Flawless zoning; complete coverage

Table S4. **Lighting Evaluation for Experts: Rhythmic Tempo.** This dimension rates lighting synchronization with performance rhythm.

Ratings	Rhythmic Tempo
1	No rhythm; disorganized timing
2	Weak rhythm; moderately off
3	Average rhythm; reasonably aligned
4	Good rhythm; accurately timed
5	Excellent rhythm; impeccably timed

$p = 0.00071$), Drama -0.366 ([−1.017, 0.286], $p = 0.260$), Entertainment $+0.748$ ([0.098, 1.398], $p = 0.025$), and the cross-scenario expert average $+0.486$ ([−0.059, 1.031], $p = 0.079$). Non-expert descriptive results are Concert 3.69 ± 0.88 versus 4.31 ± 0.73 (difference -0.62), Drama 3.27 ± 1.11 versus 4.03 ± 0.93 (difference -0.76), and Entertainment 3.39 ± 1.13 versus 3.31 ± 0.98 (difference $+0.08$); the cross-scenario non-expert difference is -0.431 (95% CI [−0.723, −0.138],

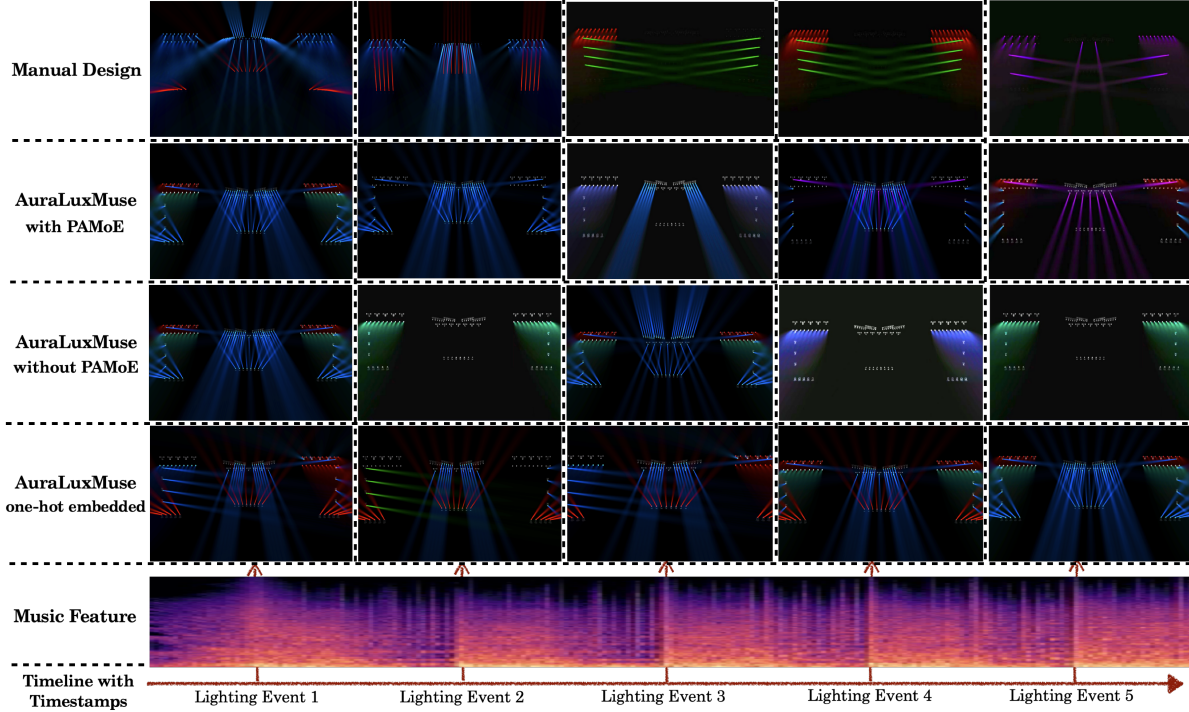


Fig. S1. **Qualitative Results for AuraLuxMuse Variants in the Drama Scenario.** Three architecture variants produce distinct retrieved/adapted lighting outputs as editable references for professional designers.

Table S5. **Lighting Evaluation for Experts: Narrativity.** This dimension rates how effectively lighting conveys plot and emotion.

Ratings	Narrativity
1	Fails to convey plot / emotion
2	Weakly conveys plot / emotion
3	Moderately conveys plot / emotion
4	Clearly conveys plot / emotion
5	Strongly conveys plot / emotion

$p = 0.0049$). EGSS is likewise scenario dependent: AuraLuxMuse is higher in Concert (3.66 ± 0.75 versus 3.09 ± 1.15 , $p = 0.010$) and Entertainment (3.26 ± 0.94 versus 2.71 ± 1.16 , $p = 0.024$), while manual design is higher in Drama (2.88 ± 0.93 versus 3.36 ± 1.04 , $p = 0.044$); the overall EGSS difference is not significant ($p = 0.282$). Based on paired ratings, 72.5% of experts and 40.0% of non-experts give AuraLuxMuse a higher overall rating. These results motivate the narrower claim of scenario-dependent competitiveness.

C Preference and Style Details

C.1 Style Dimensions Specification

This section presents the stage-lighting style dimensions and their corresponding attributes, as summarized in Table S7. The attribute definitions and assignment standards follow. Professional lighting designers assign one ordinal value for each of Color, Brightness,

Zone, and Rhythm to each labeled source program using the style taxonomy in Table S7 and the detailed level definitions in Tables S8, S9, and S10. The 33 available labeled programs provide the tuples $s = [s_c, s_b, s_z, s_r]$ used in the gate analysis. Each value is represented by a 32-D embedding, and the four vectors form the 128-D condition $\phi(s)$. During joint training, the tuple accompanies every music-cue window from that program; EGSS is never used as a training label. Inter-rater agreement was not measured for these style annotations, which limits reliability claims based on the labels.

C.2 Style Details

Color harmony is a foundational stylistic dimension in stage-lighting design because it directly shapes the viewer’s emotional perception and interpretation of a performance. Rooted in color theory and psychological aesthetics, color harmony describes the visual compatibility of color combinations: how well different hues, saturations, and brightness levels cohere within a lighting composition. We operationalize color harmony by embedding its semantics in a color-emotion mapping space informed by color psychology and cinematographic lighting design.

Brightness control is a critical dimension for evaluating and guiding stage-lighting design, offering both subjective stylistic interpretation and objective physical quantification. Lighting experts possess an intuitive grasp of illuminance ranges that aligns with defined lux intervals used in lighting-design standards. As shown in Table S8, Brightness is categorized into five discrete levels associated with distinct perceptual descriptions and lux ranges.

