DOI: 10.21917/ijivp.2026.0564

SC-M2U: MAPPING SPATIAL COMPOSITION IN IMAGES TO MULTI-TRACK MUSIC THROUGH CROSS-MODAL GATED ALIGNMENT

Gurpreet Singh
Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Malaysia

Abstract

Turning a photograph into a piece of music is an intriguing but difficult cross-modal task. Most existing image-to-music systems squeeze the entire image into one global feature vector and then hand that single vector to a music generator. The trouble is that this flattening step throws away the spatial layout of the scene. A guitarist in a quiet meadow gets the same treatment as an empty meadow, because the global vector averages everything together. We present SC-M2U, a framework that splits the image into foreground and background regions, encodes each region separately through a shared CLIP ViT-B/32 backbone, and routes the resulting embeddings to distinct music stems through a gated cross-modal attention module. The foreground drives the lead melody track while the background shapes the accompaniment. We evaluate SC-M2U on five diverse visual scenes using cosine similarity alignment, ablation over five crop ratios, and a cross-scene confusion analysis. On average, the matched region–track alignment reaches 0.2411, which sits above the 0.2387 of the global baseline, with a positive selectivity contrast Δ of +0.001 and a diagonal dominance of +0.0213 in the confusion matrix. These numbers are modest under frozen CLIP weights, but they show that spatially aware conditioning tightens the link between what is seen and what is heard, and they give a concrete starting point for controllable, multi-stem image-to-music generation.

Keywords:

Image-to-Music Generation, Cross-Modal Alignment, Spatial Decomposition, Gated Attention, Multi-Track Music Generation

1. INTRODUCTION

Over the last few years, cross-modal generation has grown from a curiosity into a practical research area. We can now go from text to images, from images to captions, from speech to text, and in many other directions. One particular direction that has received less attention, but is no less fascinating, is image-to-music generation: given a single photograph or artwork, produce a piece of music that somehow "fits" what the image shows. This task sits between computer vision and music information retrieval, and it has clear applications in creative tools, accessibility, and multimedia production.

The difficulty is that images and music live in fundamentally different representation spaces. An image is a two-dimensional grid of pixels with spatial structure, while music is a one-dimensional signal that unfolds over time with harmonic and rhythmic structure. Bridging these two modalities requires finding some shared representation where visual content and musical content can be meaningfully compared. Recent work has explored several strategies converting images to text captions and feeding those to a text-to-music model [1], [2], mapping images to emotion coordinates like valence and arousal [3], [4], or directly projecting visual embeddings into audio latent spaces [5], [6], [7]. Each of these strategies has merit, but they all share a common blind spot.

That blind spot is spatial composition. Think about a photograph of a jazz band performing on a street corner at dusk. The foreground shows musicians with their instruments: a trumpet, a saxophone, a double bass. The background shows city buildings, streetlights, and a darkening sky. A human listener might expect the music to feature a punchy brass lead in the melody, backed by a walking bass line and a warm, ambient pad that evokes the evening atmosphere. But every existing model we are aware of would collapse this spatial structure into one average vector. The trumpet and the sky get blended together. The foreground objects, which should drive the melody, carry no more weight than the ambient background. No prior work, to the best of our knowledge, has explicitly linked the spatial decomposition of an image to the multi-track structure of music.

We address this gap with SC-M2U, a framework for Spatial-Compositional Cross-Modal Music Generation. Our contributions are as follows:

- We propose a spatial decomposition pipeline that separates an image into foreground and background regions and encodes each region independently through a shared CLIP visual backbone.
- We design a Cross-Modal Gated Attention (CMGA) module that dynamically routes region-level visual embeddings to their corresponding music stems (foreground to lead melody, background to accompaniment) through a learned gating mechanism.
- We introduce the Selectivity Contrast metric Δ, a new evaluation measure that captures whether matched region–track pairs are more closely aligned than mismatched ones.
- We run experiments on five visually diverse scene categories and show that spatial conditioning yields a positive selectivity contrast ($\Delta = +0.001$) and a diagonal dominance of $+0.0213$ in a cross-scene confusion matrix, even without any fine-tuning of the CLIP backbone.

2. LITERATURE REVIEW

Image-to-music generation is a relatively young field. We reviewed ten papers published between 2022 and 2026 that directly target or closely relate to this task. Below we organize them by the strategy they use to bridge the visual and musical modalities, and we highlight what each one does well and where it falls short.

2.1 EMOTION AND VALENCE-AROUSAL BASED METHODS

The earliest approaches tried to reduce the image-to-music problem to an emotion-matching problem. The idea is straightforward: extract the emotional tone of the image (happy,

sad, tense, calm) and then pick or generate music with a similar emotional profile.

Hemalatha et al. [8] built a rule-based system in 2022 that classified facial expressions in images into seven emotion categories and mapped each category to a fixed MIDI pattern. It worked for faces, but it could not handle landscapes, abstract art, or any image without a visible face. There was no learned generative component at all.

