# 艺术品到音: 艺术品到音乐生成：通过视觉条件和大规模跨模态对齐

IVAN RINALDI，意大利巴里阿尔多·莫罗大学MATTEO MENDULA，西班牙加泰罗尼亚电信技术中心NICOLA FANELLI，意大利巴里阿尔多·莫罗大学FLORENCE LEVÉ，法国皮卡第儒勒·凡尔纳大学MATTEO TESTI，英格兰人工智能创业构建者（AIVB）GIOVANNA CASTELLANO，意大利巴里阿尔多·莫罗大学GENNARO VESSIO，意大利巴里阿尔多·莫罗大学

arXiv:2602.17599v1 [cs.CV] 19 Feb 2026

多模态深度学习推动音乐生成取得了显著进展，使模型能够根据文本合成音频，近来还能够根据图像合成音频。然而，现有的图像条件系统存在两个根本局限：（*i*）它们通常基于自然照片训练，因此难以捕捉艺术作品更丰富的语义、风格和文化内容；（*ii*）大多数系统依赖图像到文本转换阶段，将语言作为语义捷径。此方法虽简化了条件设置，却阻碍了直接的视觉到音频学习。针对这些不足，我们推出了艺术之声（ArtSound），这是一个大型多模态数据集，包含 105,884 组艺术作品—音乐配对，并附有双模态描述；该数据集通过扩展艺术图谱（ArtGraph）和 Free Music Archive 构建。我们还提出了艺术品到音乐生成框架 Art2Mus，这是首个专为直接艺术品到音乐生成而设计的框架，无需图像到文本翻译或基于语言的语义监督，即可将数字化艺术作品映射为音乐。该框架将视觉嵌入投射到潜在扩散模型的条件空间中，从而仅凭视觉信息引导音乐合成。实验结果表明，Art2Mus 生成的音乐在音乐性上连贯、风格上统一，并能体现源艺术作品的显著视觉线索。尽管绝对对齐分数仍低于文本条件系统——这符合预期，因为移除语言监督会大幅增加难度——Art2Mus 仍实现了具有竞争力的感知质量和有意义的跨模态对应。本研究确立了直接视觉到音乐生成这一独特且富有挑战性的研究方向，并提供了可支持多媒体艺术、文化遗产和人工智能辅助创作实践应用的资源。

CCS 概念： • **计算方法论** → **人工智能**； **计算机视觉**。

其他关键词和短语：艺术品到音乐生成、多模态学习、跨模态对齐、音频合成、扩散模型。

作者联系信息：Ivan Rinaldi，巴里阿尔多·莫罗大学，意大利巴里，i.rinaldi4@phd.uniba.it；Matteo Mendula，加泰罗尼亚电信技术中心，西班牙巴塞罗那，matteo.mendula@cttc.cat；Nicola Fanelli，巴里阿尔多·莫罗大学，意大利巴里，nicola.fanelli@uniba.it；Florence Levé，皮卡第儒勒·凡尔纳大学，亚眠，法国，florence.leve@u-picardie.fr；Matteo Testi，人工智能创业构建者（AIVB），英格兰伦敦，matteo.testi@aivb.ai；Giovanna Castellano，巴里阿尔多·莫罗大学，意大利巴里，giovanna.castellano@uniba.it；Gennaro Vessio，巴里阿尔多·莫罗大学，意大利巴里，gennaro.vessio@uniba.it。

特此授权免费制作本作品的全部或部分内容的数字副本或纸质副本，供个人或课堂使用，前提是不得为营利或商业利益而制作或分发副本，且副本首页须附有本声明及完整引文。对于本作品中由作者以外的其他方拥有版权的部分，必须尊重其版权。注明出处后可以摘录。若要以其他方式复制、重新出版、发布到服务器或分发至列表，须事先获得具体许可和/或支付费用。请向 permissions@acm.org 申请许可。© 2026 版权由所有者/作者持有。出版权已授权给 ACM。提交给 ACM 的稿件

**ACM 引用格式：**
Ivan Rinaldi、Matteo Mendula、Nicola Fanelli、Florence Levé、Matteo Testi、Giovanna Castellano 和 GennaroVessio。2026。艺术*2*音乐：通过视觉条件和大规模跨模态对齐实现艺术品到音乐生成。1，1（2026年2月），24 页。https://doi.org/XXXXXXX.XXXXXXX

## 1 引言

音乐生成长期以来一直是计算创造力最具表现力的形式之一。音乐融合了结构、情感与文化背景，因此其自动生成本质上极具挑战性。将视觉艺术作品转化为音乐更是难上加难，这并非因为两种模态彼此无关，而是因为它们之间的对应关系间接且高度结构化。视觉艺术通过构图、色彩、纹理、风格和历史语境传达意义，唤起对音乐表达至关重要的情感与叙事 [69]。这些共同的感知与情感维度表明，图像和声音之间存在有意义的跨模态对应关系。然而，这种关系很少是一一对应或直接可观察的；它们源于由人类感知和艺术惯例塑造的高层抽象，因此若缺少足够丰富的表征，就难以对其建模。

多模态深度学习的最新进展显著改进了生成模型，使其能够根据文本、图像或多模态输入合成符号音乐 [28, 45, 58] 和自然音频 [2, 15, 41]。基于文本的条件输入提供了强有力的语义锚定，而大型语言模型（LLM）则通过将异构输入转换为丰富的文本描述 [40, 42]，进一步强化了这一范式。然而，仍存在两个局限。

首先，大多数图像条件音乐生成系统主要基于自然照片进行训练，这些照片的语义相对字面化且以物体为中心。此类数据仅能部分反映艺术作品所具有的象征性、风格性以及文化层次丰富的语义，从而限制了现有方法在艺术领域中的适用性。

其次，许多流程依赖于图像到文本转换阶段。尽管这一策略有效，但它将语言作为中间语义瓶颈引入。语言描述不可避免地抽象掉了模糊性、纹理和风格细节——这些元素往往难以用简洁的语言加以表达，却是艺术表达的核心。因此，依赖文本中介的图像到音乐系统或许能生成音乐上连贯的输出，却无法保留源艺术作品更为丰富的表现结构。从这个意义上说，语言充当了一种语义捷径：它通过引入明确的、由人类定义的概念来简化多模态条件控制，代价是丢弃了可能与音乐生成相关、且无法用语言表达的视觉线索。

现有的图像到音乐生成尝试说明了这些问题。早期方法依赖手工设计的图像特征，并结合基于RNN的解码器 [62], ，因而表现多样性有限。在 [27], 中，艺术作品图像被转换为情感描述符和图像字幕，用以引导音乐生成。更新近的方法 [35, 60]利用多模态大语言模型（多模态大语言模型）描述图像，随后采用文本到音乐生成器。尽管有效，此类流程将核心的多模态对齐交由语言模型完成，因而无法学习视觉语义与音乐结构之间的直接关系。因此，视觉结构与生成的音乐之间的依赖关系是由语言抽象所中介的，而非直接从视觉数据中推断得出。此外，这一方向的进展受到明确将艺术作品与音乐配对的大规模数据集稀缺的制约。尽管文本到图像 [61, 64] 和文本到音频 [2, 3], 已有大量资源，但仅有小型艺术作品—音乐集合得到了探索 [53], ，从而限制了对直接跨模态学习的系统性研究。

受这些不足的启发，本研究着眼于推动艺术品到音乐生成的两个必要条件：(*i*) 通过构建大型版多模态数据集，促进系统性研究；以及 (*ii*) 引入一套 Manuscript submitted to ACM

能够直接 根据数字化艺术作品生成音乐的框架，无需文本中间表示。这一方案定义了一个难度更高的多模态问题。与能够获取明确语义监督的文本条件系统不同，仅依赖图像的模型必须直接从视觉嵌入中推断情感、风格和构成线索，并将其转化为有意义的声学模式。

为支持这项工作，我们推出了 艺术之声，这是一个包含 105,884 组艺术作品—音乐配对的大型数据集，其中还包含图像和音频描述，并经过两种新型对齐度量指标的验证。基于这一资源，我们提出了 艺术品到音乐，这是首个专为直接实现艺术品到音乐生成而设计的系统。该框架通过专用的图像到音频接口，将从艺术作品中提取的视觉嵌入注入潜在扩散模型的条件空间。通过将 CLIP [51] 或 ImageBind [23] 嵌入与 GPT-2 的 [52] 表示空间对齐，艺术品到音乐便能够仅凭视觉信息引导音乐合成。这一设计有意移除了基于语言的语义监督，迫使模型直接从视觉表征中学习跨模态对应关系，而非依赖语言抽象。

在我们的初步工作 [53], 中，我们直接以图像作为文本到音频模型的条件输入。该研究揭示了本文旨在解决的两个局限：(*i*) 简单的线性投影不足以捕捉有意义的艺术作品特征；以及 (*ii*) 包含 10,000 组配对的数据集缺乏足够的风格和音乐多样性。

本文其余部分安排如下：第2 节回顾相关文献。第3 节介绍数据集和图像描述流程。第4 节展示 $\mathcal{A}rt2\mathcal{M}us$ 架构。第5节报告实验设置和结果。第6 节总结全文。

## 2 相关工作

近年来，生成式人工智能的进展日益聚焦于对异构数据模态（包括文本、图像和音频）进行建模与整合 [6, 21, 30, 70]。条件音乐生成是一个迅速发展的研究方向，模型在此方向中根据外部输入合成音乐。这一范式涵盖文本到音乐方法 [41, 56] 以及新兴的图像到音乐方法 [12, 42, 63]。其目标是生成在音乐上连贯、在语义上有意义的音频，并体现条件模态所传达的语境。虽然文本条件通常凭借明确的语义结构实现较强的对齐，但图像条件仍更具挑战性，因为视觉信息抽象、多层次且往往不够明确。

已有若干模型探索了图像条件或跨模态音乐生成。IM2WAV [57] 通过提取 CLIP [51]嵌入，并将其输入自回归令牌式音频模型，生成与一张或多张图像相匹配的音频。M2M-Gen [56] 通过大语言模型生成文本描述，为日本漫画创作背景音乐；这些描述依据场景边界、对话以及从角色表情中推断出的情感生成。在 [36], 中，图像首先由大语言模型转换为歌词，随后由基于规则的作曲系统生成乐谱，并导出为 MusicXML。MelFusion [12] 通过在解码器层中将 DDIM 反演图像潜变量与不断演化的音频潜变量融合，将视觉语义整合进基于扩散的文本到音乐框架。Mozart's Touch [35] 支持图像、视频和文本：它首先为视觉内容生成描述文本，再将这些描述文本转换为音乐描述，最后使用文本到音乐模型生成音乐。

值得注意的是，这些系统大多使用自然照片或视觉内容直观的图像进行训练和评估，其语义主要以物体和场景为中心。这类数据只能部分体现艺术作品所特有的符号抽象、风格多样性以及层次丰富的历史语义。因此，现有的图像条件音乐生成方法可能无法有效泛化至艺术领域；在这一领域中，意义往往通过构图、风格和文化语境来表达，而非依赖明确的物体。

此外，尽管近期取得了进展，大多数图像条件音乐生成流程仍依赖文本描述作为中间模态。图像通常会先转换为图像字幕，再由文本到音频模型处理。虽然这种策略能够重复利用成熟的架构，但也带来了固有的信息瓶颈。一旦视觉内容被压缩为文本，描述文本中未能捕捉的属性——例如细粒度的风格模式、构成细微差异或氛围特质——便无法在音频生成过程中恢复。从信息论的角度来看，中间文本表示只能保留或舍弃视觉信息，却无法引入描述文本中未编码的视觉细节。因此，直接以视觉特征为条件，或许能够保留那些原本会被语言抽象过滤掉的风格与构成线索。

我们的初步研究 [53] 证明，通过视觉嵌入对文本到音频模型进行条件控制，可以实现直接的艺术品到音乐生成。然而，该方法受限于简单的投影机制和相对较小的数据集，因此促使我们开发表达能力更强的视觉—音频对齐策略，并构建规模更大的资源。