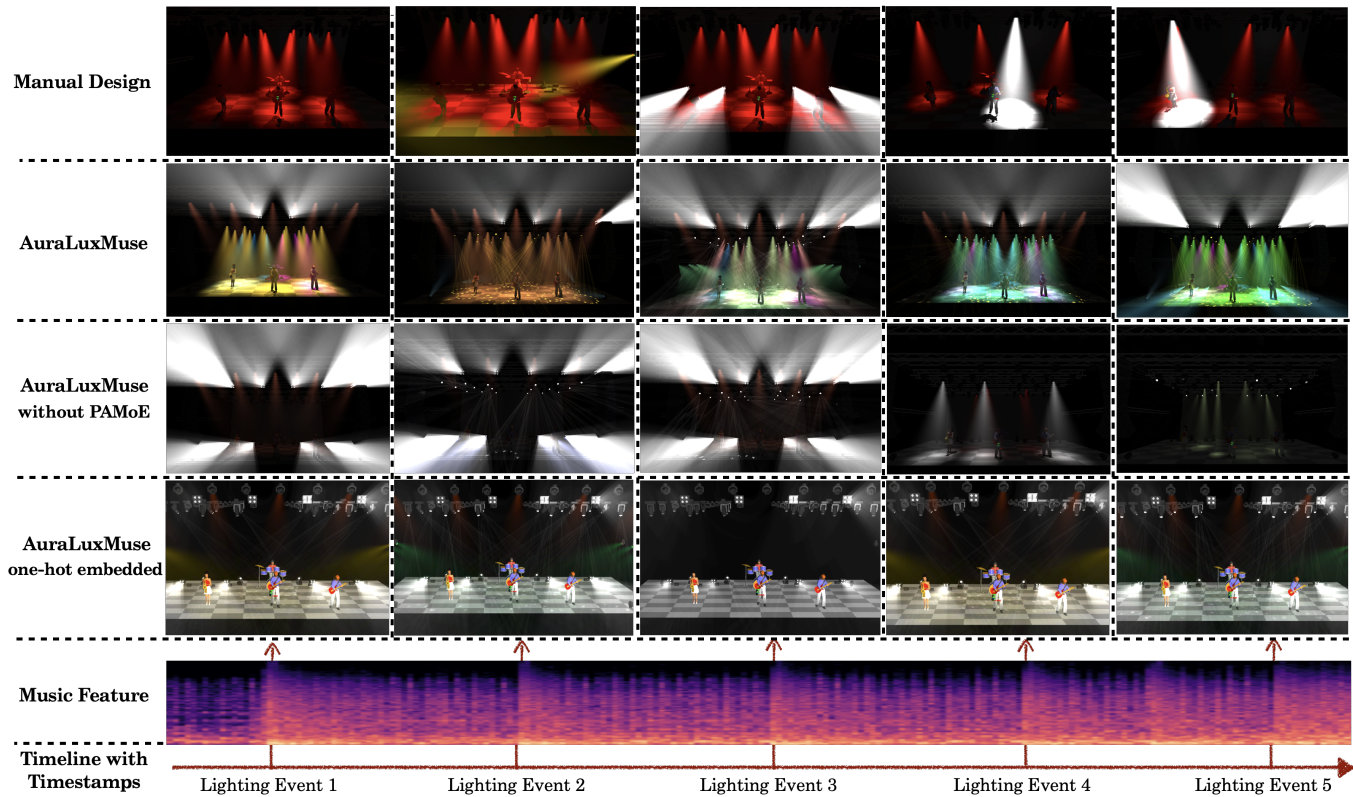


Fig. S2. **Qualitative Results in the Concert Scenario.** Consecutive frames compare manual lighting with AuraLuxMuse retrieved/adapted outputs under the same scenario; the variants show how the model structure changes cue selection.

Table S6. **Lighting Evaluation Dimensions (Relative to Music) for Non-Experts.** The criteria capture spectators' intuitive perceptions of performance quality across the three evaluated scenarios.

Ratings	Evaluation Dimensions		
	Rhythmic Alignment	Thematic Suitability	Visual Consistency
1	No alignment	Not suited	Extremely inconsistent
2	Poor alignment	Poorly suited	Somewhat inconsistent
3	Basic alignment	Basically suited	Basically consistent
4	Fair alignment	Fairly suited	Generally consistent
5	Perfect alignment	Perfectly suited	Always consistent

Table S7. **Overview of Lighting Style Dimensions.** Color, Brightness, Zone, and Rhythm are mapped to representative attributes.

Lighting Styles	Attributes
Color Harmony	<i>Joyful, Calm, Thrilled, Indifferent, Surprised, Confused, Depressed, Anxious, Angry, Fearful</i>
Brightness Control	<i>Dark, Dim, Moderate, Bright, Intense</i>
Zone Distribution	<i>Peripheral, Atmospheric, Neutral, Attention, Dominant Zone</i>
Rhythmic Tempo	<i>Static, Slow, Moderate, Fast, Rapid</i>

Table S8. **Brightness Attribute Levels.** Each level is defined by a perceptual description and illuminance range for preference specification.

Level	Description	Illuminance (lux)
Dark	Almost invisible	0–20
Dim	Barely visible	20–100
Moderate	Visually comfortable	100–300
Bright	Clear visibility	300–700
Intense	Extremely bright	700+

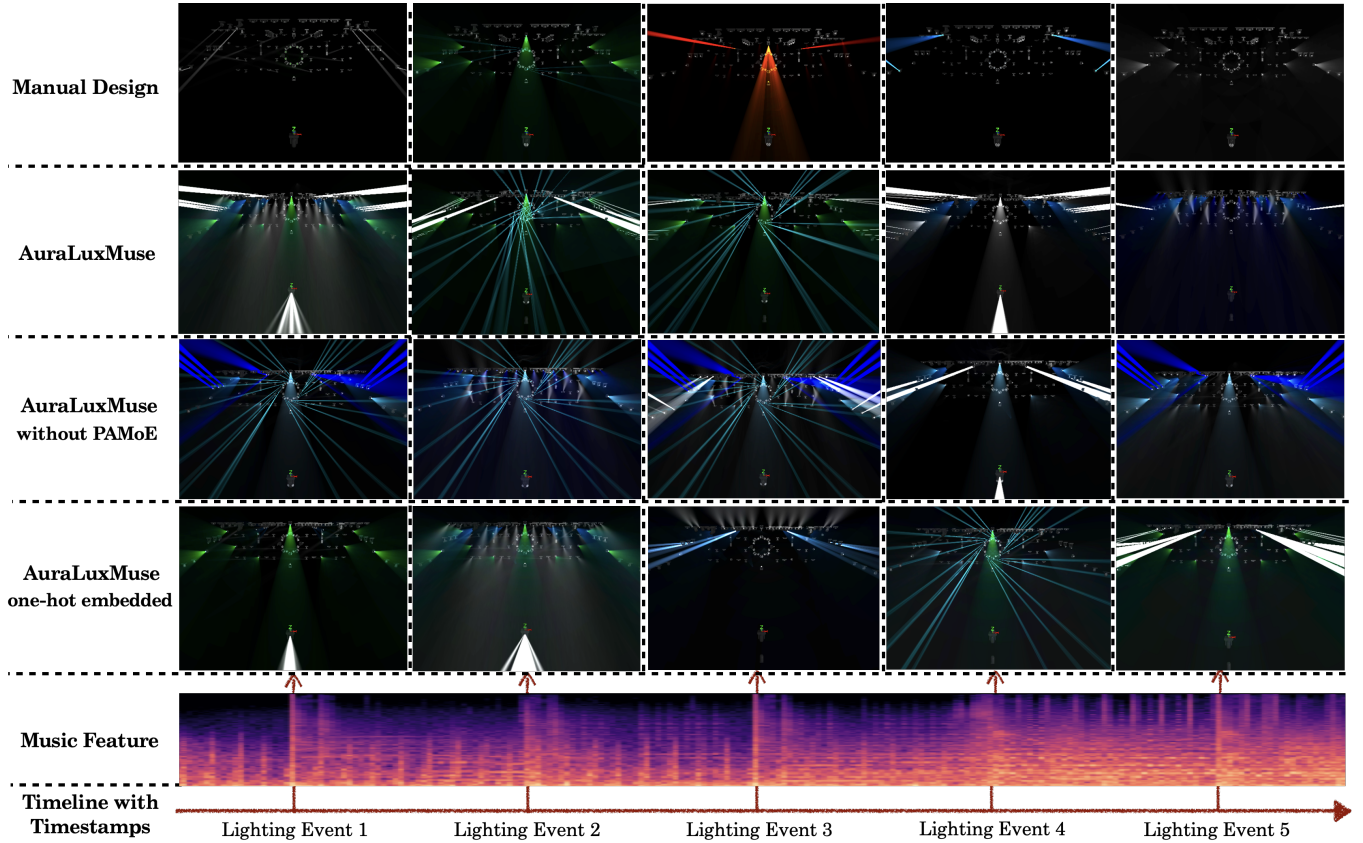


Fig. S3. **Qualitative Results in the Entertainment Scenario.** Consecutive frames compare a manual design with AuraLuxMuse retrieved/adapted cues under the same Entertainment scenario. The example illustrates different choices in visual complexity and atmosphere without implying universal superiority.