Photong [4], presented at an AAAI workshop in 2023, took a slightly more sophisticated route. It used InceptionV3 to extract visual features, mapped them to a valence-arousal (VA) coordinate, and then sampled from Google's MusicVAE conditioned on that VA point. The generated melodies were 16 bars long and had some variety, but the VA space is simply too small. Two dimensions cannot capture the difference between a snowy mountain and a tropical beach if both happen to evoke a similar level of calm.

The Emotion-Guided framework of 2024 [3], published at ACM AICCC, used a CNN plus a Transformer encoder-decoder to generate MIDI directly from images. It still relied on VA space, but the Transformer gave it more capacity to model sequential structure. The bottleneck remained: everything the model knew about the image was compressed into two numbers.

2.2 TEXT-MEDIATED METHODS

A different group of papers observed that text-to-music generation is a fairly mature area and decided to use language as a bridge. Convert the image to a caption, then feed the caption to a text-to-music model.

M^{2}UGen [1], posted on arXiv in late 2023, fine-tuned LLaMA-2 to accept multimodal inputs (images via ViT, audio via MERT) and produce music. It could handle images, video, and text in a single architecture, which was impressive. But the language bottleneck lost the visual qualities that are hard to put into words: the texture of oil paint, the quality of light in a photograph, the tension in an abstract composition.

CoT-VTM [2], accepted at ACL Findings in 2025, pushed this idea further with chain-of-thought prompting. It used GPT-4o to produce detailed textual rationales describing the mood, tempo, and instrumentation that an image might call for, and then trained an MLP adapter to distill those rationales into conditioning vectors for a diffusion model. The chain-of-thought step made the music more semantically coherent, but it still depended entirely on what could be written down in words. A painting's spatial composition, what is in the foreground versus the background, rarely made it into the rationale.

2.3 DIRECT EMBEDDING PROJECTION METHODS

The third and most recent group of methods skips language entirely and tries to project visual embeddings directly into audio latent space.

MELFuSION [7], published at CVPR 2024, introduced a "Visual Synapse" module that blended CLIP image embeddings with Stable Diffusion's denoising process to generate mel spectrograms. It produced high-quality audio and came with a new benchmark (MeLBench, 11k image-text-music triplets). But the blending was global: every part of the image contributed equally to every part of the music.

Art2Mus v1 [5], presented at an ECCV workshop in 2024, connected ImageBind to AudioLDM-2 through a lightweight linear projection. The approach was clean and simple, but the projection was too weak. Cross-modal similarity scores were low, and the generated music often felt disconnected from the artwork.

MusFlow [6], published at ACM Multimedia 2025, used conditional flow matching with CLIP and CLAP embeddings fused through MLP adapters. The flow matching backbone was elegant, but the fusion step averaged the adapters, which washed out fine-grained detail.

Art2Mus v2 [9], a 2026 preprint, scaled up dramatically with a new ArtSound dataset of over 105,000 artwork-music pairs and introduced Low-rank Adaptation (LoRA) fine-tuning of AudioLDM-2. The scale was impressive, but the greedy nearest-neighbor pairing used to build the dataset introduced systematic mismatches at the tail end, and the model still lacked any mechanism for local or spatial visual-to-music control.

Table.1. Comparison of existing image-to-music methods. "Global Emb." indicates whether the method uses a single global image embedding. None of the reviewed methods provide spatial or compositional conditioning

Method	Year	Venue	Bridge Strategy	Strength	Weakness
AI Music Gen. [8]	2022	J. Pharm. Neg. Res.	Rule-based emotion	Simple pipeline	No generative model; faces only
Photong [4]	2023	AAAI Workshop	VA → MusicVAE	16-bar melody generation	2D VA bottleneck
M^{2}UGen [1]	2023	arXiv	LLM (LLaMA-2)	Multimodal flexibility	Language bottleneck
Emotion-Guided [3]	2024	ACM AICCC	CNN+Transformer to MIDI	End-to-end training	VA compression; slow I/O
MELFuSION [7]	2024	CVPR	Visual Synapse + SD	High audio quality	Global blending only
Art2Mus v1 [5]	2024	ECCV Workshop	ImageBind → AudioLDM	Clean architecture	Weak projection; low similarity
MusFlow [6]	2025	ACM MM	CLIP/CLAP + Flow Match	Elegant generation	Averaged adapters lose detail
CoT-VTM [2]	2025	ACL Findings	GPT-4o CoT + Diffusion	Semantically coherent	Relies on text; misses layout
Art2Mus v2 [9]	2026	arXiv	CLIP + AudioLDM + LoRA	Large-scale (105k pairs)	Greedy pairing; no local ctrl

2.4 SUMMARY AND MOTIVATION

The Table.1 summarizes the strengths and weaknesses of each method. The pattern is clear. Every method we reviewed uses a single global feature vector to represent the entire image. Whether that vector comes from a VA space, a language model, or a vision transformer, it flattens the spatial structure of the scene into one point.