除了生成流程，图像描述也已成为跨模态对齐的核心机制。图像描述系统旨在将图像或音频等非结构化输入转换为文本描述，以捕捉语义内容、情感或风格属性。早期系统（包括音频描述Transformer [47]和 Midge [49], ）依赖传统架构，生成的结果往往较为笼统。近期进展则由多模态大语言模型推动，它们能够为图像（例如 LLaVA [39], 、Qwen2.5-VL [4]）和音频（例如 Pengi [18], 、Qwen2-Audio [13]）生成丰富且贴合语境的描述文本。Qwen2.5-Omni [66] 等新兴全模态架构在统一框架中整合文本、图像和音频，从而实现更复杂的跨模态推理。

这些进展得益于大型版图像描述数据集的发布 [11, 22, 32, 50], ，这些数据集为图像和音频理解提供了监督信号。人们也推出了特定领域的图像描述数据集。在视觉艺术领域，ArtCap [44] 提供众包描述，涵盖印象派和表现主义绘画；Iconclass图像描述数据集 [10] 为超过 8.6 万件艺术作品提供图像学标注；ArtEmis [1] 则收录了超过 8 万件艺术作品的情感反应和自由文本解释。在音乐领域，MusCaps [46], 、MusicBench [48], 和 MusicEval [38] 等数据集提供了有关音乐属性、情绪、流派和结构的语言描述，促进了自然语言理解和音乐评估。

尽管有这些资源，明确将艺术作品与音乐配对的数据集仍然寥寥无几，而且规模有限 [53]。因此，该领域仍主要侧重于文本中介对齐，而对直接视觉到音乐生成的系统性研究仍不充分。

我们的工作通过同时构建大型版艺术作品–音乐数据集，并提出一种直接的视觉条件框架，来弥补这些不足。该框架无需语言监督即可学习跨模态对应关系。借助 A*rt2*M*us*，我们直接以从数字化艺术作品中提取并投影的视觉嵌入为条件，驱动潜在扩散模型，从而实现以视觉表征而非文本抽象为基础、富有表现力且高效的艺术品到音乐生成。

## 3 材料

本节介绍所提出的 A*rtSound* 数据集的构建。该数据集是一种统一的多模态资源，将数字化艺术作品与音乐曲目配对，辅以图像和音频描述，并通过基于嵌入的配对策略进行对齐。我们首先介绍数据来源和描述生成流程，随后介绍用于验证描述质量的指标，最后讨论艺术品–音乐配对。

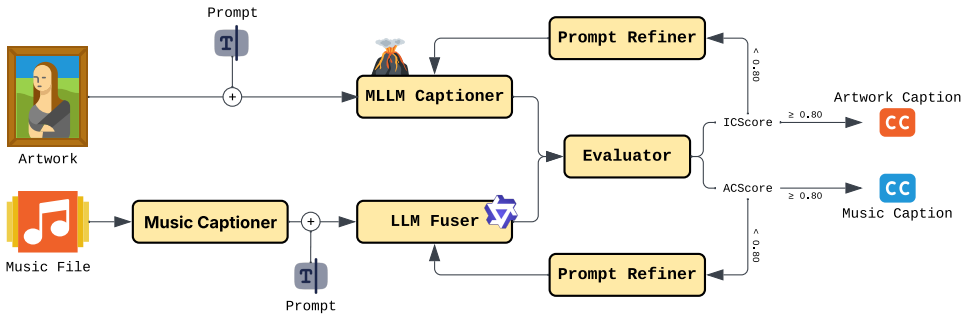



图 1. 图像和音频的统一描述生成流程。图像由多模态大语言模型在结构化提示的引导下生成描述；音频则由片段级音乐描述生成器生成描述，再由大语言模型融合。所有描述均使用所提出的 I CS$core$和 A CS$core$指标进行评估；低于阈值的描述会使用精炼提示重新生成。

### 3.1 数据来源

为了构建用于艺术品到音乐生成的大型版数据集，我们首先从两个开放访问的资源库中获取数据。对于艺术作品，我们使用了 $\mathcal{A}rt\mathcal{G}raph$ [7], 一个源自 WikiArt 和 DBpedia 的知识图谱，其中包含 116,475 件涵盖 18 个流派和 32 种风格的数字化艺术作品，并附有丰富的上下文元数据，例如艺术家、风格和流派。对于音乐，我们采用了 Free Music Archive (FMA) 数据集的大型版 [16], 该数据集包含来自 16,341 位音乐人的 106,574 首采用知识共享许可的 30 秒音轨，涵盖 16 个顶级流派和 161 个子流派。移除损坏的音频文件后，保留了 105,884 首音轨。

### 3.2 描述生成

图像和音频描述通过统一但针对不同模态的流程生成（图 1）。图像描述由 LLaVA 生成 [39], , 提示其生成丰富的描述，涵盖艺术作品的内容、情绪、风格线索及可能的艺术影响。音频描述使用 LP-MusicCaps 生成 [19], , 该模型会生成简短的 10 秒片段描述；随后使用 Qwen 将这些描述合并为连贯的整体描述3 [67]。

为确保描述可靠性，我们使用专门设计的自定义多指标评分，对两种模态进行评估，以反映对齐质量的不同方面。对于图像描述，我们将 ICS$core$ 定义为 CLIP-Score [26] 和 PAC-Score [55]的加权组合。CLIP-Score 使用 CLIP 嵌入衡量图像与其描述之间的整体语义一致性，而 PAC-Score 则评估描述中表达的正向语义属性是否存在于图像中。结合这些指标，可以均衡地评估高层语义和细粒度属性一致性。形式化表示如下：

$$\mathcal{I}\mathcal{C}\mathcal{S}core(x, y) = \gamma \cdot \text{CLIP-Score}(x, y) + (1 - \gamma) \cdot \text{PAC-Score}(x, y), \tag{1}$$

其中， $x$表示艺术作品， $y$表示生成的描述文本， $\gamma = 0.35$ 表示通过经验选取的权重。

对于音频描述，我们定义了 ACS核心 ，用于评估融合描述对 LP-MusicCaps 生成的片段级描述的概括效果。ROUGE-1 [37] 衡量词汇内容的保留程度，而 BERT-Score [68]则在嵌入空间中衡量语义相似度，从而能够适应释义改写。二者的加权组合为：

$$\mathcal{A}\mathcal{C}\mathcal{S}core(x, y) = \alpha \cdot \text{ROUGE-1}(x, y) + (1 - \alpha) \cdot \text{BERT-Score}(x, y), \tag{2}$$

其中， $x$ 表示片段级图像字幕集合， $y$ 表示融合描述， $\alpha = 0.30$ 用于控制词汇重叠与语义重叠的相对权重。

我们最初也曾考虑将 ROUGE-L 纳入 ACS核心，但后来将其弃用，因为其对最长公共子序列的强调会惩罚有效的释义，即使融合描述在语义上与片段级描述保持一致，分数也会被降低。

图像字幕若其 $\mathcal{ICS}$核心 或 $\mathcal{ACS}$核心 低于 0.80 的阈值，则使用精炼提示重新生成。精炼过程是发出一条新的提示词，明确告知大语言模型/多模态大语言模型：先前生成的描述文本未达到所需的质量阈值，并且存在具体问题。这一基于反馈的提示迫使模型重新生成描述文本，并明确确保其符合所提出的要求。精炼提示仅在首次尝试失败后才会使用，因为若事先使用，在逻辑上并不成立——彼时尚未生成或评估任何描述文本。重新生成最多允许进行 3 次迭代，因为后续尝试的收益递减。表 1 汇总了与图像字幕相关的高层统计信息。

即使达到允许的重新生成次数上限，未达到阈值的描述文本仍会保留在数据集中。此外，数据集 A*rtSound* 中还包含相应的 ICS*core* 和 ACS*core* 数值，以便进一步了解描述质量，并支持后续筛选或分析。

图像和音频的生成描述文本示例分别列于表2 和 3中。值得注意的是，图像描述的细节程度远高于音频描述。这种差异可能是由描述生成模型的选择造成的；如附录 A.1–A.2所示，其他模型也观察到了类似差异。

我们针对图像描述生成和音频描述，使用不同模型开展了额外实验，并对音频描述中的描述融合进行了实验。对于图像描述生成，最初考虑了部分模型 [5, 8] ，但后来因其过时而被舍弃；另一些模型 [4, 66] 生成的描述要么过长且不够平衡，要么过于笼统，且不支持提示条件化 [34]。示例见附录A.1。对于音频描述，我们确定了若干模型 [13, 17, 42, 66] 。然而， [42] 由于显存需求高且生成的描述质量低，因此被排除。其余模型 [13, 17, 66] 生成的描述通常准确，但过于详尽，且常常重复相同的概念；有时还会向用户提问，甚至在描述中编造回答，因此输出不适用。此类描述示例见附录A.2。对于音频描述融合，我们还评估了较小的大语言模型（LLaMa3.1-8B-Instruct [24] 和 Qwen3-8B [67]），将其作为 Qwen3-14B 的替代模型。采用$\mathcal{ACS}$核心（公式2），Qwen3-14B 的表现更佳，在保留含义和原始措辞方面均表现更强。表4 报告了这些结果，包括 $\mathcal{ACS}$核心及其底层指标 ROUGE-1 和 BERT-Score，以供完整参考。

表 1. A*rtSound*中图像和音频描述的统计数据，包括描述长度（以字符数表示）和质量指标。图像描述的指标包括：CLIP-Score、PAC-Score 和综合指标 I CS*core*。音频描述的指标包括：BERT-Score、ROUGE-1 和综合指标 ACS*core*。表中列出了高于或低于接受阈值（0.80）的数值。

| 统计数据 | 图像字幕 | | | | 音频描述 | | | |
|---|---|---|---|---|---|---|---|---|
| | 长度 | CLIP-Score | PAC-Score | $\mathcal{ICS}$核心 | 长度 | BERT-Score | ROUGE-1 | $\mathcal{ACS}$核心 |
| 最小值 | 234.00 | 0.3699 | 0.5165 | 0.4918 | 82.00 | 0.8474 | 0.1000 | 0.6234 |
| 平均值 | 950.18 | 0.7821 | 0.8431 | 0.8217 | 411.76 | 0.9321 | 0.6894 | 0.8593 |
| 最大值 | 2302.00 | 1.0000 | 1.0000 | 1.0000 | 1084.00 | 0.9940 | 0.9771 | 0.9889 |
| > 阈值 | – | – | – | 78,845 | – | – | – | 95,837 |
| < 阈值 | – | – | – | 37,630 | – | – | – | 10,047 |

表2. 使用 LLaVA1.6-Mistral-7B 生成的艺术声音图像描述示例。

| 图像 | 图像描述 |
|---|---|
| <br>沙漠<br>Richard Pousette-Dart | 这件艺术作品是一幅色彩鲜明、风格抽象的构图，看起来像是一幅画。画中呈现了密集排列的几何形状、线条和圆点，使用了多种色彩，包括红色、蓝色、黄色和黑色。形状与线条相互交叉、重叠，营造出一种复杂而深邃的感觉。这件艺术作品传达出混乱而复杂的情绪，形状与线条的抽象特质也为解读留下了空间。这幅画中展现的关键艺术技法与风格令人联想到立体主义风格，其特点是运用几何形状和线条营造深度感和维度感。鲜明的对比色也增强了这件艺术作品的整体视觉效果。这幅画可能与立体主义时期或流派有关，该流派活跃于20世纪初。这个时期以强调抽象形式以及运用几何形状来表现三维世界而闻名。这件艺术作品的抽象自然以及几何形状和线条的运用表明，它可能受到了立体主义运动的影响，或是立体主义运动的一部分。 |
| <br>蒂耶塞维尔山丘<br>卡米耶·毕沙罗 | 这件艺术作品描绘了一幅宁静的乡村景观。画面主要是一片高草地，一条蜿蜒小径穿过其中。小径将观者的视线引向一排树木在远处，营造出深度与透视感。上方的天空布满了云，为场景增添了宁静平和的氛围。艺术作品传达的情绪或情感是一种安宁与孤寂。草、树木和天空等自然元素以松散的印象派风格呈现，其特点是笔触清晰可见，并注重光线与色彩之间的交融。这种风格通常与印象主义时期联系在一起，该时期力求捕捉自然世界中光线与色彩转瞬即逝的变化。艺术作品的题材，即乡村景观，是许多艺术运动中的常见主题，其中包括印象派画家，他们常常力求捕捉自然世界之美。松散的印象派风格与宁静的题材相结合，表明这件艺术作品可能体现了印象派风格，尤其是如果它创作于19世纪末或20世纪初。 |