Table S9. **Zone Distribution Attribute Levels.** Each level defines a degree of visual prominence within a scene.

Level	Description
Peripheral	Invisible boundary or background layer
Atmospheric	Subtle lighting for mood creation
Neutral	Balanced fill without drawing attention
Attention	Guides audience attention
Dominant	Strongest visual focus

Zone distribution organizes visual-prominence regions using principles of art and perception. These regions shape audience focus and narrative structure; Table S9 defines the attributes.

Rhythmic tempo refers to the pace and regularity of lighting transitions, such as intensity changes, color shifts, and beam movements. It describes how quickly the lighting appears to change over time, often in response to musical rhythm or narrative pacing. The individual levels are defined in Table S10.

Table S10. **Rhythmic Tempo Attribute Levels.** Each level defines a transition pace for preference specification.

Level	Description
Static	Minimal change, nearly static
Slow	Gentle & Sparse transitions, rare rhythmic cues
Moderate	Regular tempo, structured but not fast
Fast	Frequent changes, strong rhythmic presence
Rapid	Intense flashes with strong beats

C.3 PAMoE Gate Sensitivity

Across 33 available preference labels, the mean weights of the four PAMoE experts are $[0.308, 0.239, 0.192, 0.261]$, with standard deviations $[0.089, 0.063, 0.042, 0.096]$. Experts 0, 3, and 1 are dominant for 16, 13, and 4 labels, respectively. Preference-expert correlations are structured: Color correlates most with Expert 3 (0.471), Brightness with Expert 1 (0.631), Zone with Expert 2 (0.513), and Rhythm with Expert 1 (0.375). In six existing preference-labeled outputs, Rhythm also correlates with motion ratio (0.700) and attribute entropy (0.691). The gate analysis directly demonstrates preference sensitivity; the output correlations are observational rather than a

controlled fixed-music causal sweep because the latter assets were unavailable for this auxiliary analysis. Experts 0–3 are identifiers for symmetric learned branches, not predefined Color, Brightness, Zone, and Rhythm experts. The reported correlations describe emergent associations and do not establish exclusive specialization. No auxiliary specialization loss or hard expert assignment is used.

D Patching Mechanism Analysis

Prior to formalizing the fixture patching procedure, it is imperative to delineate the architectural abstractions and hardware-software synchronization protocols governing stage configuration storage in industrial control frameworks (e.g., GrandMA2). Each fixture profile contains a **DMX Map**, a **Spatial Component** $\mathbf{p}_{norm} = (x, y, z)$, and a **Functional Component** $\mathbf{c}_{req} = [c_{pan}, c_{tilt}, c_{rgb}, c_{cm}, c_{gobo}, \dots]$ that encodes the required hardware features. Together, these components form a **Logical Layout Map** that stores stage-specific device details for retargeting on the destination stage.

Specifically, a DMX Map defines which channel controls each physical motor or LED within a fixture. The Spatial Component is a normalized coordinate vector that serves as the geometric anchor for an individual fixture in a three-dimensional Cartesian environment. This vector maps the fixture’s physical installation point relative to a predefined origin. The Functional Component serves as a mathematical abstraction that encodes the essential hardware specifications of a device into a requirement vector. Within this vector, each element represents a discrete physical attribute or mechanical capability. For instance, the variable c_{pan} signifies the fixture’s ability to process horizontal rotation commands; in a practical environment, this dictates whether the hardware possesses the motors and circuitry required to perform panning movements. This logic applies consistently across the entire vector, where each subsequent element corresponds to a specific hardware-enabled function.

Our flexible patching mechanism accepts an AuraLuxMuse retrieved/adapted cue $C_{out} = \langle \mathbf{p}_{norm}, \mathbf{c}_{req}, \text{Attribute}, \text{val} \rangle$ and a digital twin parsed from the MVR (My Virtual Rig) standard. The digital-physical address mapping defines a **Patch Table** for DMX fixture control. Musilux covers the fixture categories in Fig. S6; compatibility with a target stage is explicitly checked through $\mathbf{c}_{req}$ rather than assumed.² For a different stage setup, retargeting is formulated as a Linear Sum Assignment problem. A conflict occurs when multiple retrieved commands assign different values to the same fixture–attribute pair in one execution window. Real-world fixture coordinates are projected into normalized space $\hat{\mathbf{x}}$, and C_{ij} combines spatial Euclidean distance with hardware/attribute mismatch penalties. The Kuhn–Munkres solution π^* assigns commands to spatially and functionally compatible slots before console dispatch. In a deliberately conflict-heavy case with 148 fixture–attribute commands, 94 collisions occur before assignment and 0 remain afterward; attribute compatibility is preserved. The low same-fixture preservation rate in this stress test reflects the intentionally dense collision pattern. Normal exports rarely require this much reassignment because retrieved sequences usually control complementary spatial components.

²MVR is an open-source XML-based file format developed to exchange stage descriptions between CAD software and lighting consoles.

Table S11. **Ablation of LAMP Architectures.** Contrastive loss is reported after 500 training epochs of model optimization.

Music Encoder / Cue Encoder	Transformer	ResNet
VGGish + MLP	0.699(↓)	2.015
VGGish + ResNet	2.181	3.926

Table S12. **Ablation of PAMoE and Soft Alignment.** All experiments use the same learning rate, weight decay, and number of epochs.

Model Variant	FID(↓)	S _{exp}	S _{non}	EGSS(↑)
PAMoE Only	18.06	2.26	2.86	2.44
Soft Only	10.97	2.83	3.47	3.02
PAMoE + Soft	4.89	3.19	3.45	3.27

E Ablation Studies

We perform four groups of ablation studies to assess the contributions, structures, and design choices of the AuraLuxMuse components.

Training Cost and Failure Signatures. On one RTX 4090, LAMP requires approximately 50–60 minutes for 500 epochs. Joint LAMP and PAMoE training takes about 1 hour, the TCN Metadata Predictor takes 10–20 minutes, and the full pipeline takes approximately 1.0–1.5 hours. Unsuitable contrastive settings slow or destabilize alignment-loss reduction, while unsuitable density-prediction settings destabilize the cue-count loss; these effects later weaken temporal alignment or visual integrity.

E.1 Structural Variants of Modules

1. *Ablation Study on LAMP Backbone Structures.* We begin by investigating architectural variants of LAMP to identify the most effective configuration. As shown in Table S11, the combination of a VGGish-based music encoder with MLP layers and a transformer-based lighting encoder yields the most favorable performance.