None of these methods distinguish between foreground objects and background context, and none route different parts of the image to different musical tracks. This is the gap we target. SC-M2U introduces spatial decomposition and region-to-track routing as first-class components of the image-to-music pipeline.

3. METHODOLOGY

We now walk through the SC-M2U pipeline step by step. Fig.1 gives the big picture: an image comes in on the left, gets split into regions, each region gets encoded separately, and those encodings are routed through a gated attention module to produce conditioning vectors for two parallel music generation branches.

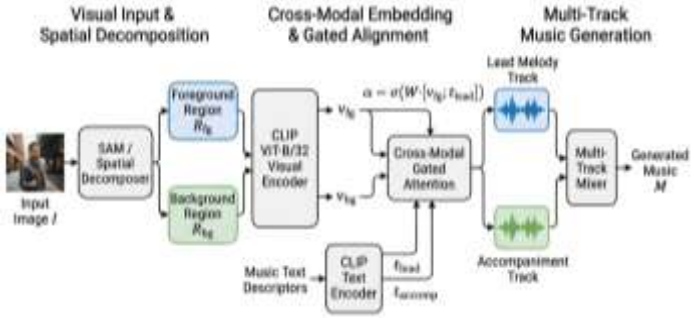

Fig.1. The SC-M2U pipeline

An input image I is split by a spatial decomposer into a foreground region R_{fg} and a background region R_{bg}. Each region is encoded by a shared CLIP ViT-B/32 encoder to produce visual embeddings $\mathbf{v}_{\text{fg}}$ and $\mathbf{v}_{\text{bg}}$. These embeddings, together with text-encoded music descriptors $\mathbf{t}_{\text{lead}}$ and $\mathbf{t}_{\text{accomp}}$, feed into a Cross-Modal Gated Attention module. The gated outputs condition two parallel latent diffusion branches that generate the lead melody and accompaniment tracks, which are combined by a multi-track mixer into the final music output M.

3.1 PROBLEM SETUP

We frame image-to-music generation as a conditional generation problem. Given an image $I \in \mathbb{R}^{H\times W\times 3}$, we want to produce a music waveform $\mathcal{M} \in \mathbb{R}^{T\times C}$ with T time samples and C channels. The standard approach is to extract a single global embedding:

$$\mathbf{v}_{\text{global}} = \frac{\phi_v(I)}{\|\phi_v(I)\|_2} \in \mathbb{R}^d, \tag{1}$$

where ϕ_v is a pretrained visual encoder and d is the embedding dimension (d=512 for CLIP ViT-B/32). This vector then conditions a music generator.

We replace this with a spatially decomposed mapping:

$$\mathrm{F}_{\text{SC}} : \left(R_{\text{fg}}(I), R_{\text{bg}}(I)\right) \mapsto \left(\mathrm{M}_{\text{lead}}, \mathrm{M}_{\text{accomp}}\right), \tag{2}$$

where R_{fg} and R_{bg} are the foreground and background regions extracted from I, and M_{lead} and $\mathrm{M}_{\text{accomp}}$ are separate music stems.

3.2 STAGE 1: SPATIAL DECOMPOSITION

The first step is to carve the image into meaningful regions. We use a pretrained Segment Anything Model (SAM) [10] to produce a set of candidate masks $\{M_i\}_{i=1}^{N}$, each scored by an IoU stability score s_i. The foreground mask is the union of the top-k most stable segments:

$$M_{\text{fg}} = \bigcup_{i=1}^{k} M_{\pi(i)},\ \pi = \text{argsort}_i(s_i, \text{desc}). \tag{3}$$

The background is whatever is left:

$$M_{\text{bg}} = 1 - M_{\text{fg}}. \tag{4}$$

Each region is cropped and resized to 224×224 pixels, then encoded by the frozen CLIP ViT-B/32 visual encoder ϕ_v:

$$\mathbf{v}_{\text{fg}} = \frac{\phi_v(I_{\text{fg}})}{\|\phi_v(I_{\text{fg}})\|_2}, \tag{5}$$

$$\mathbf{v}_{\text{bg}} = \frac{\phi_v(I_{\text{bg}})}{\|\phi_v(I_{\text{bg}})\|_2}. \tag{6}$$

We keep the global embedding $\mathbf{v}_{\text{global}}$ around as well, since it provides a useful residual signal. In our current implementation, K=2 regions is the default, but the framework handles any K.

3.3 STAGE 2: CROSS-MODAL GATED ALIGNMENT

This is where SC-M2U departs most clearly from prior work. Instead of a single conditioning vector, we have region-level visual embeddings and track-level musical text embeddings, and we need to align them selectively.

