arXiv:2603.11089v1 [cs.SD] 11 Mar 2026

V2A-DPO：全方位偏好 优化视频到音频生成

Nolan Chan[*, 1], *Timmy Gang*[*, 2], *Yongqian Wang*[3], *Yuzhe Liang*[4], *Dingdong Wang*[1]

[1] 香港特别行政区香港中文大学，中国 [2] 加拿大国家研究委员会，加拿大 [3] 华威大学，英国 [4] 上海交通大学，中国

摘要

本文介绍了 V2A-DPO，这是一种专为基于流的视频到音频生成（V2A）模型设计的新型直接偏好优化（DPO）框架，并针对性地进行了关键改进，以有效地使生成音频与人类偏好对齐。我们的方法包含三项核心创新：（1）AudioScore——一套全面的人类偏好对齐评分系统，用于评估合成音频的语义一致性、时间对齐和感知质量；（2）由 AudioScore 驱动的自动化流水线，用于生成大规模偏好对数据，以进行 DPO 优化；（3）专为基于流的生成模型量身定制的课程学习赋能 DPO 优化策略。在基准 VGGSound 数据集上的实验表明，使用 V2A-DPO 进行人类偏好对齐的 Frieren 和 MMAudio，表现优于使用去噪扩散策略优化（DDPO）进行优化的对应模型以及预训练基线。此外，经 DPO 优化的 MMAudio 在多项指标上取得最先进性能，超越了已发表的视频到音频模型。演示请见 https://seujames23.github.io/V2A-DPO/。

索引词— 视频转音频、直接偏好优化、人类偏好对齐、流匹配、课程学习

1. 引言

视频到音频生成（V2A）旨在以视频特征为条件（可选地结合文本提示），合成与视频高度对齐的音频，通常会考虑语义和时间信息。通过将无声视频与高质量且语义一致的音频相匹配，视频到音频模型能够补充现代视频生成系统；后者主要专注于视觉合成 [1]。

近年来，视频到音频模型取得了显著进展。早期基于生成对抗网络的模型 [2–4]利用生成器根据视觉特征生成音频，并使用判别器区分生成音频与真实音频。基于Transformer的自回归模型近期取得的突破，为视频到音频任务提供了强大的能力 [5–10]。其中，ThinkSound [7] 和 DeepSound-V1 [10]利用思维链（CoT）推理，实现视频到音频的逐步交互式生成与编辑。此外，其他前沿方法 [1, 11–18] 引入了新颖的基于扩散或流匹配的模型来解决这一任务，并取得了显著的性能。

尽管发展迅速，先前的视频到音频模型仍存在若干显著局限。首先，它们的风格控制受限于训练时使用的视频—音频对，限制了风格变化的灵活性和精确度 [1–18]。当测试时遇到与训练数据存在显著差异的场景时，现有模型往往会生成风格不恰当的音频。其次，合成音频的美学质量仍难以通过显式奖励建模来评估——这是一个关键方面，却被大多数视频到音频方法所忽视 [1–18]。然而，即使生成音频在语义上相关且时间对齐，若缺乏美学质量，仍可能无法让听众产生沉浸感。第三，先前的方法通常采用彼此独立的定量指标，分别评估生成音频的语义对齐、时间对齐或感知质量 [1–18]。然而，缺乏能够全面整合多项指标的综合评分体系，大大限制了对视频到音频模型性能的准确评估。

为解决这些不足，本文提出 V2A-DPO，这是一种专为基于流的V2A生成模型设计的新型直接偏好优化（DPO） [19] 框架，并针对性地进行了关键改进，以有效使生成音频与人类偏好对齐。具体而言，我们首先提出 AudioScore，这是一套全面的人类偏好对齐评分系统，旨在同时评估生成音频的语义一致性、时间对齐和感知质量。随后，我们通过 AudioScore驱动流水线，将少量人工标注偏好对与自动生成的偏好对相结合，高效构建适用于 DPO 优化的大规模数据集。通过在 DPO 优化过程中引入课程学习 [20] 范式，我们的人类偏好对齐视频到音频模型显著提升了生成音频的感知质量、语义一致性、时间对齐，甚至美学吸引力。在基准 VGGSound 数据集上的实验表明，采用 V2A-DPO 优化的人类偏好对齐 Frieren [15] 和 MMAudio [18] 优于采用去噪扩散策略优化（DDPO） [21] 优化的对应模型及预训练基线，在 IS 上最高分别提升**1.81 个绝对值（相对提升 10.4%）**，在 IB 分数上提升**0.86 个绝对值（相对提升 2.6%）**，同时 DeSync 降低**0.09 个绝对值（相对降低 20.5%）**。此外，经 DPO 优化的 MMAudio 在多项指标上达到最先进性能，超越了已发表的视频到音频模型。

我们的主要贡献总结如下：

（a）我们率先将直接偏好优化应用于基于流的 V2A 模型，解决了使音频生成结果与人类偏好保持一致所面临的独特挑战。

(b) 我们针对 DPO 优化框架进行了多项关键改进，包括：(1) AudioScore——一个全面且与人类偏好对齐的评分系统，用于评估生成音频的语义一致性、时间对齐和感知质量；(2) 一个自动化的 AudioScore驱动流水线，用于生成