2. *Ablation Study on PAMoE Influence.* Subsequently, we assess PAMoE’s impact and the combination of soft and hard alignment through ablations with and without these components. As shown in Table S12, incorporating PAMoE improves both objective metrics and subjective perception, highlighting the value of integrating expert-level artistic knowledge into the learning process. Moreover, soft alignment improves performance by encouraging shared latent lighting behavior across stages within the same scenario.

3. *Metadata Predictor Effects.* To determine dynamically the number of retrieved lighting sequences (the K value) based on musical rhythm, we formulate a cross-modal dense temporal-prediction task. Frame-level audio features (extracted using VGGish, $D = 128$) are segmented into overlapping temporal windows with a total length of $W = 64$ frames and stride $S = 4$. The supervision target is the discrete count of lighting events within each window. As shown in Table S13, the frame-wise MLP baseline fails under unified

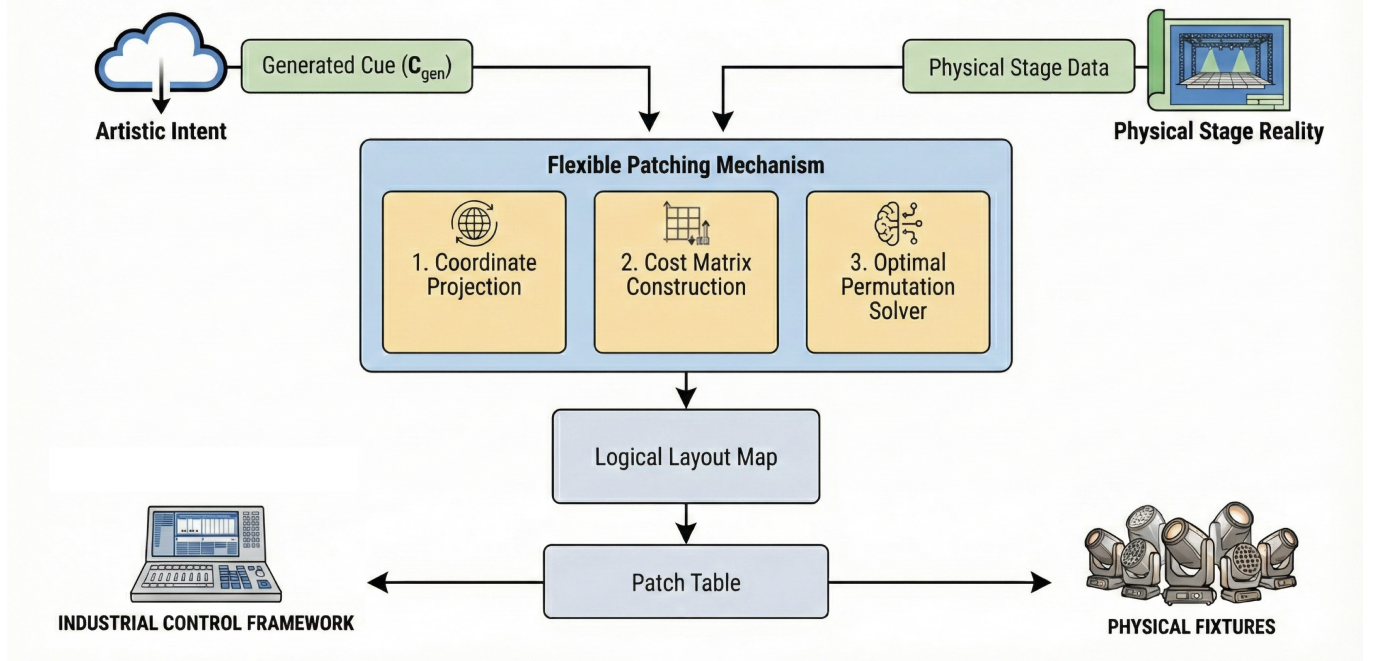


Fig. S4. **Patching Mechanism Overview.** Retrieved artistic cues are retargeted to compatible fixtures through a logical layout map and an industrial-console patch table used for dispatch to the console.

Table S13. **Metadata Predictor Performance.** We compare window-level cue counting with a baseline MLP and our non-causal TCN.

Architecture	MAE ↓	RMSE ↓	R^2 ↑	Pearson r ↑	Bias ↓
MLP	16.987	24.113	-368.715	0.061	+16.973
TCN Density	0.343	0.449	0.872	0.934	+0.017

window-level evaluation because of severe overprediction. This result highlights that mapping isolated audio frames to sparse lighting events without a sufficient contextual receptive field accumulates errors. In contrast, our non-causal TCN-based Metadata Predictor uses symmetric padding to model preceding and subsequent context within each complete input window. It aligns more closely with the sparsity of professional stage lighting and provides the density estimate λ_s that guides the downstream retrieval module.

E.2 Baseline Architecture Comparison

1. Contrastive Learning Baselines. To evaluate the effectiveness of our cross-modal alignment, we benchmark AuraLuxMuse against three representative representation learning architectures adapted for music-to-lighting retrieval: (1) CLAP [Wu et al. 2024], a dual-encoder architecture trained with symmetric InfoNCE contrastive loss to align audio and text representations in a shared embedding space; (2) SLAP [Guinot et al. 2025], which adopts a Siamese architecture that learns audio-language correspondence without explicit negative samples via self-supervised representation matching; (3) AudioCLIP [Guzhov et al. 2021], which extends CLIP-style contrastive learning to three modalities (audio, image, and text) and

Table S14. **Comparison with Contrastive Baselines.** AuraLuxMuse obtains the highest Precision@5 and EGSS under the reported setup.

Architecture	FID ↓	EGSS ↑	Precision@5 ↑	mAP ↑
CLAP	5.06	2.52	0.11	0.043
SLAP	7.20	2.11	0.02	0.085
AudioCLIP	5.21	2.18	0.13	0.05
AuraLuxMuse	4.89	3.27	0.170	0.100

learns a unified embedding space through multi-way cross-modal contrastive objectives. As shown in Table S14, the adapted generic contrastive models struggle with the discrete, physically constrained stage-lighting space. AudioCLIP is the strongest of these baselines, while AuraLuxMuse obtains the best semantic-retrieval and expert-guided subjective scores under this evaluation.

2. End-to-End Generative Baselines. Because lighting-cue generation has not previously been formulated as a sequence-to-sequence task, we introduce a custom CueTokenizer to bridge continuous VGGish audio features and structured lighting controls. It converts lighting events into token sequences by quantizing continuous parameters into 256 bins and mapping categorical attributes to an NLP-like vocabulary. Using this formulation, we train four autoregressive baselines with token-level cross-entropy and teacher forcing: lightweight adaptations of BART [Lewis et al. 2019] and T5 [Raffel et al. 2020], which inject linearly projected audio frames as encoder embeddings; a domain-specific SpecTNT [Lu et al. 2021] surrogate with

Table S15. **Comparison with Generative Baselines.** Against four sequence-to-sequence models, AuraLuxMuse achieves lower inference latency and a higher EGSS under the reported setup.