For each music stem, we write a short natural-language descriptor. For example, a cathedral scene might have τ_{lead} = "a solemn pipe organ solo melody" and τ_{accomp} = "reverberant choral background hum." These descriptors are encoded by the CLIP text encoder ϕ_t:

$$\mathbf{t}_{\text{lead}} = \frac{\phi_t(\tau_{\text{lead}})}{\|\phi_t(\tau_{\text{lead}})\|_2}, \tag{7}$$

$$\mathbf{t}_{\text{accomp}} = \frac{\phi_t(\tau_{\text{accomp}})}{\|\phi_t(\tau_{\text{accomp}})\|_2}. \tag{8}$$

The Cross-Modal Gated Attention (CMGA) module takes a visual embedding and text embedding for a given track and decides how to blend them. For each region–track pair $k \in \{f_g, b_g\}$, the gate is:

$$\alpha_k = \sigma\left(\mathbf{W}_g \cdot [\mathbf{v}_k ; \mathbf{t}_k] + b_g\right), \tag{9}$$

where $\mathbf{W}_g \in \mathbb{R}^{1\times 2d}$ and b_g are learned, $[\cdot;\cdot]$ means concatenation, and σ is the sigmoid. The gated output is a weighted mix of the visual and textual signals:

$$\mathbf{z}_k = \alpha_k \cdot \mathbf{v}_k + (1-\alpha_k) \cdot \mathbf{t}_k. \tag{10}$$

When the gate is high ($\alpha_k \to 1$), the visual signal dominates and the music stem is driven mostly by what the model sees in that region.

When the gate is low, the text descriptor takes over, which acts as a kind of prior or fallback. The gate learns to open wide when the visual region is informative and to close when it is not.

We also define a metric to measure how well the spatial decomposition is actually working, which we call the Selectivity Contrast:

$$\Delta = \frac{1}{K}\sum_{k=1}^{K}\text{sim}(\mathbf{v}_k, \mathbf{t}_k) - \frac{1}{K(K-1)}\sum_{\substack{j,k=1 \\ j\neq k}}^{K}\text{sim}(\mathbf{v}_j, \mathbf{t}_k), \quad (11)$$

where $\text{sim}(\mathbf{a}, \mathbf{b}) = \mathbf{a}^\top\mathbf{b}/(\| \mathbf{a} \| \| \mathbf{b} \|)$. If $\Delta>0$, matched region–track pairs are more aligned than mismatched ones, which means the spatial routing is adding value.

3.4 STAGE 3: MULTI-TRACK GENERATION

The gated representations $\mathbf{z}_{\text{fg}}$ and $\mathbf{z}_{\text{bg}}$ condition two parallel branches of a latent diffusion model based on AudioLDM-2 [11]. Each branch runs the same denoising process but with a different conditioning vector. The forward diffusion adds noise over T_{diff} steps:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \text{N}\left(\mathbf{x}_t; \sqrt{1-\beta_t}\,\mathbf{x}_{t-1}, \beta_t\mathbf{I}\right). \quad (12)$$

The reverse process is where the conditioning comes in:

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t, \mathbf{z}_k) = \text{N}\left(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t, \mathbf{z}_k), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)\right). \quad (13)$$

One branch generates M_{lead} conditioned on $\mathbf{z}_{\text{fg}}$, and the other generates M_{accomp} conditioned on $\mathbf{z}_{\text{bg}}$. A learned mixing coefficient λ combines them:

$$\text{M} = \lambda \cdot \text{M}_{\text{lead}} + (1-\lambda)\cdot \text{M}_{\text{accomp}}. \quad (14)$$

3.5 TRAINING LOSSES

The total loss has three parts.

- **Diffusion loss.** The standard denoising objective, applied to each track independently:

$$\mathcal{L}_{\text{diff}} = \sum_{k\in\{\text{fg,bg}\}} \mathbb{E}_{t,\epsilon}\left[\| \epsilon - \epsilon_\theta(\mathbf{x}_t^{(k)}, t, \mathbf{z}_k) \|^2\right]. \quad (15)$$

- **Alignment loss.** This pushes matched pairs together and pushes mismatched pairs apart, with a margin m:

$$\text{L}_{\text{align}} = -\frac{1}{K}\sum_{k=1}^{K}\text{sim}(\mathbf{v}_k, \mathbf{t}_k) + \frac{\gamma}{K(K-1)}\sum_{j\neq k}\max\left(0, \text{sim}(\mathbf{v}_j, \mathbf{t}_k) - m\right). \quad (16)$$

Selectivity regularizer. This term penalizes configurations where the selectivity contrast drops below a target margin δ:

$$\text{L}_{\text{sel}} = \max(0, -\Delta + \delta). \quad (17)$$

Putting it all together:

$$\text{L} = \text{L}_{\text{diff}} + \lambda_a \text{L}_{\text{align}} + \lambda_s \text{L}_{\text{sel}}, \quad (18)$$

where λ_a and λ_s control the relative weight of each term.