表3. 使用LP-MusicCaps + Qwen3生成的A*rtSound* 音频描述示例。

| 音频 | 图像字幕 |
|---|---|
| **音频 #1 (001/001556)** | 这是一段低质量录音，呈现了一首氛围歌曲，由持续的合成垫音和弦、持续的合成贝斯、柔和的钢琴和弦以及一段简短的拨弦旋律组成。听起来令人平静、放松，且令人着迷——就像你会在瑜伽课程或水疗中心听到的音乐。 |
| **音频 #2 (001/001140)** | 这是一首由男声演唱主旋律的音乐作品，并配有打击乐演奏简单的四四拍节奏，贝斯演奏和弦的根音，钢琴则在背景中演奏分解和弦。情绪浪漫，适合用于浪漫电影。此外，它还带有一种一段布鲁斯风格的器乐摇滚音乐作品，由男歌手在主旋律中深情演唱，主旋律由钢琴演奏，电吉他和贝斯吉他则在背景中伴奏。节奏背景由缓慢节奏速度的原声鼓点构成，氛围忧郁，适合用于剧情电视剧的原声带，尤其适合角色试图挣脱束缚的场景。 |

### 3.3 艺术品–音乐配对

鉴于已有的带字幕图像和音频集合，我们的目标是构建语义上有意义的图像–音频对。为此，我们使用 ImageBind [23]的表示，评估了几种基于嵌入的相似度方法。由于 ImageBind 将多个模态（例如文本、图像和音频）的表示嵌入到共享表示空间中，因此可以直接比较这三种模态。事实上，我们通过比较这些不同模态的嵌入，评估了多种配对策略。

表 4. 音频描述融合的大语言模型评估期间计算的平均指标。指标包括 ROUGE-1、BERT-Score 和自定义的 ACS*core*。最佳分数以**粗体**显示，下划线标注的数值表示接近最佳值。

| LLM | ROUGE-1 | BERT-Score | $\mathcal{ACS}core$ |
|---|---|---|---|
| LLaMa3.1-8B-Instruct | 0.5373 | 0.9056 | 0.7951 |
| Qwen3-8B | 0.6301 | 0.9260 | 0.8372 |
| Qwen3-14B | **0.6870** | **0.9312** | **0.8579** |

配对依据的是某一数据源中嵌入的余弦相似度；该数据源可以是图像或音频文件，具体模态则可以是原始数据或图像字幕。配对过程从相似度最高的项目开始。我们特别关注音频项目，将每个音频项目与最相似的图像配对。

该过程可形式化如下。设 $\mathrm{A} = \{a_1, \ldots, a_N\}$ 表示音频项目集合，$\mathrm{I} = \{i_1, \ldots, i_M\}$ 表示图像集合。每个项目 $x \in \mathrm{A} \cup \mathrm{I}$ 都会嵌入到一个维度为 $d$的共享空间中，生成嵌入$e(x^{v} \in \mathbb{R}^d)$，其中 $v \in \{\text{raw}, \text{caption}\}$ 表示所使用的模态。图像 $i$与音频项目 $a$在模态 $v_i$ 和 $v_a$下的余弦相似度定义为：

$$sim(i, a; v_{\mathrm{i}}, v_{\mathrm{a}}) = \frac{e_{\mathrm{i}}^{(v_i)} \cdot e_{\mathrm{a}}^{(v_a)}}{\|e_{\mathrm{i}}^{(v_i)}\| \, \|e_{\mathrm{a}}^{(v_a)}\|}. \tag{3}$$

对于每个音频项目 $a_j \in \mathrm{A}$，配对时选择图像：

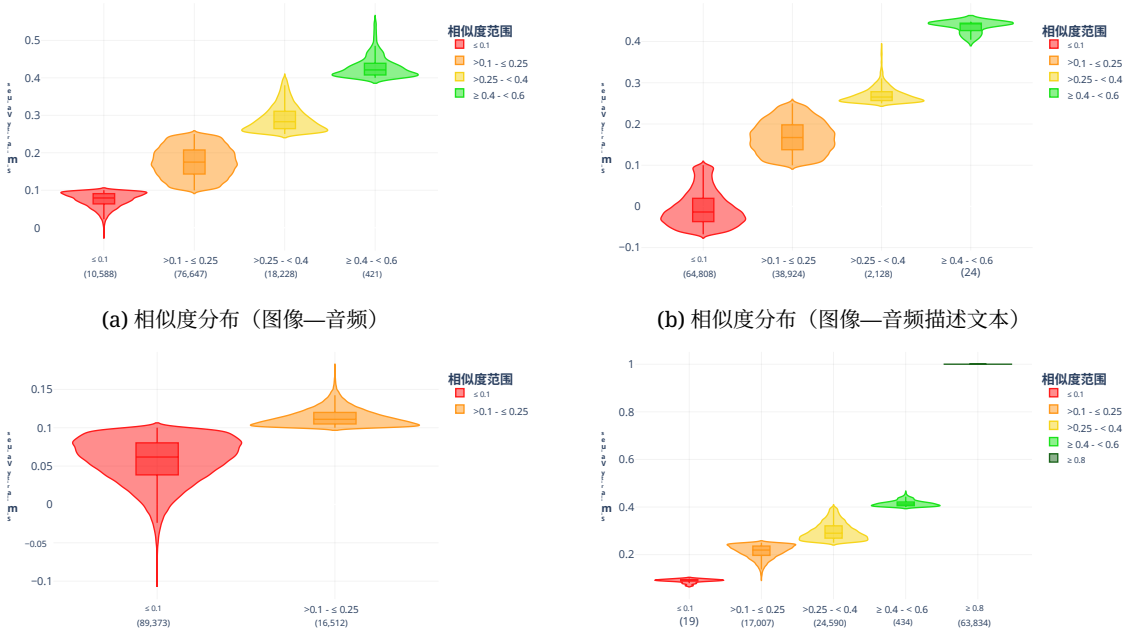
$$p_j^* = \underset{i \in I_{\text{avail}}}{\operatorname{argmax}} \, sim(i, a_j; v_{\mathrm{i}}, v_{\mathrm{a}}), \tag{4}$$

其中，$\mathrm{I}_{\text{avail}}$ 是当前未配对图像的集合，选中的图像 $p_j^*$ 随后会从$\mathrm{I}_{\text{avail}}$中移除。该过程持续进行，直到所有 $a_N \in \mathrm{A}$ ($N = 105,\ 884$)均完成配对。贪心策略的一个局限出现在配对过程接近尾声时：匹配完相似度值最高的项目后，剩余项目无论相似度分数多低，都必须进行配对。为便于后续分析这一影响，我们将计算所得的相似度值纳入最终数据集。

我们分析了计算得到的相似度，以评估潜在的模态差距，并探究同一模态内的比较（如文本–文本）是否比跨模态比较（如图像–文本或音频–文本）产生更高的相似度分数。各配对方法的客观结果见表 5，图 2 则展示了相似度值在固定区间内的分布，清晰呈现了数值落在较低区间和较高区间的数量。

表5. 各配对方法在图像、音频、图像描述和音频描述上的相似度统计（最小值、最大值、均值和标准差）汇总。所有方法均使用相应模态的ImageBind嵌入。**粗体**数值表示最高值，下划线标出的数值表示第二高值。

| 配对方法 | 图像 | 图像描述文本 | 音频 | 音频描述 | 相似度值 Min | Max | Avg | 标准差 |
|---|---|---|---|---|---|---|---|---|
| 图像–音频 | ✓ | – | ✓ | – | −0.0216 | 0.5518 | 0.1864 | 0.0697 |
| 图像–音频描述 | ✓ | – | – | ✓ | −0.0676 | 0.4482 | 0.0644 | 0.0970 |
| 图像描述–音频 | – | ✓ | ✓ | – | −0.1011 | 0.1794 | 0.0658 | **0.0344** |
| 图像描述–音频描述 | – | ✓ | – | ✓ | **0.0699** | **1.0000** | **0.7081** | 0.3613 |




(a) 相似度分布（图像—音频）

(b) 相似度分布（图像—音频描述文本）

(c) 相似度分布（图像描述文本—音频）

(d) 相似度分布（图像描述文本—音频描述文本）

图 2. 图像-音频配对策略的相似度得分分布小提琴图。使用 ImageBind 根据图像、音频和描述文本嵌入的组合计算得分，并分为低（≤0.25）、中（0.25–0.6）和高（≥0.6）范围，以显示哪些模型-模态配对产生了较强或较弱的对齐。

总体而言，评估的配对策略得到的相似度分数相对较低（均值范围为 0.0644 至 0.7081），表明图像与音频之间存在显著的模态差距。图像描述与音频描述配对时，对齐效果最强，平均相似度最高（0.7081），这说明相比原始模态，文本表示能够提供更一致的共享空间。相比之下，原始图像与音频描述的组合表现最差，凸显了通过不匹配的表示直接弥合视觉信息与听觉信息之间差距的难度。

最后，对计算出的配对进行定性检查后，我们选择了图像-音频方法作为最佳方法。尽管其平均相似度得分并非最高，但这些配对项在语义上比其他方法产生的配对更为一致。因此，通过该方法获得的配对被用于构建 *ArtSound*，即我们的最终数据集，其中包含 105,884 个艺术品-音乐配对，每一对都配有高质量的图像和音频描述。

我们还评估了其他嵌入模型，包括 CLIP [51],CLAP [20],以及两个文本编码器（RoBERTa [43]和 Modern-BERT [65]），但定性分析表明，这些模型经常产生不相关的图像-音频配对，因此我们将其舍弃。此外，我们还测试了其他配对方法，详见附录B，其中包括：利用一小部分*ArtGraph*艺术作品以及来自 Free Music Archive数据集的音乐曲目，将艺术作品与音频频谱图和音频生成图像进行配对。尽管总体结果较差，但部分生成的配对表明，艺术品风格与音乐流派具有相似的理念。更多细节见附录 B。

## 4 艺术*2*音乐

本节介绍 艺术*2*音乐 （图 3）的架构。这是一个受 AudioLDM 2 [40], 生成范式启发的多模态框架，但经过重新设计，以解决一个根本不同的问题：直接从艺术作品生成音乐。AudioLDM 2 遵循文本到音频范式，而 艺术*2*音乐 则围绕视觉到音频形式构建，并有意去除语言监督。我们没有扩展原有任务，而是采用其潜在扩散骨干网络作为预训练生成先验，并重新设计条件接口，以支持视觉表征的直接跨模态对齐。核心贡献不在于修改扩散架构本身，而在于重新定义条件信息的表示与对齐方式。具体而言，我们提出一种视觉到 LoA（音频语言）接口，学习将图像嵌入编码到最初为音频语义开发的潜在令牌空间中，使扩散模型能够在新的跨模态条件机制下运行。

### 4.1 基础架构

AudioLDM 2 [40] 将潜在扩散建模 [54] 应用于音频生成。音频信号首先被转换为梅尔频谱图，并使用变分自编码器（VAE）编码为潜在向量 $z_0$ 。扩散过程逐渐将 $z_0$ 扰动为 $z_T \sim \mathrm{N}(0,\ 1)$，并训练基于 UNet 的去噪器 $\epsilon_\theta$ 来逆转这一过程。其目标是

$$\mathcal{L}_{\mathrm{LDM}} = \mathbb{E}_{z_0,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, c)\|_2^2\right], \tag{5}$$

其中， $c$表示条件信息， $t$是从 {1、.....、 $T$}中采样的扩散时间步。与文本到图像扩散模型不同，AudioLDM 2 在 LoA 嵌入中对用于生成的条件信息进行编码