Architecture	FID	EGSS $\uparrow$	Inference Time (s/file) $\downarrow$
T5	4.28	1.79	25.74
BART	4.21	1.89	35.20
Autoformer	3.98	1.24	207.27
SpecTNT	3.95	1.24	150.18
AuraLuxMuse	4.89	3.27	7.65

a lightweight causal Transformer decoder; and an Autoformer [Wu et al. 2022] surrogate with a higher-capacity sequence-modeling variant. As shown in Table S15, the tested autoregressive baselines have much higher inference latency and can emit invalid attribute combinations. AuraLuxMuse instead retrieves expert-authored sequences and patches them to compatible fixtures. This design preserves corpus-level cue validity and yields 7.65-second average inference under the reported setup; final physical validity still depends on the compatibility checks and conflict-resolution procedure. The generative baselines perform poorly because attributes and values are highly correlated, while discrete tokenization produces a combinatorial number of possible token sequences. Without explicit retrieval guidance, the autoregressive models struggle to learn the dependencies and constraints of stage-lighting design, resulting in incoherent outputs and lower subjective scores.

E.3 Retrieval Baseline Diagnostics

The rule-based baseline partitions the cue corpus with coarse audio-duration and complexity heuristics and then samples compatible professional cues. It is physically executable and interpretable but does not learn music–lighting correspondence. Random retrieval shares the Metadata Predictor’s event timing and samples from the same expert-authored, console-valid corpus; this explains why it is a strong corpus-prior lower bound rather than a zero-quality baseline. Its weakness is semantic alignment: random retrieval scores 2.70 in narrativity versus 3.64 for similarity retrieval, while visual consistency is closer (3.30 versus 3.55). Auxiliary cue statistics support this distinction: random outputs have a substantially larger mean cue-statistic distance from similarity retrieval than the rule-based outputs reported and analyzed above.

E.4 Delicate Component Mechanisms

1. Ablation Study on K -Value Design. The symbol K denotes the number of retrieved sequences computed by integrating the predicted event density. All top- K sequences are considered because an individual sequence may control only one stage region; conflicts are resolved before dispatch. Table S16 compares adaptive K with fixed values, and the adaptive setting obtains the best reported objective and subjective scores among those tested configurations.

2. Ablation Study on Soft Alignment and Weights. To investigate the contribution of the soft alignment mechanism in our contrastive learning objective, we perform an ablation study on the soft positive

Table S16. **Ablation Study: Static vs. Dynamic K Values.** Dynamic K yields a lower FID and a higher EGSS than either static setting.

Strategy	K	FID ($\downarrow$)	EGSS ($\uparrow$)
Fixed Small	1	8.42	2.85
Fixed Large	10	9.82	3.02
Adaptive	Dynamic	4.89	3.27

weight, denoted as w . While fixing the standard hard alignment weight at 1.0, we systematically sweep $w \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$. The introduction of soft alignment is highly beneficial for the stage lighting domain; unlike standard text-image pairs, multiple visually distinct lighting cues can be aesthetically compatible with the same musical rhythm. By assigning a partial target probability (w) to these style-consistent samples, the soft alignment loss provides a smoothed learning signal that prevents the model from overly penalizing valid stylistic variations, thereby enriching the semantic continuity of the shared embedding space. As shown in Table S17, setting $w = 0.3$ provides the best configuration. When w is too small (e.g., 0.1), strict hard alignment dominates and the model does not fully exploit latent stylistic correlations. Conversely, increasing the soft weight to $w \geq 0.5$ causes soft positives to overpower the exact ground-truth anchors. This excessive smoothing blurs discriminative boundaries in the feature space, as evidenced by a substantial drop in retrieval recall and an increase in FID to 61.21 at $w = 0.7$. At $w = 0.3$, the network achieves the highest overall retrieval accuracy while preserving a stable representation manifold.

3. Ablation Study on Lighting Embedding Strategies. We evaluate two cue embedding strategies: attribute mapping, which assigns predefined indices to categorical lighting attributes, and one-hot encoding, which uses high-dimensional binary vectors. As shown in Table S18, attribute mapping consistently outperforms one-hot encoding across objective metrics and subjective evaluations.

E.5 Metric Validity and Reliability

1. Sensitivity Analysis. To validate our evaluation protocol, we investigate the sensitivity of the Expert-Guided Subjective Score (EGSS) to the expert-weight parameter λ . As illustrated in Fig. S5, varying λ systematically adjusts the emphasis placed on professional lighting principles relative to general visual perception. The manual design serves as an approximate upper bound rather than a competing baseline. Across the entire λ range, AuraLuxMuse outperforms the tested automated baselines. We adopt $\lambda = 0.7$ as the default configuration because it balances expert alignment with general aesthetic perception under the reported study design.

2. Correlation Analysis between FID and EGSS. We examine FID and EGSS across ten automated model variants. Pearson correlation is positive but not significant ($r = 0.3955$, $p = 0.2579$), whereas Spearman correlation is positive and significant in this sample ($\rho = 0.6848$, $p = 0.0289$). Because this analysis has only ten variants, excludes manual design, and evaluates FID in one learned feature space, it indicates exploratory rank association rather than

Table S17. **Ablation on Soft Alignment Weight (w).** We evaluate the impact of the soft-alignment weight w on cross-modal retrieval performance and Fréchet Inception Distance (FID). The hard-alignment weight is fixed at 1.0. Results indicate that $w = 0.3$ yields the best balance, achieving the strongest overall retrieval metrics by leveraging stylistic similarities without blurring discriminative boundaries.

Soft Weight (w)	Avg FID ↓	R@1 ↑	R@5 ↑	R@10 ↑	MRR ↑	nDCG ↑
$w = 0.1$	11.85	0.045	0.223	0.338	0.144	0.303
$w = 0.3$	4.89	0.046	0.249	0.374	0.149	0.309
$w = 0.5$	10.10	0.062	0.202	0.280	0.143	0.300
$w = 0.7$	61.21	0.033	0.159	0.301	0.118	0.279
$w = 0.9$	47.09	0.020	0.127	0.258	0.097	0.261

Table S18. **Ablation of Cue-Embedding Strategies.** The architecture is fixed, and only the embedding strategy varies.

Embedding	FID(↓)	S_{exp}	S_{non}	EGSS(↑)
One Hot	12.24	2.51	3.29	2.74
Attribute Map	4.89	3.19	3.45	3.27

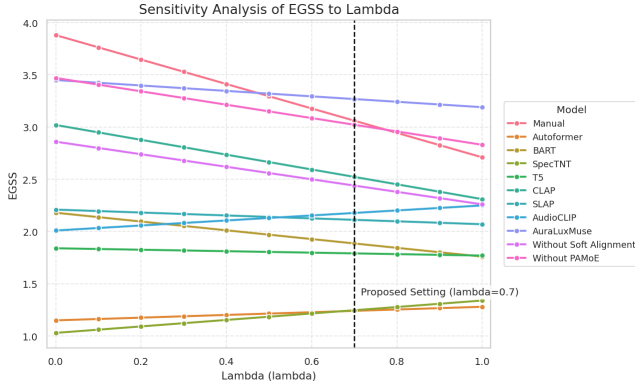


Fig. S5. **Sensitivity analysis of the EGSS metric with respect to λ .** The parameter λ controls the weight assigned to expert preferences in the subjective evaluation process. The manual lighting design serves as an approximate upper bound. AuraLuxMuse consistently outperforms all automated baselines across the full λ range, demonstrating strong robustness to this hyperparameter. The vertical dashed line indicates our selected default configuration ($\lambda = 0.7$).

robust metric validity. FID and EGSS are therefore complementary diagnostics, not interchangeable aesthetic measures.

F Stage Lighting Environments

F.1 Software

For high-dimensional parameter modulation, sequence orchestration, and real-time output synchronization, we employ GrandMA2 as our primary lighting-control interface. For synthetic data generation and previsualization, MA3D renders complex volumetric lighting effects in a simulated environment. The system also supports hardware-in-the-loop execution: control signals generated through GrandMA2 are mapped identically to both the virtual engine and physical stage luminaires. This unified protocol ensures

spatiotemporal consistency between simulated and real-world lighting distributions, providing a framework for cross-domain validation.