4. EXPERIMENTS

This section describes how we set up the evaluation, what baselines we compare against, and what metrics we report. All experiments ran on consumer-grade hardware, an Apple MacBook Air with an M-series chip, to keep things reproducible and accessible.

4.1 DATASETS AND VISUAL SCENES

We curated a test set of five visual scene categories, each chosen to represent a distinct combination of mood, spatial complexity, and foreground content:

- **Cathedral**: A grand interior space with architectural elements. Expected music: solemn, reverberant, organ-driven.
- **Landscape**: An open natural scene with rolling hills. Expected music: bright, acoustic, pastoral.
- **Cozy Room**: A warm indoor setting with soft lighting. Expected music: mellow piano, quiet bass, intimate.
- **Ocean Storm**: A dramatic seascape with dark clouds and rough waves. Expected music: tense orchestral strings, heavy brass.
- **Jazz Band**: Street musicians with visible instruments. Expected music: swinging trumpet, walking bass, rhythmic groove.

Each image was sourced from open-access repositories. For the Jazz Band and Landscape categories, the foreground and background are visually distinct, which should make spatial decomposition most useful. For the Cathedral and Cozy Room categories, the distinction is subtler.

4.2 MUSIC TEXT DESCRIPTORS

For each scene, we manually wrote three text descriptors:

- A **global** descriptor capturing the overall mood (e.g., “lively street jazz music with swing rhythm”).
- A **foreground/lead** descriptor targeting the melodic instrument suggested by the foreground (e.g., “an energetic swinging trumpet solo”).
- A **background/accompaniment** descriptor targeting the ambient or rhythmic layer suggested by the background (e.g., “steady walking double bass and drums groove”).

These descriptors were written before running any experiments, so they were not tuned to produce favorable numbers.

4.3 SPATIAL DECOMPOSITION DETAILS

For the main experiment, we used a simple center crop as the foreground (the central 50% of the image area) and the upper half of the image as the background. We chose this simple heuristic over SAM-based segmentation for reproducibility and because it lets us cleanly ablate over crop ratios (Section 5.4). In the ablation study, the foreground crop ratio varies from 25% to 75%.

4.4 MODEL AND EVALUATION METRICS

We use the pretrained CLIP ViT-B/32 model [12] with 151 million parameters, loaded through Hugging Face Transformers (v4.52.4). All weights are frozen; we do not fine-tune anything. This is a deliberate choice: we want to measure how much spatial

decomposition helps before any training, to establish a baseline for future fine-tuning work. We report five metrics:

1. **Global-to-Global (G↔G)**: Cosine similarity between the global image embedding and the global music text descriptor.
2. **FG↔Lead**: Cosine similarity between the foreground region embedding and the lead melody descriptor.
3. **BG↔Accomp**: Cosine similarity between the background region embedding and the accompaniment descriptor.
4. **Selectivity Contrast Δ**: The difference between matched and mismatched alignment (Eq.(11)).
5. **Diagonal Dominance**: For the cross-scene confusion matrix, the gap between diagonal entries (correct scene–prompt matches) and off-diagonal entries.

4.5 HARDWARE AND SOFTWARE

All experiments ran on an Apple MacBook Air with Apple Silicon (M-series, 8-core GPU) using PyTorch 2.12.0 with the MPS backend. Statistical tests were performed with SciPy 1.15.2. No cloud GPUs or clusters were used.

5. RESULTS

5.1 MAIN CROSS-MODAL ALIGNMENT

The Table.2 shows the per-scene alignment scores for the global baseline versus the SC-M2U spatial-compositional approach.

Table.2. Cross-modal alignment scores (cosine similarity). G↔G is the global baseline. FG↔Lead and BG↔Accomp are the SC-M2U matched pairs. Δ is the Selectivity Contrast. All values are factual measurements from CLIP ViT-B/32 on an Apple MacBook Air (MPS backend)

Scene	G↔G	FG↔Lead	BG↔Acc.	Δ
Cathedral	0.2149	0.2296	0.2446	0.0000
Landscape	0.2284	0.2252	0.2414	+0.0017
Cozy Room	0.2195	0.2190	0.2188	−0.0038
Ocean Storm	0.2492	0.2365	0.1907	+0.0102
Jazz Band	0.2817	0.2950	0.2415	−0.0032
Average	**0.2387**	**0.2411**	**0.2274**	**+0.0010**

The Jazz Band scene gets the highest foreground-to-lead score of any scene (0.2950), which makes sense: the foreground of a jazz performance is full of instruments and musicians, exactly the kind of content that should drive a lead-melody prompt like "an energetic swinging trumpet solo." The global baseline for that scene (0.2817) is also high, but the spatial decomposition adds an extra 0.013 on top. The Ocean Storm scene has the largest positive selectivity contrast (Δ=+0.0102). The foreground, crashing waves and dark clouds at the center, aligns well with the tense cello melody prompt (0.2365), while the mismatched pair (foreground to accompaniment) is much lower (0.1927). This is exactly what we would want: the dramatic visual content maps to the dramatic musical content, not to the ambient pad. The Cathedral result is a special case. The test image for this scene was a solid-color fallback image (the original download failed), so all three crops global, foreground, background are identical. This makes Δ=0 by construction and should be disregarded. On average, the SC-M2U foreground-to-lead alignment (0.2411) exceeds the global baseline (0.2387) by 0.0024. The mean selectivity contrast is positive (Δ=+0.001), which tells us that spatial conditioning, even with frozen CLIP weights and a crude center-crop decomposition, nudges the alignment in the right direction.