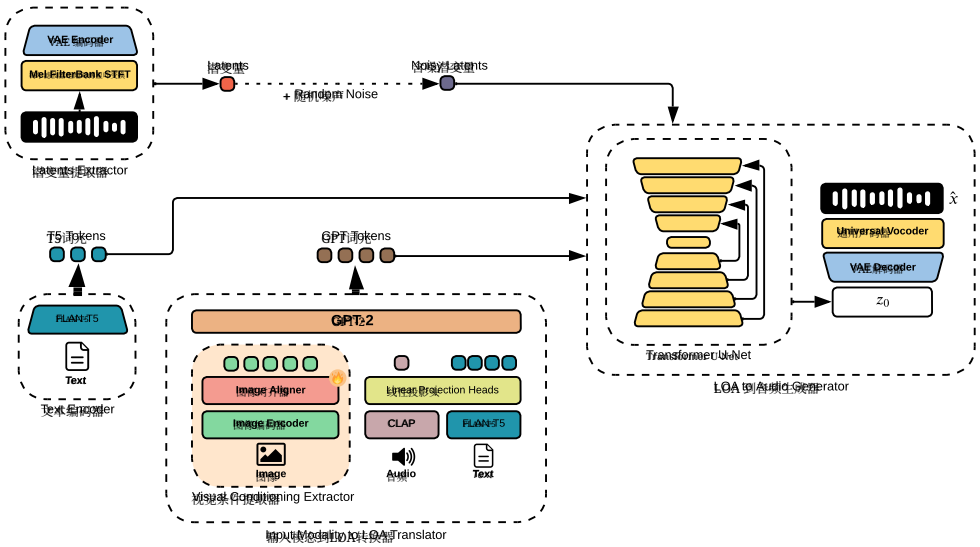
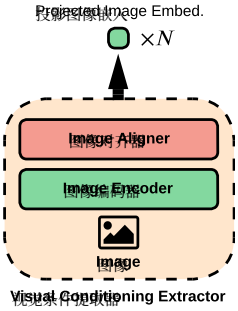
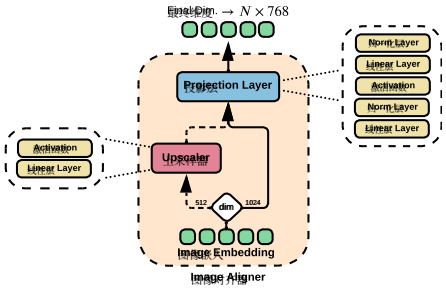



图 3. A艺术*2*音乐 架构概览。受 AudioLDM 2 生成范式的启发，A艺术*2*音乐 重新设计了条件接口，以实现直接的艺术品到音乐生成。视觉条件提取器 和图像对齐器 通过学习将图像嵌入映射到 GPT-2 的 LoA 嵌入空间，建立跨模态对齐，无需文本中间表示。预训练的 AudioLDM 2 主干网络用作冻结的生成先验。GPT-2 作为模态无关的 LoA 转换器，在潜在扩散过程中整合视觉条件。音乐由潜在表示提取器 编码为梅尔频谱图潜变量，再由*LoA* 到音频生成器合成为音频。




(a) 视觉条件提取器。一个预训练图像编码器（例如，CLIP 或 ImageBind）生成一个嵌入嵌入，随后映射到 GPT-2 的词元空间（维度为 768）。




（b）图像对齐器。如果输入嵌入为 512 维（CLIP），则会升维至 1024 维；否则（ImageBind）则跳过此步骤。一个 pro- 投影层随后将所有嵌入映射为 $N$ 个词元映射到 GPT-2 的 768 维空间中。

图 4。视觉条件提取器和图像对齐器模块。

空间——一种连续的音频信号表示，同时捕捉语义和声学信息。该空间使用 AudioMAE 构建 [29], 这是一种自监督掩码自编码器，通过重建被掩码的频谱图块来训练，并学习有意义的音频表示。

对于文本到音频生成，我们对 GPT-2 模型进行微调，使其能够将文本提示词直接映射到 LoA 空间。这种建模方式简化了目标函数，因为扩散模型以 LoA 嵌入而非文本表示作为条件。在训练过程中，条件信号在真实 LoA 特征和 GPT-2 预测的 LoA 特征之间交替使用，从而促进跨模态泛化，同时保留 LoA 嵌入丰富的语义信息。

去噪器 $\epsilon_\theta$ 是 T-UNet（Transformer UNet），与图像生成中使用的模型类似 [54]。卷积下采样和上采样层与包含自注意力和交叉注意力的 Transformer 模块交替排列，使模型能够将 LoA 特征融入去噪过程。去噪后的潜在表示由 VAE 解码器解码为梅尔频谱图，预训练的 HiFiGAN 声码器 [33] 再将频谱图转换为最终的音频波形。

### 4.2 视觉条件提取器

AudioLDM 2 最初仅用于处理文本条件，这意味着其内部的 LoA 表示缺乏原生机制来纳入视觉信息。仅通过图像–音频对对模型进行微调，并不能实现跨模态对齐，因为去噪器和 GPT-2 需要以分词文本空间表示的条件。为克服这一架构限制，$\mathcal{A}rt2\mathcal{M}us$ 引入了视觉条件提取器，这是一个新组件，为视觉语义进入 LoA 空间建立了专用通路（图 4a）。

给定一张艺术作品图像，预训练视觉编码器会提取一个维度为 512（CLIP）或 1024（ImageBind）的嵌入。随后，这些嵌入会被传入图像对齐器 （图 4b），其由以下部分组成：

- 上采样器：将 512 维 CLIP 嵌入扩展至 1024 维；

- 投影层：将每个 1024 维嵌入映射为 GPT-2 的 768 维嵌入空间中的 $N$ 个词元。

这种模块化设计确保了与异构视觉编码器的兼容性，并提供了稳定的中间表示。实证结果表明，在投影前使用 1024 维的过渡空间，比直接投影低维嵌入生成的音乐更忠实，风格也更协调。

### 4.3 音乐生成过程

给定一幅艺术作品，视觉条件提取器会生成源自图像的条件词元，GPT-2 将其转换为 LoA 嵌入 $c_{\text{vis}}$。与此同时，文本提示词“代表这幅艺术作品内容的音乐”由 FLAN-T 编码5 [14],，生成 $c_{\text{text}}$。最终的条件向量 $c$ 结合了这两种模态。

推理时，生成从高斯噪声 $z_T \sim \mathrm{N}(0,\ 1)$ 开始，并经历反向扩散过程。为提高效率并实现确定性采样，我们采用 DDIM调度器 [59],，以更少的去噪步骤实现高质量生成。设 $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ 表示方差调度系数的累积乘积，其中每个 $\alpha_t = 1 - \beta_t$ 控制时间步 $t$保留的信号量，$\beta_t$ 为预定义的噪声方差。因此，$1 - \bar{\alpha}_t$ 表示累积噪声水平。在每个时间步 $t$，模型预测噪声分量 $\epsilon_\theta(z_t,\ t,\ c)$，并计算干净潜变量的估计值：

$$\hat{z}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}} \left( z_t - \sqrt{1 - \bar{\alpha}_t}\, \epsilon_\theta(z_t, t, c) \right). \tag{6}$$

随后，DDIM更新规则会将 $z_t$ 确定性地映射为 $z_{t-1}$，如下所示

$$z_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\, \hat{z}_0 + \sqrt{1 - \bar{\alpha}_{t-1}}\, \epsilon_\theta(z_t, t, c). \tag{7}$$

迭代至 $t = 0$后，最终潜变量 $\hat{z}_0$ 会通过 VAE解码器和神经声码器解码为梅尔频谱图并转换为波形。

### 4.4 训练算法

训练 A*rt2*M*us*需要学习一种跨模态对齐，而 AudioLDM 2 从根本上缺乏这种能力：模型必须将细粒度的视觉嵌入与 LoA空间中编码的音乐结构关联起来。这构成了一项新的学习任务，因为 AudioLDM 2 从未接触过图像，也没有关联视觉语义与音乐属性的先验知识。

具体而言，训练依赖由艺术作品、与之配对的音乐曲目以及固定引导提示组成的三元组；该提示与推理时使用的提示相同（即 “代表这幅艺术作品内容的音乐”）。此外，还会提供负向提示（例如 “低质量”），以启用无分类器引导；还可以进一步补充内容，以抑制不希望出现的流派。负向提示有助于引导模型避开不理想的方面，从而提升整体质量，并确保输出更贴合预期意图。每首配对的音乐曲目都会转换为梅尔频谱图，并使用 AudioLDM 2 VAE 编码为干净潜变量 $z_0$。时间步 $t$ 从均匀分布中采样，并添加噪声以获得 $z_t$。扩散模型以视觉 LoA 词元和辅助文本嵌入为条件，按照公式（5）训练，以预测所添加的噪声。

至关重要的是，仅更新图像对齐器 的参数。AudioLDM 2 的所有组件，包括去噪器和 GPT-2 LoA 翻译器，均保持冻结。这在保留预训练生成先验的同时，使系统能够学习从艺术作品到 LoA 条件特征的稳健映射。通过这一重新表述，$\mathcal{A}rt2\mathcal{M}us$稿件已提交至 ACM

证明，文本条件扩散先验可通过表示对齐被重新用于直接的视觉到音乐生成，而无需改变底层生成架构。

在实践中，根据使用 ImageBind 还是 CLIP 作为视觉编码器，要么只训练投影层，要么同时训练上采样器和投影层。训练遵循标准的去噪扩散概率模型目标，即模型预测添加到干净潜变量 $z_0$上的噪声。前向扩散过程定义为

$$z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\, \epsilon, \qquad \epsilon \sim \mathcal{N}(0, I), \tag{8}$$

其中 $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$ 是方差调度系数的累积乘积， $1-\bar{\alpha}_t$ 表示累积噪声水平。对于给定时间步 $t$，基础噪声预测损失为

$$\mathcal{L}_{\text{MSE}}(t) = \|\epsilon - \epsilon_\theta(z_t, t, c)\|_2^2 . \tag{9}$$

因此，时间步 $t$ 处的信噪比为

$$\text{SNR}_t = \frac{\bar{\alpha}_t}{1-\bar{\alpha}_t}, \tag{10}$$

这直接反映了保留信号的相对强度 $\sqrt{\bar{\alpha}_t}$与注入噪声 $\sqrt{1-\bar{\alpha}_t}$。遵循[25]，时间步损失被重新加权为

$$\mathcal{L}_t = \frac{\min(\text{SNR}_t, \gamma)}{\text{SNR}_t} \cdot \|\epsilon - \epsilon_\theta(z_t, t, c)\|_2^2, \tag{11}$$

与 $\gamma = 5.0$。

该公式保留了原有的噪声预测目标，同时根据信噪比调节每个时间步的贡献。由噪声主导的时间步（ $\bar{\alpha}_t$ 较小）会被降低权重，而信号内容较强的时间步则会受到相对更高的重视，从而实现更稳定的优化并提升样本质量。

## 5 实验

本节介绍用于训练和评估 A*rt2*M*us* 的实验方案。我们首先对 A*rtSound* 数据集进行探索性分析，随后详细说明训练设置、性能指标和模型配置。之后报告客观和主观评估结果。

### 5.1 探索性数据分析

我们对 *ArtSound* 进行了初步分析，以评估艺术作品—音乐相似度的分布，并探讨艺术风格与音乐流派之间可能存在的关联。我们使用 ImageBind 嵌入构建了艺术作品—音乐配对（第 3.3 节），为每首音乐曲目选取与之最相似的艺术作品。

相似度值是通过计算使用 ImageBind 提取的艺术作品（图像）与音乐（音频）嵌入之间的余弦相似度得出的，范围为 −0.0216 至 0.5518，均值为 0.1864 （表 5）。尽管相似度总体不高——这与任务的跨模态特性一致——但超过 50,000组配对高于均值，表明相似度分数呈平滑连续的分布，而非低值与高值之间存在明显分界。对少量样本（30 组）进行人工检查后发现，许多配对直观上具有意义，但也有一些仍不明确，这反映了跨模态对齐本身具有主观性。

接下来，我们通过考察 32 种艺术风格在 161 种音乐流派中的分布，分析艺术作品与音乐之间的关联（图 5）。我们发现了几种总体趋势：（*i*）实验音乐、电子音乐和先锋派等流派往往与现代且以感知为导向的运动（如印象主义、表现主义）共同出现；（*ii*）代表性较低的