GrandMA2 offers a robust, scalable solution for professional stage lighting control, integrating high-performance hardware with versatile software to support a wide range of production needs. The system supports seamless show file management, real-time cue execution, and parameter expansion across a unified control interface. The GrandMA2 software enables rich programming functionality including presets, effects, and time-synchronized lighting behaviors, with consistent UI across hardware configurations. Designed for demanding live environments, GrandMA2 provides reliable performance for productions ranging from small installations to arena-scale events.

MA3D is a high-fidelity real-time rendering and visualization engine designed for complex lighting configurations. It serves as a digital twin of the physical stage, enabling simulation of volumetric light beams, non-Lambertian shadows, and dynamic spectral distributions across three-dimensional occluders. The engine maintains bidirectional data synchronization with GrandMA2 through a high-speed network interface, ensuring that virtual lighting states are spatiotemporally aligned with hardware-level DMX output.

F.2 Lighting Fixtures

Stage-lighting fixtures can be classified by the morphology of their emitted light into four classes: imaging lights, spotlights, soft lights, and floodlights. Fig. S6 shows the fixture models represented in Musilux. Each fixture type serves a distinct purpose in shaping the visual atmosphere of a performance and offers different degrees of beam control, diffusion, and intensity.

Imaging lights, often equipped with advanced optical systems, allow precise projection of patterns, gobos, or textured effects. Spotlights use focused, directional beams to highlight performers or key stage elements with sharp, well-defined edges. In contrast, soft lights produce diffuse, even illumination that minimizes harsh shadows, making them suitable for naturalistic or ambient conditions. Floodlights deliver broad, uniform coverage and are frequently used to wash large stage areas or backdrops. These four fixture classes cover the professional-stage configurations represented in Musilux.

Fig. S7 depicts fixture parameters stored in the stage-lighting control interface. The interface maps hardware IDs to lighting and spatial-transform vectors, recording extrinsic parameters alongside the multimodal data. This deterministic mapping from logical

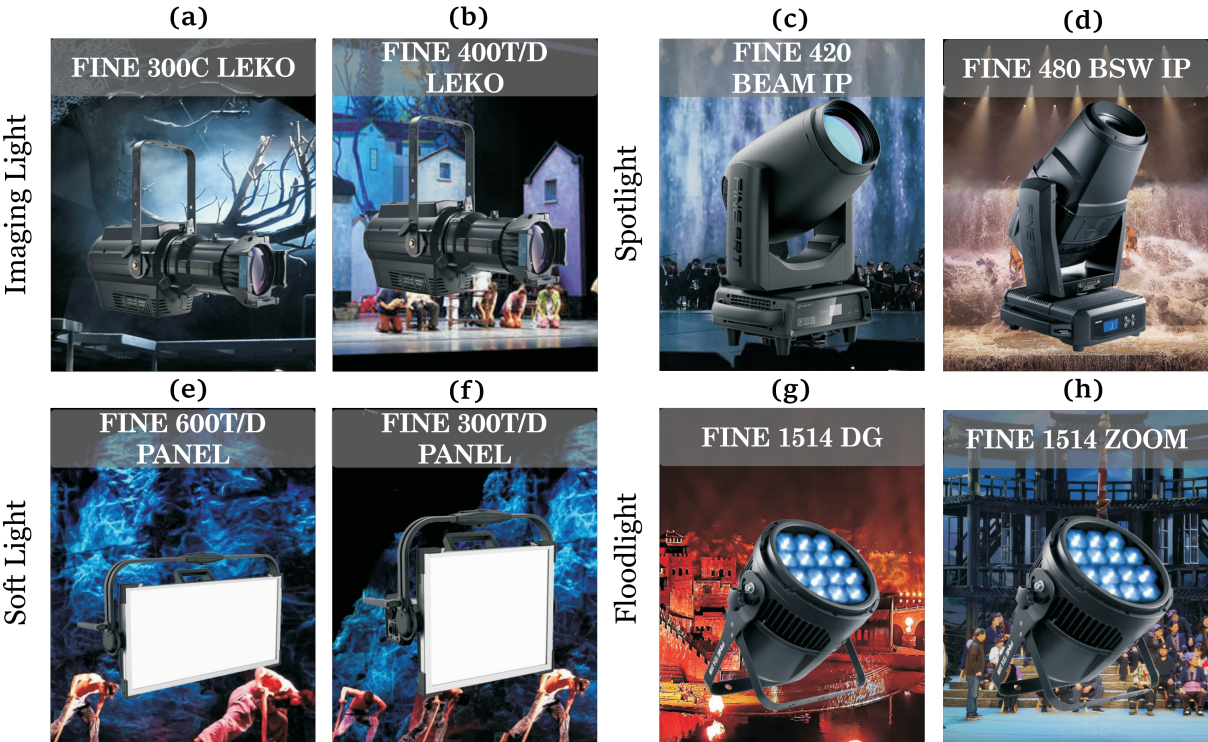


Fig. S6. **Examples of Stage Lighting Fixtures:** (a-b) Imaging Lights: FINE 300C LEKO and FINE 400T/D LEKO for sharp projection and patterned illumination; (c-d) Spotlights: FINE 420 BEAM IP and FINE 480 BSW IP for dynamic focus and concentrated beams; (e-f) Soft Lights: FINE 600T/D PANEL and FINE 300T/D PANEL for diffused illumination and even coverage; (g-h) Floodlights: FINE 1514 DG and FINE 1514 ZOOM for broad lighting and wide effects.

Fixture Sheet

ID	F/C	Name	Dimmer	Dim W	DirCurve	DirPanel	Position	Tilt	Pan	Tilt F	Gobo1	Gobo2
102:		Alpha Spo	close	0.0			center				open	open zero
103:		Alpha Spo	close	0.0			center				open	open zero
104:		Alpha Spo	close	0.0			center				open	open zero
105:		Alpha Spo	close	0.0			center				open	open zero
106:		Alpha Spo	close	0.0			center				open	open zero
107:		Alpha Spo	close	0.0			center				open	open zero
108:		Alpha Spo	close	0.0			center				open	open zero
109:		Alpha Spo	close	0.0			center				open	open zero
110:		Alpha Spo	close	0.0			center				open	open zero
111:		Alpha Spo	close	0.0			center				open	open zero
112:		Alpha Spo	close	0.0			center				open	open zero
113:		Alpha Spo	close	0.0			center				open	open zero
114:		Alpha Spo	close	0.0			center				open	open zero
115:		Alpha Spo	close	0.0			center				open	open zero

Fixture ID **Attributes**

Fig. S7. **Fixture Parameter Visualization.** The Fixture Sheet records instantaneous lighting-infrastructure states. Its fields match the functional components in Section D, enabling precise state tracking during acquisition.

control to physical fixture behavior supports tasks such as inverse rendering and scene decomposition.

F.3 Stage Configurations

In addition to software and fixtures, the stage provides the physical environment in which lighting operates. A stage configuration is stored from multiple perspectives, including two-dimensional geometric information established through three orthographic views in the industrial design process. To establish a robust mapping between the virtual simulation and the physical capture space, the spatial configuration within the GrandMA2 ecosystem contains complex geometric occluders. As depicted in Fig. S8 (Top View) and Fig. S9 (Front & Side View), this integrated representation facilitates the precise calibration of light transport paths and ensures that the extrinsic properties of the illumination sources are preserved across both synthetic and real-world domains.