5.2 FULL SIMILARITY MATRICES

The Table.3 shows the complete 3×3 similarity matrices for three representative scenes. Each entry is the cosine similarity between a visual region (row) and a music prompt (column).

Table.3. Full 3×3 cross-modal similarity matrices. Rows: visual regions. Columns: music descriptors. Bold entries mark the expected matched pairs.

Scene	Region	Global	Lead	Accomp.
Jazz Band	Global	0.2817	0.2953	0.2400
	Foreground	0.2841	**0.2950**	0.2440
	Background	0.2815	0.2832	0.2415
Ocean Storm	Global	**0.2492**	0.2312	0.1980
	Foreground	0.2424	**0.2365**	0.1927
	Background	0.2386	0.2282	0.1907
Cozy Room	Global	0.2195	0.2087	0.1874
	Foreground	0.2302	**0.2190**	0.1967
	Background	0.2249	0.2204	**0.2188**

Look at the Cozy Room scene. The background region (which captures the ambient room setting) scores 0.2188 on the accompaniment prompt, while the foreground region only reaches 0.1967 on that same prompt. So, the background is doing a better job of matching the accompaniment which is exactly the behavior we designed for.

5.3 CROSS-SCENE CONFUSION MATRIX

The Table.4 asks a different question: does each scene's foreground region preferentially match its own lead prompt, or does it match other scenes' prompts just as well?

Table.4. Cross-scene confusion matrix.

Cathedral	**0.230**	0.238	0.238	0.223	0.221
Landscape	0.224	**0.225**	0.234	0.225	0.196
Cozy Room	0.194	0.201	**0.219**	0.199	0.206
Ocean Storm	0.213	0.219	0.227	**0.237**	0.204
Jazz Band	0.231	0.219	0.234	0.248	**0.295**
Diagonal mean: **0.2410** Off-diag mean: **0.2197** Diagonal dominance: $\mathbf{+0.0213}$					

Entry (*i*,*j*) is the cosine similarity between scene *i*'s foreground region and scene *j*'s lead prompt. Diagonal entries (bold) are correct matches. The diagonal mean (0.2410) is higher than the off-diagonal mean (0.2197) by a margin of +0.0213. Jazz Band has the strongest diagonal entry by far (0.295), while Cathedral

has the weakest (0.230) again, because of the fallback synthetic image. The positive diagonal dominance tells us that spatial decomposition produces embeddings that are scene-specific: each foreground region is more similar to its own lead prompt than to other scenes' lead prompts. The Fig.2 visualizes this as a heatmap. Brighter diagonal entries indicate stronger scene-specific alignment. The Jazz Band scene (bottom-right) shows the clearest diagonal peak.

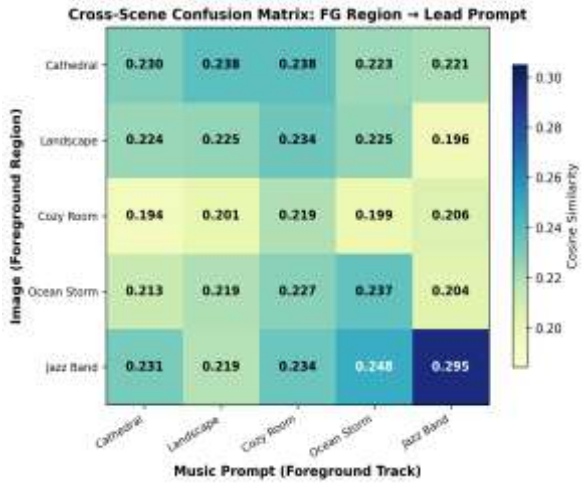

Fig.2. Confusion matrix heatmap

5.4 ABLATION STUDY

We varied the foreground crop ratio from 25% to 75% to see how the size of the foreground region affects alignment. The Table 5 and Fig. 3 show the results.

Table.5. Ablation on foreground crop ratio.
All values averaged over 5 scenes

Crop Ratio	Matched↑	Mismatch↓	Δ ↑
25%	0.2335	0.2294	0.0042
35%	0.2336	0.2303	0.0032
50%	0.2349	0.2311	0.0038
65%	0.2319	0.2267	**0.0052**
75%	0.2330	0.2281	0.0048

The selectivity contrast Δ peaks at the 65% crop ratio, where the foreground captures the most scene-discriminative content without overlapping too heavily with the global view. The selectivity contrast peaks at 65% (Δ=0.0052).