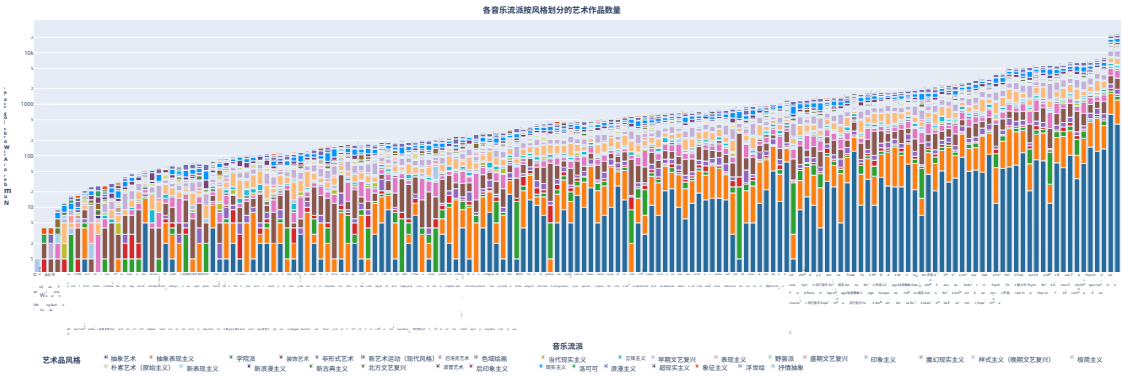



图 5. 按音乐流派划分的艺术作品分布，并按艺术风格细分。**抽象艺术**、**抽象表现主义**、**新艺术运动（现代风格）**和**表现主义**在大多数流派中都十分常见。相比之下，**早期文艺复兴**和**象征主义**出现频率较低。

流派（例如深度放克、西部摇摆）的风格分布更为多样；而（*iii*）流行流派（嘻哈、爵士乐、布鲁斯）往往与强调情感或叙事表现力的风格相契合。这些趋势并非旨在确立明确的审美对应关系，而是表明基于嵌入的配对捕捉到了微弱但可解释的跨模态规律。

### 5.2 实验设置

A*rtSound* 数据集按艺术风格分层，以 90/10 的比例划分为 $D_{train}$ 和 $D_{test}$。从 $D_{test}$中抽取 100 件艺术作品，组成验证集 $D_{val}$。验证期间，模型使用 200 个推理步骤，并搭配负向提示词“低质量”生成 10 秒样本。

训练共进行了 50 个周期，采用默认超参数的 AdamW，学习率为 $2\times10^{-5}$，包含 1,324 个预热步骤、37,200 个总步骤、批量大小为 128（在 4×块 A100 GPU 上，每块 GPU 处理 32 个样本），并进行 1 步梯度累积。仅更新图像到 LoA 接口；AudioLDM 2 的所有组件均保持冻结。完整的 50 周期训练约耗时 60 小时。

### 5.3 性能指标

描述质量通过 $\mathcal{ICS}core$ 和 $\mathcal{ACS}core$ (第3节）进行评估。生成的音频则通过三项互补指标进行评估。

KL散度（KL 散度）用于衡量生成音乐的特征分布与真实音频的特征分布有多接近。给定分布 P（真实数据）和 Q（生成数据），数值越低表示对齐效果越好：

$$\text{KL-Div}(\mathcal{P}\|\mathcal{Q}) = \sum_i \mathcal{P}(i)\log\frac{\mathcal{P}(i)}{\mathcal{Q}(i)}. \tag{12}$$

Fréchet音频距离（FAD） [31] 通过比较生成音频和参考音频的 VGGish 嵌入（建模为高斯分布）来评估感知真实感：

$$\text{FAD} = \|\mu_b - \mu_e\|^2 + \text{Tr}\left(\Sigma_b + \Sigma_e - 2(\Sigma_b\Sigma_e)^{1/2}\right), \tag{13}$$

其中，$(\mu_b,\ \Sigma_b)$ 和 $(\mu_e,\ \Sigma_e)$ 分别是参考嵌入和生成的嵌入的均值和协方差。分数越低，表示与真实音频的感知相似度越高。

ImageBind 分数（IBSc）通过计算艺术作品与生成音频的 ImageBind 嵌入之间的余弦相似度来衡量跨模态一致性：

$$\mathrm{IBSc}(e_{\mathrm{art}}, e_{\mathrm{gen}}) = \frac{e_{\mathrm{art}} \cdot e_{\mathrm{gen}}}{\|e_{\mathrm{art}}\|\,\|e_{\mathrm{gen}}\|}. \tag{14}$$

我们分别报告 IBSc 相对于艺术作品和配对真实值音轨的数值；数值越高，表示视觉—音频语义对齐越强。

这些指标共同衡量分布相似度（KL 散度）、感知真实感（FAD）和跨模态连贯性（IBSc）。

### 5.4 实验结果

表 6 报告了针对不同视觉编码器和词元数量在训练过程中计算的客观指标。对于 ImageBind，IBSc 随着词元数量增加而持续提升，并在每张图像 16 个词元时达到峰值。CLIP 则未呈现类似的单调趋势。不同配置下，FAD 和 KL 散度均保持稳定。这些结果表明，ImageBind 更高维的多模态嵌入空间受益于更大的词元容量，从而能够提供更丰富的条件信号。

随后，我们将表现最佳的 A*rt2*M*us* 配置（采用 16 个词元的 ImageBind 和采用 6 个词元的 CLIP）与 AudioLDM 2 [40], AudioLDM 2-Music [40], 以及 Mozart's Touch [35]进行了比较。AudioLDM 2 模型以文本为输入，其中 AudioLDM 2 可生成音乐、语音或音效，而 AudioLDM 2-Music 专注于音乐。Mozart's Touch 接收图像并生成描述文本，再将其转换为文本，最终用于引导音乐生成。相比之下，A*rt2*M*us* 无需文本，也不会将图像转换为描述，而是直接利用图像信息引导音乐生成。对于基于文本的基线方法，我们将 A*rtSound* 艺术作品的描述文本作为输入；对于 Mozart's Touch，则先按照其默认提示词，使用 Qwen3-14B 将相同的描述文本转换为简短的音乐描述，以符合该模型预期的输入格式。

表 6. 不同视觉编码器和词元数量的平均训练指标。箭头表示较高 (↑) 或较低 (↓) 的数值更理想。各编码器的最佳结果以**粗体**标示；次佳结果加下划线。缩写：Tok/Img – 每张图像的词元数；艺术作品 – 艺术作品；生成音乐 – 生成的音乐；真实音乐 – 真实值音乐。

| 图像编码器 | 每张图像的词元数 | FAD ↓ | KL 散度 ↓ | $\mathrm{IBSc}_{\mathrm{Artw–GeMus}}$ ↑ | $\mathrm{IBSc}_{\mathrm{GtMus–GeMus}}$ ↑ |
|---|---|---|---|---|---|
| ImageBind | 4 | **19** | **0.00266** | 0.05716 | 0.19363 |
| | 6 | **19** | 0.00273 | 0.05693 | 0.21398 |
| | 8 | **19** | 0.00284 | 0.05946 | 0.21341 |
| | 10 | **19** | 0.00386 | 0.05288 | 0.18045 |
| | 16 | **19** | 0.00279 | **0.05986** | **0.21661** |
| CLIP | 4 | 22 | 0.00320 | 0.05185 | 0.18021 |
| | 6 | 23 | 0.00317 | **0.05663** | 0.16921 |
| | 8 | 22 | 0.00310 | 0.05130 | 0.17757 |
| | 10 | **21** | **0.00307** | 0.05378 | **0.19169** |
| | 16 | 22 | 0.00309 | 0.05546 | 0.17374 |

表 7。A*rt2*M*us* 与基线模型的客观评估结果：AudioLDM2（音效、人类语音和音乐）、AudioLDM 2- Music（仅音乐）以及 Mozart's Touch（仅音乐）。箭头表示较高（↑）或较低（↓）的数值更理想。每个行块中的最佳数值以**粗体**标出；次佳数值带下划线。缩写：嵌入模型 – 图像嵌入模型；每张图像的词元数 – 每张图像的词元数；艺术作品 – 艺术作品；生成音乐 – 生成的音乐；真实音乐 – 真实值音乐。

| 模型 | 嵌入模型 | 训练规模（小时） | 每张图像的词元数 | FAD ↓ | KL 散度 ↓ | $\text{IBSc}_{\text{Artw–GeMus}}$ ↑ | $\text{IBSc}_{\text{GtMus–GeMus}}$ ↑ |
|---|---|---|---|---|---|---|---|
| *Art2Mus* | ImageBind | ≈2.6 | 16 | **13** | 0.00296 | 0.04671 | 0.20368 |
| 艺术2音乐 | CLIP | | 6 | 15 | 0.00374 | 0.04173 | 0.16716 |
| AudioLDM 2 | – | ≈1,150,000 | – | 14 | 0.00425 | 0.09822 | 0.17412 |
| AudioLDM 2-Music | – | ≈665,000 | – | 15 | 0.08449 | **0.18683** | 0.18683 |
| Mozart's Touch | – | ≈20,000 | – | 19 | **0.00238** | 0.08348 | **0.21629** |

基于 ImageBind 的 A*rt2*M*us* 取得了最低的 FAD（13）、具有竞争力的 KL 散度（0.00296），以及与真实音乐相比第二高的 IBSc（0.20368）。相比之下，基于 CLIP 的变体在所有指标上的表现都略差，证实 ImageBind嵌入能够更好地捕捉艺术作品与音乐之间的跨模态关系。AudioLDM 2 在海量数据集（≈1, 150,000 小时）上训练，取得了较低的 FAD（14）和具有竞争力的艺术作品 IBSc（0.09822），但其 KL 散度高于 A*rt2*M*us*。AudioLDM 2-Music 尽管训练时长较短（≈665，000 小时），却取得了最高的艺术作品–生成音乐 IBSc（0.18683），表明其与艺术作品高度对齐，但其 KL 散度明显较差（0.08449）。Mozart's Touch 在较小的数据集（≈20，000 小时）上训练，取得了最低的 KL 散度（0.00238）和最高的真实音乐–生成音乐 IBSc（0.21629），凸显了其捕捉真实音乐特征的有效性。不过，它的 FAD 最高（19）。

需要注意的是，基于文本的基线模型受益于从图像字幕和语言模型中获得的显式语义条件。从这个意义上说，与其将它们视为纯视觉条件系统的直接竞争者，不如将其视为语义监督下对齐效果的上限。总体而言，在难度严格更高的设定下，A*rt2*M*us* 在感知质量与跨模态连贯性之间实现了良好的权衡，因为它直接根据艺术作品生成音乐，不依赖从文本中提取的语义监督。

我们对 *Art2Mus* 生成的音乐进行了主观评估，以分析其定性特征。评估共有十五名参与者（3名音乐家、2名视觉艺术家以及10名没有艺术背景的参与者）参加。每位参与者通过回答表8中的四个问题，对15件艺术作品进行评估，为音频质量、表现力、相关性和对齐情况打分，并为每件艺术作品选出偏好曲目。结果汇总于表9。

两种 A*rt2*M*us* 配置都表现出音频质量与语义对齐之间一贯的权衡。基于 ImageBind 的生成结果在音频质量上略胜一筹，这表明其音色特征更加清晰或稳定。同时，基于 CLIP 的输出在表现力、相关性和对齐方面得分更高，说明其能够更充分地捕捉艺术作品的语义和情感内容。总体而言，A*rt2*M*us* 在将生成的音乐与视觉艺术作品进行对齐方面展现出具有竞争力的性能，同时保持了合理的音频质量。

表8. 主观评估所用的问题。参与者对每首生成曲目的音频质量、表现力、相关性和对齐情况进行评分。

| ID | 问题 | 评分标准 |
|---|---|---|
| Q1 | 这段音乐听起来如何? | 1–5 |
| Q2 | 这段音乐的表现力和情感感染力如何? | 1–5 |
| Q3 | 这段音乐与艺术作品的相关性如何? | 1–5 |
| Q4 | 这段音乐对艺术作品的情绪／内容反映得如何? | 1–5 |

表9. 主观评估结果（15 位参与者的平均分）。每条音轨均从音频质量、表现力、与艺术作品的相关性以及与艺术作品的对齐程度进行评估。对于每个模型，我们计算每位参与者的平均评分，并对所有参与者的评分取平均值，得出每个问题的最终分数。最佳数值以**粗体**显示。

| 模型 | 音频质量（Q1）↑ | 表现力（Q2）↑ | 相关性 (Q3) ↑ | 对齐 (Q4) ↑ |
|---|---|---|---|---|
| A*rt2*M*us*（仅 CLIP 艺术作品） | 2.72 | **3.10** | **2.91** | **2.88** |
| A*rt2*M*us*（仅 ImageBind 艺术作品） | **2.74** | 3.02 | 2.87 | 2.79 |