The stage environment serves as a persistent spatial repository in which the extrinsic parameters of heterogeneous fixtures are represented in a unified three-dimensional coordinate system. These geometric priors allow the architectural configuration to be serialized in GrandMA2 show files and persistent data archives. This hierarchical representation supports rapid initialization of complex lighting configurations and import of existing stage-lighting programs. The serialized format also preserves the spatial distribution of the lighting infrastructure across project instances, supporting consistent multimodal data acquisition.

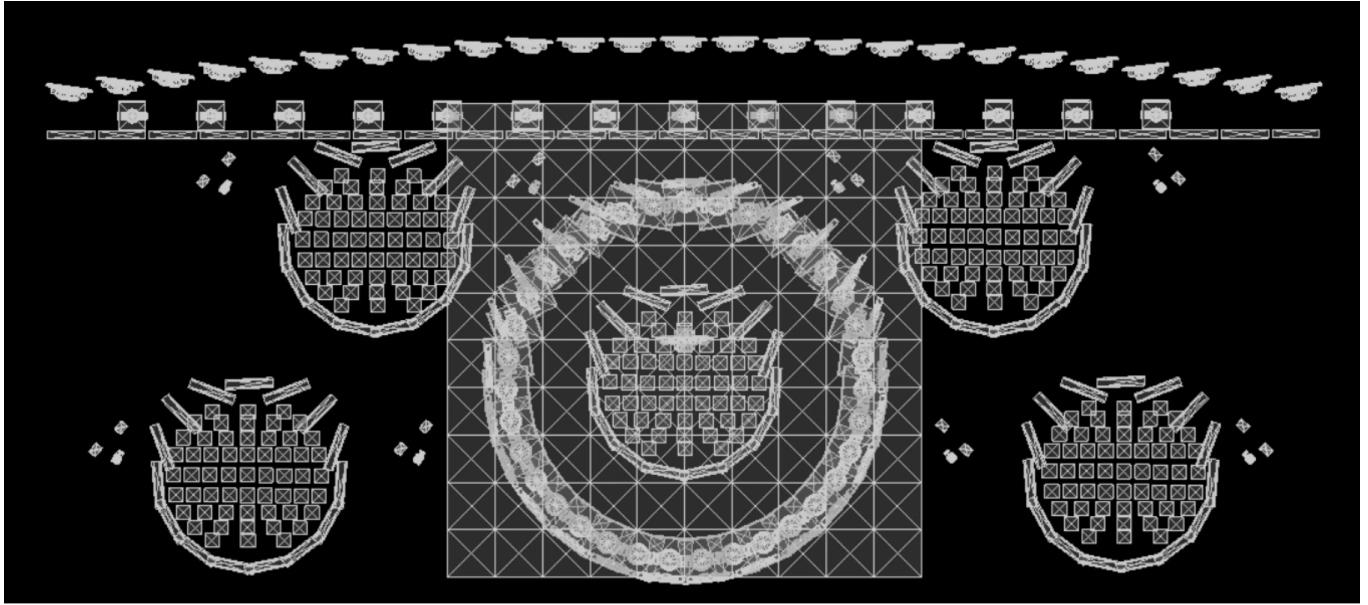


Fig. S8. **Software-Defined Spatial Manifold: Top-down View.** The top-down orthographic projection of the experimental environment as registered within the GrandMA2 configuration space. Each luminaire instance and structural element is precisely localized within a unified global coordinate system, providing the necessary extrinsic parameters for spatiotemporal alignment. This geometric blueprint serves as the architectural ground truth, ensuring that synthesized radiance fields are spatially consistent with the physical acquisition environment.

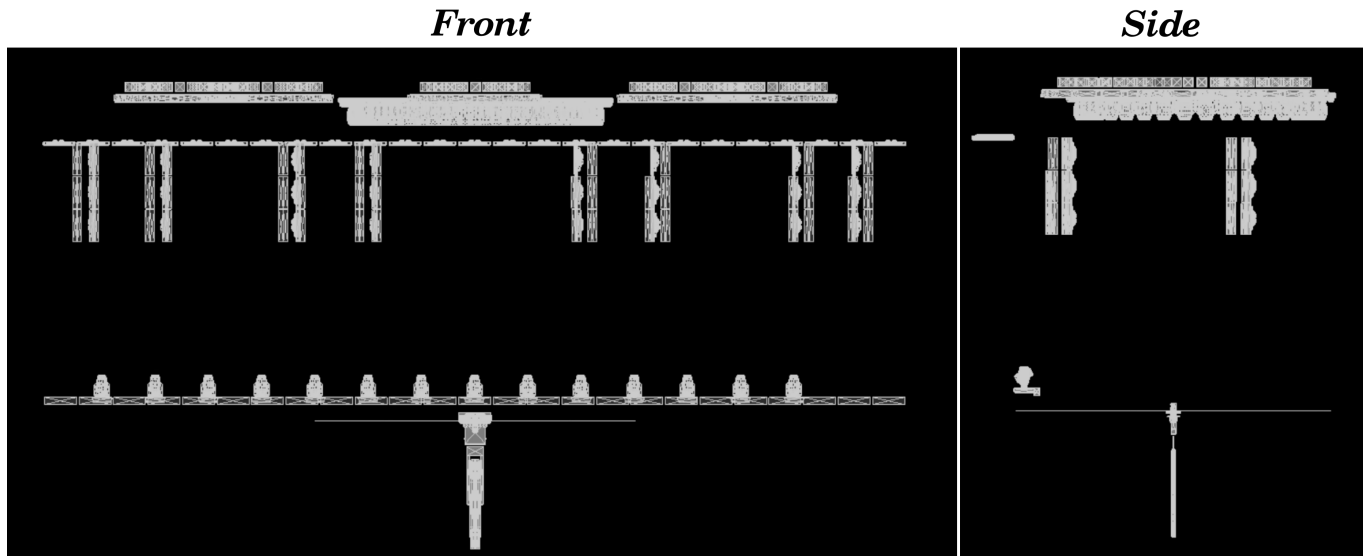


Fig. S9. **Software-Defined Spatial Manifold: Front and Side Views.** Registered orthographic projections map fixture elevations and lateral offsets, providing ground truth for spatial registration.

F.4 Stage Lighting Console

Fig. S10 illustrates the physical control surface used to orchestrate the lighting infrastructure. GrandMA2 provides a structured framework for temporal sequencing, per-fixture mapping, and low-level

parameter modulation. Coupling the hardware surface with the software environment enables control of complex lighting trajectories. This configuration supports hardware-in-the-loop actuation, allowing designers to manipulate real-world stage lighting and refine its state in real time through a tactile control surface.



Fig. S10. **Example of a GrandMA2 Console.** The professional console provides robust stage-lighting control.

The full-size console supports 8,192 parameters natively and scales up to 65,536 via MA Network Processing Units (NPUs), while the light and ultra-light versions offer 4,096 parameters, expandable similarly. Each console features multi-touch displays, extensive DMX I/O (up to 6 outputs), MIDI/SMPTE integration, and ethernet-based protocols such as Art-Net and sACN.

F.5 Stage Lighting Interface

The interactive software interface is demonstrated in Fig. S11. It is a hierarchical state-space for managing dynamic radiance environments and provides the following functional components:

- **Temporal Synchronization (Timecode):** The global time-code interface in Fig. S12 maps discrete event triggers to a high-resolution SMPTE reference.
- **Hierarchical Cue Sequence Representation:** The repository in Fig. S13 stores illumination events as structured Cue

Sequences; each cue is a state vector that parameterizes lighting behavior.