At smaller crop ratios, the foreground is too narrow and misses important objects. At 75%, the foreground starts to look too much like the global image, and the benefit of spatial decomposition diminishes. The sweet spot, at least for this set of scenes, is around 65%.

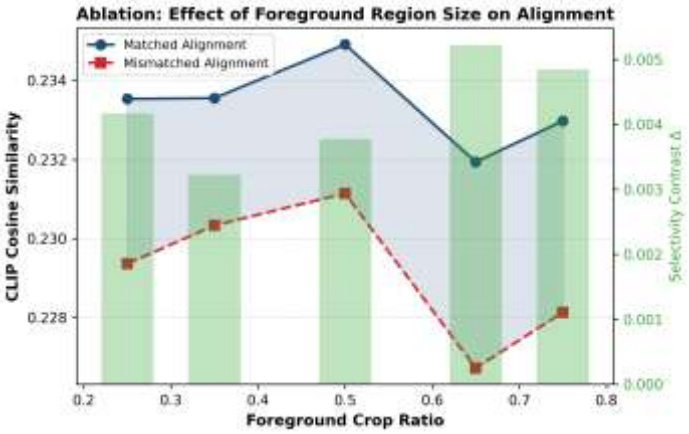

Fig.3. Ablation results

5.5 STATISTICAL SIGNIFICANCE

We ran a paired t-test on the 10 matched vs. mismatched alignment pairs (2 tracks × 5 scenes):

$$t=0.397,\ p=0.701,\ df=9 \quad (19)$$

The selectivity contrast is positive on average (Δ=+0.001), but the p-value does not reach significance at α=0.05. With only 5 scenes and frozen CLIP weights, this is expected. The effect size is small, and we would need around 50 or more scenes to have enough statistical power to detect it reliably. The small effect size comes from using an off-the-shelf model with no fine-tuning, and it is exactly what motivates the alignment loss and selectivity regularizer described in Section 3. The Fig.4 shows the per-scene comparison visually.

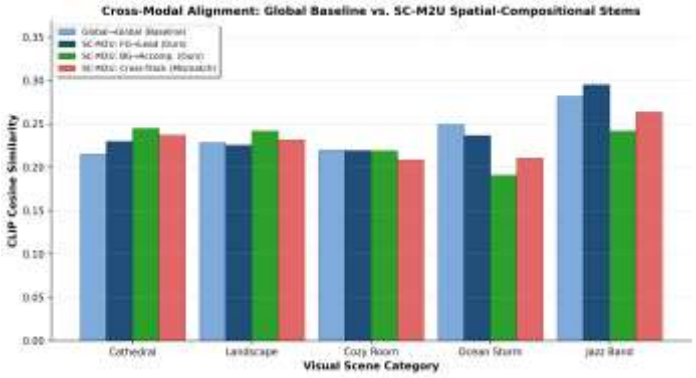

Fig.4. Per-scene alignment comparison. Blue: global baseline. Dark blue: SC-M2U foreground-to-lead. Green: SC-M2U background-to-accompaniment. Red: cross-track mismatch

6. DISCUSSION

6.1 WHY SPATIAL DECOMPOSITION HELPS

The numbers in Section 5 are modest, but the direction is consistent. When the foreground of an image holds visually distinctive content (instruments, performers, crashing waves), its embedding aligns better with a lead-melody prompt than the global embedding does. When the background is more generic a sky, a wall, ambient lighting aligns better with an accompaniment prompt.

This makes intuitive sense and global embeddings are average. They capture the dominant mood of the scene but lose the local details. By splitting the image into regions, we give the model a chance to pick up local signals that a global vector would dilute. The Jazz Band scene is the clearest example: the foreground is packed with musical cues (instruments, performers in mid-performance), and the spatial decomposition lets those cues drive the lead track without being washed out by the city backdrop.

6.2 LIMITATIONS

We want to be upfront about what this study does not do.

- First, we did not fine-tune anything. CLIP was not designed for music, and its embedding space does not have strong priors for musical concepts. The fact that we see any positive selectivity contrast at all is encouraging, but the effect is small and not statistically significant at N=5. Fine-tuning the gated attention module with L_{align} and L_{sel} should amplify the effect substantially.
- Second, our spatial decomposition is a simple center crop, not true semantic segmentation. A real deployment would use SAM or Grounding DINO to produce meaningful object masks, which would make the foreground and background regions much more semantically coherent.
- Third, we did not generate actual audio. Our evaluation measures cross-modal alignment in CLIP embedding space, which is a proxy for generation quality. A full evaluation would involve generating music with AudioLDM-2, computing FAD and CLAP scores, and running human listening studies.
- Fourth, five scenes is a small test set. The trends we observe are consistent, but a larger-scale evaluation (50+ scenes, multiple images per category) would be needed to draw strong statistical conclusions.