### 5.5 讨论

A*rt2*M*us* 无需依赖文本条件即可直接进行艺术品到音乐生成，这构成了一个严格意义上比基于描述文本的基线更具挑战性的设置。因此，与文本条件系统相比，较低的绝对得分并不表示模型能力较弱，而是反映了在缺乏显式语义监督的情况下学习视觉—音频对应关系的更高难度。尽管如此，A*rt2*M*us* 仍实现了具有竞争力的感知质量和跨模态对齐效果；鉴于其缺少源自语言的语义线索，这一结果尤为突出。

重要的是，文本条件管线本质上会将视觉信息压缩为语言抽象。因此，在基于文本的系统中，两幅图像字幕相似但视觉结构截然不同的艺术作品可能会坍缩为相似的音乐输出，因为难以用语言表达的细粒度构图与风格线索会在图像描述过程中被丢弃。通过直接在视觉嵌入空间中运行，A*rt2*M*us* 保留了原本会因文本中介而丢失的结构信息。

总体而言，A*rt2*M*us* 证明了大规模直接生成艺术作品到音乐是可行的，并为未来的多模态创意系统奠定了基础。

## 6 结论

本文介绍了 A*rt2*M*us*，这是一个用于直接生成艺术作品到音乐的框架，它摆脱了对文本中间表示的依赖，转而直接从图像嵌入中学习视觉—音频对应关系。受潜在扩散模型生成范式的启发，A*rt2*M*us* 通过视觉条件提取器 和图像对齐器重新设计条件接口，使预训练音频扩散骨干网络能够在纯视觉条件下运行。除模型之外，我们还介绍了 A*rtSound*，这是一个包含 10.5 万组艺术作品—音乐配对的大型版数据集，并附有图像和音频描述；同时，我们提出了两种对齐度量指标 ICS*core* 和 ACS*core*，用于评估描述可靠性。A*rtSound* 不仅提供训练数据，也为大规模研究跨模态创意生成建立了基准。

实验结果表明，直接进行视觉到音乐生成是可行的，并且能够实现有竞争力的感知质量和有意义的跨模态对齐。文本条件系统受益于明确的语义监督，因此代表了语言介导条件控制下的上限；而 A*rt2*M*us* 则直面一个难度更高的问题：直接从视觉结构中学习对应关系。这一区别至关重要：将艺术作品压缩为描述文本不可避免地会丢失细粒度的构图和风格信息，而直接视觉条件控制则能在生成过程中保留这些线索。

更广泛而言，本研究确立了直接从艺术作品生成音乐，作为多模态生成式人工智能领域一个切实可行的研究方向。未来的工作包括扩大数据集以提升风格覆盖范围，探索更丰富的视觉标记化和自适应 LoA 表示，整合检索增强或分层条件控制机制，并采用可解释人工智能技术 [9] 以更好地理解视觉特征如何影响音乐结构。

通过视觉控制或视觉-文本混合控制实现用户引导编辑，也为交互式个性化创意系统的发展提供了一个有前景的方向。

**致谢**

我们感谢 ISCRA 授予我们使用 LEONARDO 超级计算机的权限。该计算机归欧洲高性能计算共同事业所有，由 CINECA（意大利）托管。

Ivan Rinaldi 的研究得到博士奖学金资助，该奖学金属于“企业高等教育地域协议”项目，由巴里阿尔多·莫罗大学“图像、文本与音频转换高级生成模型”博士项目提供，并由 AI Factory有限责任公司共同资助（CUP F61B23000370006）。

Nicola Fanelli 的研究得到意大利“第118/23号部长令”项下的博士奖学金资助，该奖学金属于国家复苏与韧性计划、任务4、组成部分2、投资4.1框架下的博士项目“利用人工智能技术分析和 valorización 数字化艺术遗产”（CUP H91I23000690007）。

**参考文献**

[1] Panos Achlioptas、Maks Ovsjanikov、Kilichbek Haydarov、Mohamed Elhoseiny 和 Leonidas J Guibas。2021。ArtEmis：用于视觉艺术的情感语言。载于*IEEE/CVF*计算机视觉与模式识别会议论文集。计算机视觉基金会/IEEE，11569–11579。[2] Andrea Agostinelli、Timo I Denk、Zalán Borsos、Jesse Engel、Mauro Verzetti、Antoine Caillon、Qingqing Huang、Aren Jansen、Adam Roberts、Marco Tagliasacchi 等。2023。MusicLM：从文本生成音乐。*arXiv*预印本 *arXiv:2301.11325* abs/2301.11325（2023）。[3] Jisheng Bai、Haohe Liu、Mou Wang、Dongyuan Shi、Wenwu Wang、Mark D Plumbley、Woon-Seng Gan 和 Jianfeng Chen。2025。AudioSetCaps：利用大型音频和语言模型及自动生成流程构建的增强型音频描述数据集。*IEEE*音频、语音与语言处理汇刊（2025）。[4] Shuai Bai、Keqin Chen、Xuejing Liu、Jialin Wang、Wenbin Ge、Sibo Song、Kai Dang、Peng Wang、Shijie Wang、Jun Tang 等。2025。Qwen2.5-VL技术报告。*arXiv*预印本 *arXiv:2502.13923* abs/2502.13923（2025）。[5] Zechen Bai、Yuta Nakashima 和 Noa Garcia。2021。解释一下这幅画：多主题知识型艺术描述生成。载于*IEEE/CVF*国际计算机视觉会议论文集。IEEE，5422–5432。[6] Zalán Borsos、Raphaël Marinier、Damien Vincent、Eugene Kharitonov、Olivier Pietquin、Matt Sharifi、Dominik Roblek、Olivier Teboul、DavidGrangier、Marco Tagliasacchi 等。2023。AudioLM：一种音频生成的语言建模方法。*IEEE/ACM* 音频、语音与语言处理汇刊31（2023），2523–2533。[7] Giovanna Castellano、Vincenzo Digeno、Giovanni Sansaro 和 Gennaro Vessio。2022。利用知识图谱和深度学习实现自动艺术分析。基于知识的系统248（2022），108859。[8] Giovanna Castellano、Nicola Fanelli、Raffaele Scaringi 和 Gennaro Vessio。2023。探索视觉-语言预训练与ChatGPT在艺术品描述生成中的协同作用：初步研究。载于国际图像分析与处理会议。Springer，309–321。[9] Giovanna Castellano、MARIA GRAZIA Miccoli、Raffaele Scaringi、Gennaro Vessio、Gianluca Zaza 等。2024。利用大语言模型通过Grad-CAM热力图解释人工智能生成艺术分类。载于第五届意大利可解释人工智能研讨会论文集（与意大利人工智能协会第*23*届国际会议同期举行），意大利博尔扎诺。CEUR-WS.org，65–74。[10] Eva Cetinic。2021。艺术作品的图像志图像描述生成。载于国际模式识别会议。Springer，502–516。[11] Xinlei Chen、Hao Fang、Tsung-Yi Lin、Ramakrishna Vedantam、Saurabh Gupta、Piotr Dollár 和 C Lawrence Zitnick。2015。Microsoft COCO描述数据集：数据收集与评估服务器。*arXiv*预印本 *arXiv:1504.00325* abs/1504.00325（2015）。[12] Sanjoy Chowdhury、Sayan Nag、KJ Joseph、Balaji Vasan Srinivasan 和 Dinesh Manocha。2024。MelFusion：利用扩散模型根据图像和语言线索合成音乐。载于*IEEE/CVF*计算机视觉与模式识别会议论文集。IEEE，26826–26835。[13] Yunfei Chu、Jin Xu、Qian Yang、Haojie Wei、Xipin Wei、Zhifang Guo、Yichong Leng、Yuanjun Lv、Jinzheng He、Junyang Lin 等。2024。Qwen2-Audio技术报告。*arXiv*预印本 *arXiv:2407.10759* abs/2407.10759（2024）。[14] Hyung Won Chung、Le Hou、Shayne Longpre、Barret Zoph、Yi Tay、William Fedus、Yunxuan Li、Xuezhi Wang、Mostafa Dehghani、Siddhartha Brahma 等。2024。扩展指令微调语言模型。机器学习研究杂志25（2024），70:1–70:53。[15] Jade Copet、Felix Kreuk、Itai Gat、Tal Remez、David Kant、Gabriel Synnaeve、Yossi Adi 和 Alexandre Défossez。2023。简单且可控的音乐生成。载于神经信息处理系统进展第*36*卷：*2023*年神经信息处理系统年会，*NeurIPS 2023*，美国路易斯安那州新奥尔良，*2023*年*12*月*10*日至*16*日，第36卷。47704–47720。[16] Michaël Defferrard、Kirell Benzi、Pierre Vandergheynst 和 Xavier Bresson。2017。FMA：音乐分析数据集。载于第*18*届国际音乐信息检索学会会议论文集，*ISMIR 2017*，中国苏州，*2017*年*10*月*23*日至*27*日。316–323。[17] Zihao Deng、Yinghao Ma、Yudong Liu、Rongchen Guo、Ge Zhang、Wenhu Chen、Wenhao Huang 和 Emmanouil Benetos。2024。Musilingo：利用预训练语言模型连接音乐与文本，以实现音乐描述生成和查询响应。载于计算语言学协会会议论文集：计算提交给ACM的稿件