* 通讯作者：yjchen@se.cuhk.edu.hk

用于 DPO 优化的大规模偏好对数据；(3) 一种针对基于流的生成模型、由课程学习赋能的 DPO优化策略。

(c) 据我们所知，我们构建了首个高质量视频–文本提示–音频偏好对数据集，旨在使视频到音频模型与人类偏好对齐，并同时考虑语义一致性、时间对齐、感知质量和美学吸引力。

(d) 我们通过在两个开源预训练视频转音频模型上开展大量实验，验证了所提出的 V2A-DPO 框架。结果表明，我们的方法在多项指标上均具有稳健性和有效性。

2. 方法

我们提出的 V2A-DPO（面向视频到音频生成的全方位偏好优化）框架如图 1 所示，包括一个全面的人类偏好对齐评分系统、一个自动化偏好对数据生成流程，以及一种面向视频到音频模型的课程学习赋能的 DPO 优化策略。

2.1. AudioScore

生成音频的质量受多种因素影响，包括语义一致性、时间对齐、感知质量和美学吸引力。语义一致性指生成音频与视频或文本提示之间的内容一致性。另一方面，时间对齐关注生成音频是否准确地跟随视频。第三，感知质量包括生成音频的清晰度和丰富度。最后，美学吸引力主要关注生成音频能否为听众营造沉浸式体验，即使其在前三项标准上得分很高。首先，我们从 VGGSound [22] 训练集中随机抽取 2K 个带有文本提示的视频，并通过我们计划进行对齐的预训练视频到音频模型生成 10K 个音频样本。随后，人工标注员对这些生成样本进行评分，并将其分为“优”“中”和“差”三类，如图 1(a) 所示。

为解决人工标注成本高这一难题，我们提出了 AudioScore，这是一套与人类偏好对齐的综合评分系统，由若干权重冻结的基础模型、MLP 和 Softmax 模块组成。如图 1(a) 所示，AudioScore 计算 ImageBind 从输入视觉内容和生成音频中提取的特征之间的平均余弦相似度 [23]，作为视频-音频语义一致性分数（IB 分数 [5]； $score_1$）。同样，若有文本提示，AudioScore 会利用 CLAP [24] 测量生成音频与文本提示之间的语义一致性分数（$score_2$）。此外，AudioScore 采用文献 [18], 提出的同步分数（DeSync； $score_3$），该分数由 Synchformer [25] 预测，表示音频与视频之间未对齐的时长（秒）。此外，AudioScore 根据 [15], , 使用基于 PANNs 的 Inception 分数 [26] （$score_4$）评估生成质量，并结合客观指标语音感知质量评估分数（PESQ； $score_5$），评估人类语音类别的音频质量。在获得五维分数向量后，AudioScore 使用两个线性层（中间设有一个 ReLU ）和一个 Softmax 模块，并通过交叉熵损失 $\mathcal{L}_{CE}$，使生成音频样本的自动分类结果与人工标注结果对齐。

2.2. 全模态偏好对数据生成

为了构建大规模偏好对数据集，我们结合使用 AudioScore 和基于分类概率的“最佳与最差”选择策略。如图 1(b) 所示，对于每个提示，我们使用旨在对齐的预训练视频到音频模型生成多个样本，并选取一个偏好对，其中“优”的概率最高的样本为胜出样本，“差”的概率最高的样本为落败样本。我们从 VGGSound 训练集中随机抽取 50K 个带有文本提示的视频，经过再次筛选后，获得约 46K 个全方位偏好对。鉴于通过 AudioScore 定量评估美学吸引力颇具挑战性，我们将自动生成的偏好对与 2K 个人工标注偏好对相结合，构成一个总计约 48K 个偏好的 DPO训练数据集。**音频生成与评分**。 给定一个视频 v （可选带有文本提示 c），我们使用预训练的视频到音频模型生成一组 N音频 a_1， a_2，……， a_N。对于每段生成音频 a_i，我们应用 AudioScore S(·)，根据视频 v （可选带有文本提示 c）对其质量进行评分。该评分系统会预测每段音频的偏好概率向量$\mathbf{p}_i$：

$$\begin{aligned}\mathbf{p}_i &= \Big(p_i(\text{Good}|a^i, v, c), p_i(\text{Medium}|a^i, v, c), p_i(\text{Bad}|a^i, v, c)\Big) \\ &= \text{S}(a^i, v, c), \text{ for } i = 1, 2, ..., N.\end{aligned} \tag{1}$$

其中， $p_i(*|a^i,\ v,\ c)$表示在视频 v 以及可选文本提示 c 的条件下，第 i个生成音频样本属于相应类别的预测概率，其中 $*$ 属于 {“优”“中”“差”}之一。