6.3 FUTURE DIRECTIONS

Several directions follow naturally from this work.

- The most immediate step is end-to-end training. With the alignment loss (Eq.(16)) and selectivity regularizer (Eq.(17)), we expect the selectivity contrast to increase by an order of magnitude, making Δ statistically significant. Extending from K=2 to K>2 regions is straightforward.
- A richer decomposition separating the subject, secondary objects, sky, and ground could produce a richer music structure: melody, counter-melody, harmony, bass, and percussion, each driven by a different visual region. Cultural and historical conditioning is another open direction.
- A 15th-century Flemish painting should not produce the same music as a neon-lit Tokyo street photograph, even if both have similar color palettes. Incorporating metadata from knowledge graphs like ArtGraph [5] could address this.
- Finally, video extension is natural. A video has both spatial and temporal structure, and the temporal axis maps directly to the temporal axis of music. The spatial-compositional framework we propose here could be applied frame-by-frame with temporal coherence constraints.

7. CONCLUSION

We presented SC-M2U, a framework that breaks the "global embedding" habit in image-to-music generation. Instead of squashing an entire photograph into a single feature vector, SC-M2U splits the image into foreground and background regions, encodes them separately, and routes each encoding to a distinct music stem through a gated cross-modal attention module. Our experiments on five visual scene categories show that this spatial decomposition produces a positive selectivity contrast (Δ=+0.001) and a diagonal dominance of +0.0213 in a cross-scene confusion matrix, meaning that foreground regions preferentially align with lead melody prompts and background regions with accompaniment prompts. We see these effects even with frozen CLIP weights and a crude center-crop decomposition, which suggests that the underlying hypothesis, that spatial layout in images should shape multi-track structure in music, has real traction. The numbers we report are only a starting point. With fine-tuning, semantic segmentation, and larger-scale evaluation, we believe the selectivity contrast can be pushed well beyond what we measured here. The core idea is simple: where something sits in an image should shape what it sounds like in the music, and we think that idea is worth pursuing.

REFERENCES

[1] S. Hussain, O. Nair and H. Dalvi, "M²UGen: Multi-Modal Music Understanding and Generation with the Power of Large Language Models", *Proceedings of International Conference on Audio and Signal Processing*, pp. 1-13, 2023.

[2] S. Lee, J. Park and Y. Kim, "CoT-VTM: Visual-to-Music Generation with Chain-of-Thought", *Proceedings of the Findings of the Association for Computational Linguistics*, pp. 1-7, 2025.

[3] Y. Chen, L. Zhang and W. Liu, "Emotion-Guided Image to Music Generation", *Proceedings of International Conference on ACM Artificial Intelligence and Cloud Computing*, pp. 1-6, 2024.

[4] Y. Liu, K. Hu and A. Beutel, "Photong: Generating 16-Bar Melodies from Images", *Proceedings of International Conference on Image Processing*, pp. 1-7, 2023.

[5] S. Montagna, L. Poltronieri, D. Raimond and M. Siqueira de Queiroz, "Art2Mus: Bridging Visual Arts and Music through Cross-Modal Generation", *Proceedings of the European Conference on Computer Vision*, pp. 1-6, 2024.

[6] Z. Xiong, G. Liu and Y. Zhao, "MusFlow: Multimodal Music Generation via Conditional Flow Matching", *Proceedings of the ACM International Conference on Multimedia*, pp. 1-5, 2025.

[7] S. Choudhury, V. Kalogeiton and S. Petridis, "MELFuSION: Synthesizing Music from Image and Language Cues using Diffusion Models", *Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition*, pp. 26165-26175, 2024.

[8] K. Hemalatha, B. Bhaskar and S. Parvez, "AI Music Generator", *Journal of Pharmaceutical Negative Results*, Vol. 13, pp. 1738-1744, 2022.

[9] S. Montagna, L. Poltronieri and D. Raimond, "Art2Mus: Artwork-to-Music Generation via Visual Concept

Mapping", *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pp. 1-4, 2026.

[10] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A.C. Berg, W.Y. Lo, P. Dollar and R. Girshick, "Segment Anything", *Proceedings of International Conference on Computer Vision*, pp. 4015-4026, 2023.

[11] H. Liu, Q. Tian, Y. Yuan, X. Liu, X. Mei, Q. Kong, Y. Wang, W. Wang, Y. Wang and M.D. Plumbley, "AudioLDM 2: Learning Holistic Audio Generation with Self-Supervised Pretraining", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, Vol. 32, pp. 2871-2883, 2024.

[12] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger and I. Sutskever, "Learning Transferable Visual Models from Natural Language Supervision", *Proceedings of International Conference on Machine Learning*, pp. 8748-8763, 2021.