语言学：NAACL 2024。计算语言学协会，3643–3655。[18] Soham Deshmukh、Benjamin Elizalde、Rita Singh 和 Huaming Wang。2023。Pengi：用于音频任务的音频语言模型。载于神经信息处理系统进展第36卷：2023年神经信息处理系统年会，NeurIPS 2023，美国路易斯安那州新奥尔良，2023年12月10日至16日，第36卷。18090–18108。[19] SeungHeon Doh、Keunwoo Choi、Jongpil Lee 和 Juhan Nam。2023。LP-MusicCaps：基于大语言模型的伪音乐描述。载于第24届国际音乐信息检索学会会议论文集，ISMIR 2023，意大利米兰，2023年11月5日至9日。409–416。[20] Benjamin Elizalde、Soham Deshmukh、Mahmoud Al Ismail 和 Huaming Wang。2023。CLAP：通过自然语言监督学习音频概念。载于ICASSP 2023——2023年IEEE国际声学、语音与信号处理会议（ICASSP）。IEEE，1–5。[21] Songwei Ge、Taesung Park、Jun-Yan Zhu 和 Jia-Bin Huang。2023。利用丰富文本实现富有表现力的文本到图像生成。载于IEEE/CVF国际计算机视觉会议论文集。IEEE，7545–7556。[22] Jort F Gemmeke、Daniel PW Ellis、Dylan Freedman、Aren Jansen、Wade Lawrence、R Channing Moore、Manoj Plakal 和 Marvin Ritter。2017。Audio set：音频事件本体与人工标注数据集。载于2017年IEEE国际声学、语音与信号处理会议（ICASSP）。IEEE，776–780。[23] Rohit Girdhar、Alaaeldin El-Nouby、Zhuang Liu、Mannat Singh、Kalyan Vasudev Alwala、Armand Joulin 和 Ishan Misra。2023。ImageBind：用一个嵌入空间绑定所有模态。载于IEEE/CVF计算机视觉与模式识别会议论文集。IEEE，15180–15190。[24] Aaron Grattafiori、Abhimanyu Dubey、Abhinav J auhri、Abhinav Pandey、Abhishek Kadian、Ahmad Al-Dahle、Aiesha Letman、Akhil Mathur、Alan Schelten、Alex Vaughan 等。2024。Llama 3模型家族。arXiv预印本 arXiv:2407.21783 abs/2407.21783（2024）。[25] Tiankai Hang、Shuyang Gu、Chen Li、Jianmin Bao、Dong Chen、Han Hu、Xin Geng 和 Baining Guo。2023。通过最小信噪比加权策略实现高效扩散训练。载于IEEE/CVF国际计算机视觉会议，ICCV 2023，法国巴黎，2023年10月1日至6日。IEEE，7407–7417。[26] Jack Hessel、Ari Holtzman、Maxwell Forbes、Ronan Le Bras 和 Yejin Choi。2021。CLIPScore：一种无参考图像描述生成评估指标。载于2021年自然语言处理实证方法会议论文集，EMNLP 2021，线上活动/多米尼加共和国蓬塔卡纳，2021年11月7日至11日。计算语言学协会，7514–7528。[27] Tanisha Hisariya、Huan Zhang 和 Jinhua Liang。2024。连接绘画与音乐——探索通过绘画实现基于情感的音乐生成。arXiv预印本 arXiv:2409.07827（2024）。[28] Zhejing Hu、Xiao Ma、Yan Liu、Gong Chen、Yongxu Liu 和 Roger B Dannenberg。2023。重复之美：一种结合动机级重复生成器和大纲到音乐生成器的符号音乐算法作曲模型。IEEE多媒体汇刊26（2023），4320–4333。[29] Po-Yao Huang、Hu Xu、Juncheng Li、Alexei Baevski、Michael Auli、Wojciech Galuba、Florian Metze 和 Christoph Feichtenhofer。2022。会聆听的掩码自编码器。神经信息处理系统进展35（2022），28708–28720。[30] Rongjie Huang、Jiawei Huang、Dongchao Yang、Yi Ren、Luping Liu、Mingze Li、Zhenhui Ye、Jinglin Liu、Xiang Yin 和 Zhou Zhao。2023。Make-an-audio：利用提示增强扩散模型实现文本到音频生成。载于国际机器学习会议，ICML 2023，2023年7月23日至29日，美国夏威夷州檀香山（机器学习研究会议论文集，第202卷）。PMLR，13916–13932。[31] Kevin Kilgour、Mauricio Zuluaga、Dominik Roblek 和 Matthew Sharifi。2019。Fr\' echet音频距离：一种评估音乐增强算法的指标。载于国际语音通信协会第20届年会，Interspeech 2019，奥地利格拉茨，2019年9月15日至19日。ISCA，2350–2354。[32] Chris Dongjoo Kim、Byeongchang Kim、Hyunmin Lee 和 Gunhee Kim。2019。AudioCaps：为真实场景中的音频生成描述。载于2019年计算语言学协会北美分会会议：人类语言技术论文集，NAACL-HLT 2019，美国明尼苏达州明尼阿波利斯，2019年6月2日至7日，第1卷（长篇与短篇论文）。计算语言学协会，119–132。[33] Jungil Kong、Jaehyeon Kim 和 Jaekyoung Bae。2020。HiFi-GAN：用于高效、高保真语音合成的生成对抗网络。载于神经信息处理系统进展第33卷：2020年神经信息处理系统年会，NeurIPS 2020，2020年12月6日至12日，线上，第33卷。17022–17033。[34] Dongxu Li、Junnan Li、Hung Le、Guangsen Wang、Silvio Savarese 和 Steven CH Hoi。2023。LAVIS：一站式语言视觉智能库。载于计算语言学协会第61届年会论文集：系统演示，ACL 2023，加拿大多伦多，2023年7月10日至12日。计算语言学协会，31–41。[35] Jiajun Li、Tianze Xu、Xuesong Chen、Xinrui Yao、Jingchou Han 和 Shuchang Liu。2025。Mozart's Touch：一种基于预训练大模型的轻量级多模态音乐生成框架。载于人工智能生成内容国际会议（AIGC 2024），第13649卷。SPIE，198–207。[36] Callie C Liao、Duoduo Liao 和 Ellie L Zhang。2025。一种由算法驱动的音乐核心支持的图像到音乐生成框架。（2025）。[37] Chin-Yew Lin。2004。ROUGE：用于自动评估摘要的软件包。载于文本摘要的拓展。计算语言学协会，西班牙巴塞罗那，74–81。[38] Cheng Liu、Hui Wang、Jinghua Zhao、Shiwan Zhao、Hui Bu、Xin Xu、Jiaming Zhou、Haoqin Sun 和 Yong Qin。2025。MusicEval：包含专家评分、用于自动文本到音乐评估的生成式音乐数据集。载于2025年IEEE国际声学、语音与信号处理会议，ICASSP 2025，印度海得拉巴，2025年4月6日至11日。IEEE，1–5。[39] Haotian Liu、Chunyuan Li、Qingyang Wu 和 Yong Jae Lee。2023。视觉指令微调。载于神经信息处理系统进展第36卷：2023年神经信息处理系统年会，NeurIPS 2023，美国路易斯安那州新奥尔良，2023年12月10日至16日，第36卷。34892–34916。

[40] Haohe Liu、Yi Yuan、Xubo Liu、Xinhao Mei、Qiuqiang Kong、Qiao Tian、Yuping Wang、Wenwu Wang、Yuxuan Wang 和 Mark D. Plumbley。2024。AudioLDM 2：通过自监督预训练学习整体音频生成。《IEEE/ACM 音频、语音与语言处理汇刊》32（2024），2871–2883。[41] Shansong Liu、Atin Sakkeer Hussain、Chenshuo Sun 和 Ying Shan。2023。M$^2$M-Gen：借助大语言模型的强大能力实现多模态音乐理解与生成。arXiv 预印本 arXiv:2311.11255（2023）。[42] Shansong Liu、Atin Sakkeer Hussain、Qilong Wu、Chenshuo Sun 和 Ying Shan。2026。Mumu-llama：通过大语言模型实现多模态音乐理解与生成。《专家系统与应用》305（2026），130688。[43] Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Mandar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettlemoyer 和 Veselin Stoyanov。2019。RoBERTa：一种稳健优化的 BERT 预训练方法。arXiv 预印本 arXiv:1907.11692（2019）。[44] Yue Lu、Chao Guo、Xingyuan Dai 和 Fei-Yue Wang。2024。ArtCap：用于艺术绘画图像描述生成的数据集。《IEEE 计算社会系统汇刊》11，1（2024），576–587。[45] Xichu Ma、Yuchen Wang 和 Ye Wang。2024。基于图学习的偏好建模和文本查询的符号音乐生成。《IEEE多媒体汇刊》26（2024），10545–10558。[46] Ilaria Manco、Emmanouil Benetos、Elio Quinton 和 György Fazekas。2021。MusCaps：为音乐音频生成描述文本。载于《2021年国际联合神经网络会议》（IJCNN）。IEEE，1–8。[47] Xinhao Mei、Xubo Liu、Qiushi Huang、Mark D. Plumbley 和 Wenwu Wang。2021。音频描述Transformer。载于《第六届声学场景与事件检测及分类研讨会论文集》（DCASE2021）。西班牙巴塞罗那，211–215。[48] Jan Melechovsky、Zixun Guo、Deepanway Ghosal、Navonil Majumder、Dorien Herremans 和 Soujanya Poria。2024。Mustango：迈向可控的文本到音乐生成。载于《2024年计算语言学协会北美分会年会：人类语言技术论文集》（第1卷：长篇论文），NAACL 2024，墨西哥墨西哥城，2024年6月16日至21日。计算语言学协会，8293–8316。[49] Margaret Mitchell、Jesse Dodge、Amit Goyal、Kota Yamaguchi、Karl Stratos、Xufeng Han、Alyssa Mensch、Alexander Berg、Tamara Berg 和 Hal Daumé III。2012。Midge：根据计算机视觉检测结果生成图像描述。载于EACL 2012，即第13届计算语言学协会欧洲分会会议，法国阿维尼翁，2012年4月23日至27日。计算语言学协会，747–756。[50] Bryan A Plummer、Liwei Wang、Chris M Cervantes、Juan C Caicedo、Julia Hockenmaier 和 Svetlana Lazebnik。2015。Flickr30k 实体：收集区域到短语对应关系，以构建更丰富的图像到句子模型。载于《2015年IEEE国际计算机视觉会议》，ICCV 2015，智利圣地亚哥，2015年12月7日至13日。IEEE计算机学会，2641–2649。[51] Alec Radford、Jong Wook Kim、Chris Hallacy、Aditya Ramesh、Gabriel Goh、Sandhini Agarwal、Girish Sastry、Amanda Askell、Pamela Mishkin、Jack Clark 等。2021。通过自然语言监督学习可迁移视觉模型。载于《第38届国际机器学习会议论文集》，ICML 2021，2021年7月18日至24日，线上会议（《机器学习研究论文集》，第139卷）。PMLR，8748–8763。[52] Alec Radford、Jeffrey Wu、Rewon Child、David Luan、Dario Amodei、Ilya Sutskever 等。2019。语言模型是无监督的多任务学习器。《OpenAI 博客》1，8（2019），9。[53] Ivan Rinaldi、Nicola Fanelli、Giovanna Castellano 和 Gennaro Vessio。2024。Art2Mus：通过跨模态生成连接视觉艺术与音乐。载于《计算机视觉——ECCV 2024研讨会：论文集，第V部分》，意大利米兰，2024年9月29日至10月4日（《计算机科学讲义》，第15627卷）。Springer，173–186。[54] Robin Rombach、Andreas Blattmann、Dominik Lorenz、Patrick Esser 和 Björn Ommer。2022。利用潜在扩散模型进行高分辨率图像合成。载于IEEE/CVF计算机视觉与模式识别会议，CVPR 2022，美国路易斯安那州新奥尔良，2022年6月18日至24日。IEEE，10674–10685。[55] Sara Sarto、Manuele Barraco、Marcella Cornia、Lorenzo Baraldi 和 Rita Cucchiara。2023。用于图像与视频描述生成评估的正向增强对比学习。载于IEEE/CVF计算机视觉与模式识别会议，CVPR 2023，加拿大不列颠哥伦比亚省温哥华，2023年6月17日至24日。IEEE，6914–6924。[56] Megha Sharma、Muhammad Taimoor Haseeb、Gus Xia 和 Yoshimasa Tsuruoka。2024。M2M-Gen：利用大语言模型为日本漫画自动生成背景音乐的多模态框架。arXiv 预印本 arXiv:2410.09928（2024）。[57] Roy Sheffer 和 Yossi Adi。2023。我听见了你真实的色彩：图像引导的音频生成。载于ICASSP 2023——2023年IEEE国际声学、语音与信号处理会议（ICASSP）。IEEE，1–5。[58] Yi-Jen Shih、Shih-Lun Wu、Frank Zalkow、Meinard Müller 和 Yi-Hsuan Yang。2022。主题变换器：使用主题条件变换器生成符号音乐。《IEEE多媒体汇刊》25（2022），3495–3508。[59] Jiaming Song、Chenlin Meng 和 Stefano Ermon。2021。去噪扩散隐式模型。载于《国际学习表征会议》。https://openreview.net/forum?id=St1giarCHLP[60] Jiahao Song 和 Yu-Zhao Wang。2025。MusFlow：通过条件流匹配实现多模态音乐生成。载于《第33届ACM国际多媒体会议论文集》，MM 2025，爱尔兰都柏林，2025年10月27日至31日。ACM，10200–10209。[61] Krishna Srinivasan、Karthik Raman、Jiecao Chen、Michael Bendersky 和 Marc Najork。2021。Wit：用于多模态多语言机器学习的基于维基百科的图像文本数据集。载于SIGIR '21：第44届国际ACM SIGIR信息检索研究与发展会议，线上会议，加拿大，2021年7月11日至15日。ACM，2443–2449。[62] Xiaodong Tan、Mathis Antony 和 H Kong。2020。通过情感为视觉艺术自动生成音乐。载于第十一届计算创造力国际会议论文集，ICCC 2020，葡萄牙科英布拉，2020年9月7日至11日。计算创造力协会（ACC），247–250。稿件已提交至 ACM