- **Attribute Control Surface:** Eight encoders (Dimmer, Position, Gobo, Color, Beam, Focus, Control, and Shapers) control fixture parameters during show execution.
- **Tactile Command Emulation:** A 1:1 digital twin includes the Numeric Keypad, Command Buttons (e.g., Store, Update, Clear), and Execution Faders.
- **Navigation and Workspace Management:** Vertical tabs support rapid switching among fixture sheets and layouts.

G Stage Lighting Implementation Procedure

This section details how to export learnable data structures from existing lighting shows and import AuraLuxMuse-generated stage lighting into real-world simulations and applications.



Fig. S11. **GrandMA2 Software Interface.** The software provides real-time visualization, cue management, and timeline editing.

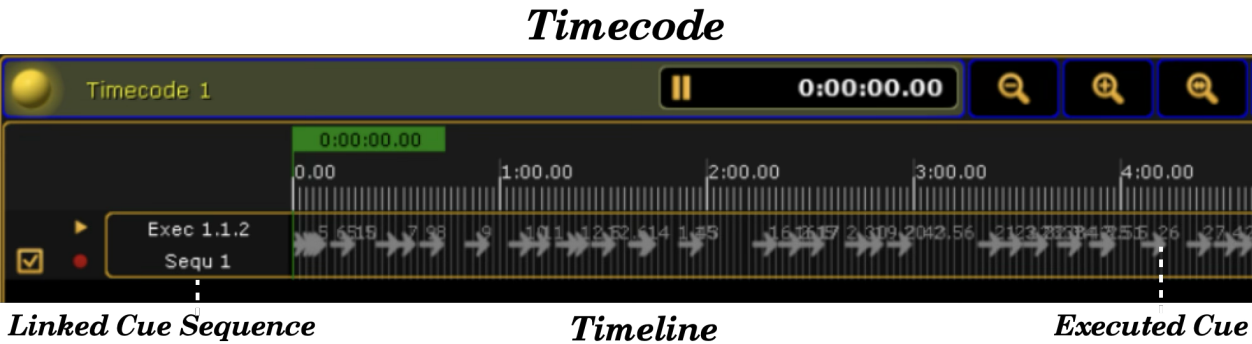


Fig. S12. **Timecode Encoding in GrandMA2.** The interface precisely synchronizes multimodal lighting triggers; each marker denotes a discrete cue execution.

G.1 Stage Lighting Input

This section describes how to export XML files from existing stage-lighting designs in GrandMA2 and convert them into readable documents for constructing Musilux.

Step 1: Import Lighting Show. To begin, the show file corresponding to the specific stage lighting setup must be imported into the console. This is done by selecting the Load Show option, which

allows the user to load the desired lighting configuration file into the GrandMA2 system.

Step 2: Export XML Files. For one sequence or timecode, execute:

- SelectDrive 4
- Export Sequence (Number of Sequence) ("File Name")
- Export Timecode (Number of Timecode) ("File Name")

For multiple sequences or timecodes, execute:

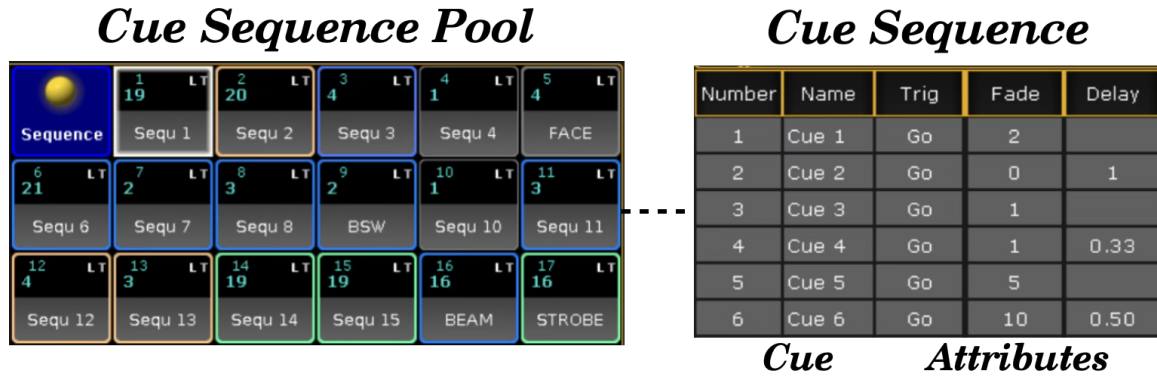


Fig. S13. **Structured Parameter Space for Cue Sequences.** (Left) The Sequence Pool stores illumination trajectories. (Right) An individual Cue Sequence specifies fixture behavior through an attribute vector that includes transition duration (Fade) and temporal offset (Delay), enabling deterministic execution.

- SelectDrive 4
- Export Sequence (Start) Thru (End) ("File Name")
- Export Timecode (Start) Thru (End) ("File Name")

Step 3: Data Conversion. The model cannot directly use the original XML stage-lighting data. Run `read_sequence_cue.py` and `read_timecode.py` to extract structured information and produce JSON files used for model training.

G.2 Stage Lighting Output

This section describes how to integrate AuraLuxMuse-generated XML files into the **GrandMA2** console for stage-lighting control during deployment on target stages.

Step 1: File Preparation. AuraLuxMuse generates a cue sequence XML file and a timecode XML file, which should be placed in the USB drive's

`\gma2\importexport` directory.

Step 2: Console Setup. Open a blank show file and configure the lighting fixtures.

Step 3: Import Files. Execute these macros:

- SelectDrive 4
- Import ("File Name") At Sequence 1
- Import ("File Name") At Timecode 1

Replace "File Name" with the actual XML file names. For simplicity in subsequent explanations, the numbers are assumed to be 1, though they can be adjusted as needed.

Step 4: Assign Sequence. Run Assign Sequence 1 At Executor 1 to link the cue sequence to Executor 1, which is also adjustable.

Step 5: Bind Timecode. Edit Timecode 1 and select the corresponding Executor 1.

Step 6: Playback. Click the timecode play icon to start synchronized lighting control.

This workflow enables efficient AuraLuxMuse deployment.

H Dataset Split and Generalization Analysis

Training excludes the validation and test assets and contains approximately 150 minutes of paired music and lighting cues. Validation

covers six styles with two held-out examples per scenario and supplies manual references for the main-paper comparison. Test uses six unseen tracks across the six styles on a stage excluded from training. This is a controlled, style-balanced check of unseen music and unseen stage layout, not a population-scale benchmark; full-length expert/non-expert scoring limits its practical sample size. A cue-statistic t-SNE analysis over 14 stage-level output vectors yields a mean pairwise z-scored distance of 5.409 ± 1.746 , indicating multiple statistic regimes rather than collapse to a single template. The real-world demo stage is unseen; its nearest cue-statistic neighbors are the dance and rock stage cases. Because the t-SNE uses available cue-output vectors rather than the complete raw Musilux corpus, we treat it as supplementary descriptive evidence, not as a standalone generalization metric. The approximately 150-minute training portion is about 80.1% of total Musilux duration. Validation comprises six held-out programs, two per scenario, selected for six-style coverage and the availability of human-authored references. Test comprises six additional tracks, one per style, and an unseen stage layout. These count- and duration-level descriptions are both reported because program lengths vary.