[63] Zhaokai Wang、Chenxi Bao、Le Zhuo、Jingrui Han、Yang Yue、Yihong Tang、Victor Shea-Jay Huang 和 Yue Liao。2025。从视觉生成音乐：综述。arXiv：2503.21254 [cs.CV] https://arxiv.org/abs/2503.21254[64] Zijie J Wang、Evan Montoya、David Munechika、Haoyang Yang、Benjamin Hoover 和 Duen Horng Chau。2023。Diffusiondb：面向文本到图像生成模型的大规模提示词图库数据集。载于计算语言学协会第*61*届年会论文集（第*1*卷：长篇论文），*ACL 2023*，加拿大多伦多，*2023*年*7*月*9*日至*14*日。计算语言学协会，893–911。[65] Benjamin Warner、Antoine Chaffin、Benjamin Clavié、Orion Weller、Oskar Hallström、Said Taghadouini、Alexis Gallagher、Raja Biswas、Faisal Ladhak、Tom Aarsen 等。2025。更智能、更出色、更快速、更长效：面向快速、内存高效、长上下文微调与推理的现代双向编码器。载于计算语言学协会第*63*届年会论文集（第*1*卷：长篇论文），*ACL 2025*，奥地利维也纳，*2025*年*7*月*27*日至*8*月*1*日。计算语言学协会，2526–2547。[66] Jin Xu、Zhifang Guo、Jinzheng He、Hangrui Hu、Ting He、Shuai Bai、Keqin Chen、Jialin Wang、Yang Fan、Kai Dang 等。2025。Qwen2.5-Omni 技术报告。*arXiv* 预印本 *arXiv:2503.20215* abs/2503.20215（2025）。[67] An Yang、Anfeng Li、Baosong Yang、Beichen Zhang、Binyuan Hui、Bo Zheng、Bowen Yu、Chang Gao、Chengen Huang、Chenxu Lv 等。2025。Qwen3 技术报告。*arXiv* 预印本 *arXiv:2505.09388* abs/2505.09388（2025）。[68] Tianyi Zhang、Varsha Kishore、Felix Wu、Kilian Q Weinberger 和 Yoav Artzi。2020。BERTScore：使用 BERT 评估文本生成。载于第八届国际学习表征会议，*ICLR 2020*，埃塞俄比亚亚的斯亚贝巴，*2020*年*4*月*26*日至*30*日。OpenReview.net。[69] Sicheng Zhao、Yaxian Li、Xingxu Yao、Weizhi Nie、Pengfei Xu、Jufeng Yang 和 Kurt Keutzer。2020。效价-唤醒度空间中图像与音乐之间基于情感的端到端匹配。载于*MM '20*：第*28*届*ACM*国际多媒体会议，线上会议／美国华盛顿州西雅图，*2020*年*10*月*12*日至*16*日。ACM，2945–2954。[70] Yuanzhi Zhu、Zhaohai Li、Tianwei Wang、Mengchao He 和 Cong Yao。2023。利用扩散模型进行条件文本图像生成。载于*IEEE/CVF*计算机视觉与模式识别会议论文集。IEEE，14235–14245。

## 图像描述

如正文（第 3.2 节）所述，我们评估了用于图像描述和音频描述的其他模型，但由于描述长度或内容质量方面的问题，这些模型未纳入最终流程（图 1）。以下小节展示了这些弃用模型针对两项任务生成的示例描述文本。

表10。使用**弃用模型**生成的艺术作品描述文本示例。LAVIS 不需要提示词。其余模型均收到以下指令：“你是一位艺术专家。请简洁地描述这件艺术作品，谈及其主题、传达的情绪或情感、主要艺术技巧与风格，以及可能所属的艺术时期或流派”。

| 图像 | 图像描述 |
|---|---|
| <br>蒂耶塞维尔山丘<br>卡米耶·毕沙罗 | **Qwen2.5-VL：** 这幅艺术作品描绘了一幅宁静的乡村景观，可能是以印象派风格创作的。画面展现了起伏的山丘、蜿蜒的溪流和零星分布的树木，笼罩在柔和的天空下。光线与色彩的运用营造出宁静的氛围。画面传达出平和与安宁的情绪。艺术家运用松散的笔触，色彩以大地色调为主，这有助于营造整体的印象派风格。光线似乎是自然光。风格和技法表明，这件作品可能属于印象主义运动。这一时期的特点是注重捕捉光线和氛围转瞬即逝的变化，通常通过运用鲜艳的色彩和清晰可见的笔触来实现。<br>**Qwen2.5-Omni：** 嗯，这件艺术作品似乎是一幅风景画。画面描绘的主题是一片田野，一条蜿蜒小径或溪流穿行其间。田野里有一些树木背景。它传达的情绪相当平和宁静，几乎能让人感受到场景中的平静。画中可能运用了松散的笔触，赋予作品一种印象主义的感觉。色彩柔和且彼此融合，营造出和谐的视觉效果。至于风格，可能是印象主义。创作时期可能是19世纪后期。如果你想进一步了解这幅画或其他艺术作品，欢迎随时问我。<br>**LAVIS:** 一幅描绘溪流穿过草地的画作。 |

表11。使用**弃用模型**生成的音频描述示例。所有模型均收到以下提示词：“你是一位音乐专家。请分析提供的音频片段，并给出准确的描述（约200字）。回答应包括：（一）识别出的音乐流派，（二）可辨识乐器列表，以及（三）对整体声音和风格的简要描述。请勿提问或寻求更多信息”。

| 音频 | 图像字幕 |
|---|---|
| **音频 #1 (001/001556)** | **Qwen2-Audio:** 这段提供的音频片段是一首属于电子音乐流派的实验性无人机音乐作品。它具有合成器、贝斯、吉他和鼓。这些乐器编制营造出空间感与超越时间感。这首作品的结构遵循重复模式。这种方式使每种乐器的音色都能充分共鸣并产生深层回响。<br>**Qwen2.5-Omni:** 嗯，据我所知，这首音乐的音乐流派似乎是氛围音乐。你可以听到许多合成器其中还加入了一些电子铺底音色。整体声音十分梦幻且富有氛围感。它的节奏速度较慢这增添了平静与开阔的感觉。就像漂浮在一个梦幻而超现实的地方。你想要你觉得呢？还有其他想让我分析的音乐片段吗？<br>人类：什么什么<br>**Musilingo:** 这段音频片段是一首歌曲现场演出的录音。这首歌融合了摇滚和流行音乐。录音中可以听到的乐器包括吉他、贝斯吉他和合成器。这首歌的整体声音充满活力。这首歌的风格轻快而活泼。录音的音质很高。 |

表 12. 各种配对方法（使用频谱图或音频生成图像）下，CLIP 嵌入之间余弦相似度分数的汇总统计量（最小值、最大值、均值和标准差）。最佳结果以**粗体**显示，次佳结果带下划线。

| 方法 | CLIP 嵌入相似度 | | | |
|---|---|---|---|---|
| | Min | Max | Avg | 标准差 |
| 艺术作品-频谱图 | **0.2466** | 0.7183 | 0.4355 | **0.0783** |
| 重新着色的艺术作品频谱图 | 0.2100 | **0.9200** | **0.7045** | 0.1422 |
| 艺术作品-音乐图像 | 0.1274 | 0.7759 | 0.4578 | 0.0931 |

### A.1 图像描述生成

表 10 展示了弃用模型生成的图像描述示例，包括 Qwen2.5-VL [4], Qwen2.5-Omni [66],和 LAVIS [34]。

### A.2 音频描述

表 11 展示了弃用模型生成的音频描述示例，包括 MusiLingo [17], Qwen2-Audio [13],和 Qwen2.5-Omni [66]。

## B 配对方法

我们测试了另外两种配对方法：根据频谱图将数字化艺术作品与音乐曲目配对；以及根据音乐曲目的描述生成有意义的图像，并将生成的图像与数字化艺术作品进行比较。实验采用分层抽样，从 A*rt*G*raph* 数据集中选取了100件艺术作品，并从 Free Music Archive数据集中选取了100条音轨，样本按艺术风格和音乐流派进行平衡。

### B.1 按声波匹配

在第一种方法中，我们将音乐曲目（音频）转换为频谱图，并将其与原始艺术作品图像和重新着色版本（即使用不同颜色映射渲染的艺术作品）进行比较。重新着色旨在模拟频谱图的行为，即较亮的像素表示更高的能量，较暗的像素表示更低的能量。稿件已提交至 ACM



(a) Henryk Rodakowski *"Stróż"* (1868年)

(b) Antoine Blanchard *"Théâtre du Vaudeville"*

(c) Henri Fantin-Latour "花卉、大丽花与剑兰"（1879年）

图6. 比较 艺术作品与其使用*Inferno* 颜色映射的重新着色版本。每个子图展示原始 I
艺术作品（左）与其重新着色版本（右）。

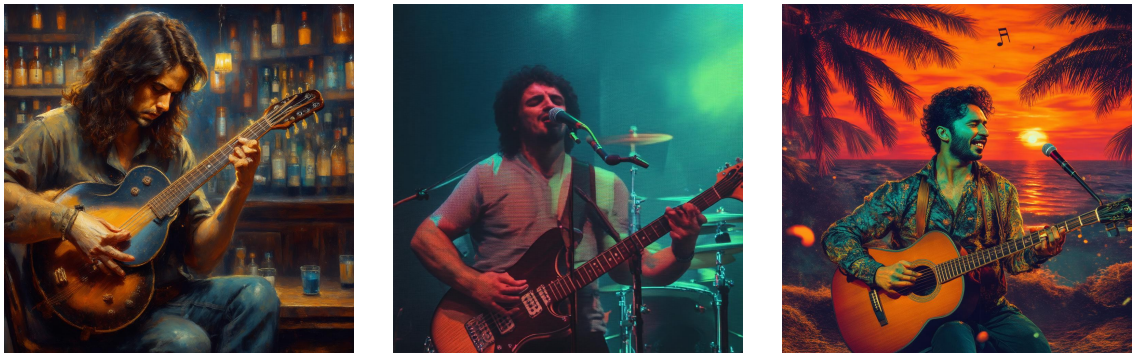
我们使用 CLIP 从艺术作品和频谱图中提取图像嵌入，并计算它们之间的余弦相似度。表 12 总结了结果。将艺术作品与频谱图进行比较（艺术作品-频谱图）产生了较低的平均相似度（0.4355），且标准差最低（0.0783）。尽管分数较为一致，但语义对齐较弱，因为大多数匹配对并不具有实际意义。

随后，我们将*Inferno* 颜色映射应用于每件艺术作品，以更好地模拟频谱图能量模式，从而有效地将每个像素映射为声音强度表示。示例见图6。尽管这种方法（重新着色的艺术作品频谱图）实现了更高的均值相似度（0.7045），定性评估仍表明，所得匹配并无实际意义。

### B.2 看见声音

在第二种方法中，我们根据*ArtSound* 数据集中的描述文本，为每条音乐曲目生成一幅图像。这些描述文本被提供给 Stable Diffusion [54], , 具体而言是 `stabilityai/stable-diffusion-3.5-medium`，以生成具有代表性的图像。生成图像的示例见图7。

与艺术作品–频谱图比较一样，我们计算了艺术作品与音频生成图像的 CLIP 嵌入之间的余弦相似度，但未观察到任何改进（见表12）。



(a) 图像 生成 for 音频 file 105718.mp3

(b) 图像 生成的 for 音频 file 063250.mp3

(c) 图像 生成的 for 音频 file 002099.mp3

图 7. 基于生成的音频描述文本，由 Stable Diffusion 3.5 生成的图像示例。

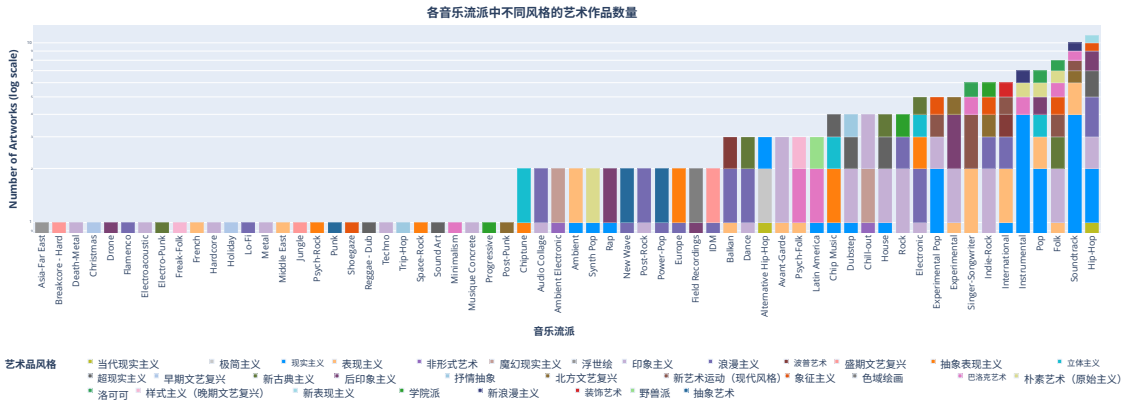



图 8. 艺术作品的分布 s
大多数流派，w 按音乐流派划分的艺术作品，按艺术风格分段。浪漫主义和表现主义在大多数流派中都很常见
极简主义和抽象艺术则较少出现。

不过，这种方法为探索艺术作品与音乐流派之间的关联提供了见解（图8）。例如，浪漫主义常与嘻哈和流行音乐联系在一起，因为二者都强调情感表达和个人叙事；表现主义则与重金属音乐相关联，反映出它们都着重表现强烈而原始的情感。这些模式表明，根据音乐描述文本生成的图像可能会唤起更广泛的文化或情感原型，不过，实验样本规模较小可能影响了相关性。后续研究应评估这种配对策略的稳健性。