偏好对选择。 在具有相应概率向量的 N 个生成音频中，$\mathbf{p}_1$，$\mathbf{p}_2$，……，$\mathbf{p}_N$，我们选择“优”类别概率最高的音频作为胜出样本a^w，并选择“差”类别概率最高的视频作为落败样本 a^l。这一选择过程形式化如下:

$$\begin{aligned}(a^w, a^l) = (a^i, a^j),\ i = arg\ max\ p_i(\text{Good}|a^i, v, c), \\ j = arg\ max\ p_j(\text{Bad}|a^j, v, c),\ \text{for } i, j = 1, 2, ..., N.\end{aligned} \tag{2}$$

2.3. 课程学习赋能的DPO

尽管DPO在文本转语音、文本生成图像和文本生成视频任务中取得了令人鼓舞的成果 [27–37], ，但研究已证明，训练时不考虑偏好对的难度、而是随机排序现有偏好对并非最优，因为困难偏好对之间的差异很小，模型难以有效区分 [32, 37]。为此，我们为视频到音频模型引入了一种课程学习赋能的DPO优化策略，根据偏好对的复杂度分数将训练过程划分为两个不同阶段。第一阶段使用差异明显、易于区分的偏好对对模型进行对齐，第二阶段则使用差异较细微、更加微妙的偏好对。通过课程学习，模型首先从较简单的偏好对中学习更有意义的对齐线索，再在后续阶段逐步处理更复杂的偏好对。这种结构化方法使模型的生成能力能够更加稳定、循序渐进地提升。

基于概率的复杂度分数。对于输入视频 v 和可选的文本提示 c，给定一个偏好对（a^w， a^l），该偏好对的复杂度分数 $score_c$ 可根据公式 3 计算。如图 1(c) 所示，当 $score_c$ 超过预设阈值 $score_\Delta$时，偏好对数据将输入第一阶段训练；否则则输入第二阶段。请注意，所有 2K 个人工标注偏好对的复杂度分数均设为零，因为我们希望优化后的模型在第二阶段专注于生成音频的美学吸引力。

$$\begin{aligned}score_c = \frac{1}{2}\Big[\Big(p^w(\text{Good}|a^w, v, c) - p^l(\text{Good}|a^l, v, c)\Big) \\ + \Big(p^l(\text{Bad}|a^l, v, c) - p^w(\text{Bad}|a^w, v, c)\Big)\Big]\end{aligned} \tag{3}$$

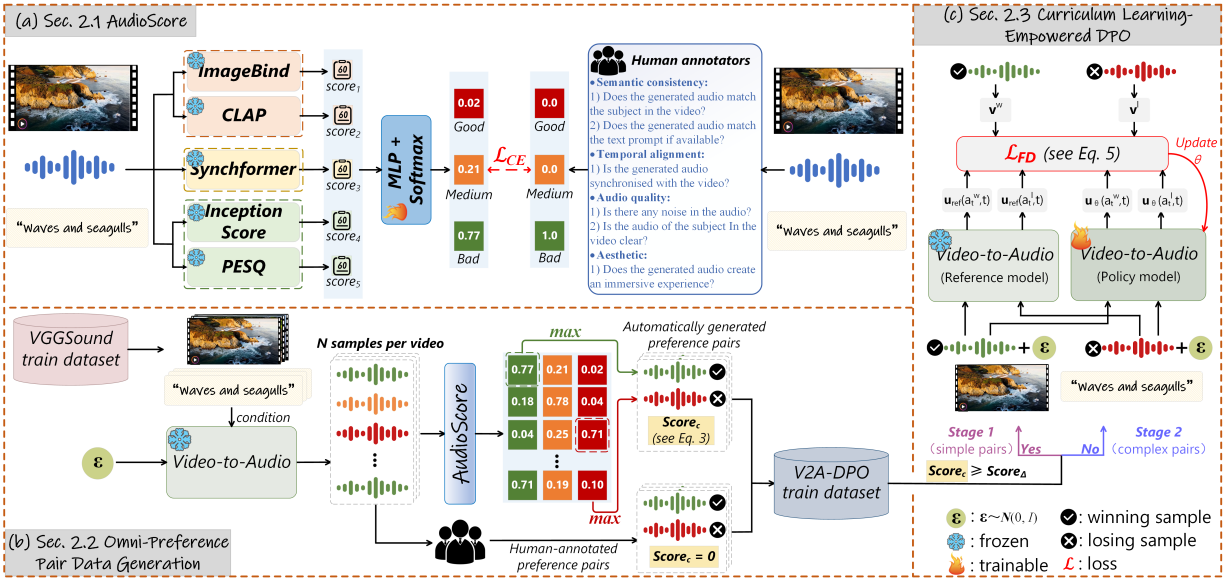


图 1。所提出的 V2A-DPO 框架示意图，包括：(a) 我们提出的 AudioScore，利用多维分数对生成音频进行评估；(b) 全模态偏好对数据生成，将基于 AudioScore 自动生成的偏好对与少量人工标注偏好对相结合；(c) 课程学习赋能的DPO，根据复杂度分数 $score_c$ 划分复杂配对和简单配对，并逐步对视频到音频模型进行优化。

其中，$p^w(优|a^w,\ v,\ c)$ 和 $p^l(优|a^l,\ v,\ c)$ 分别表示获胜样本和落败样本的“优”概率，而 $p^w(差|a^w,\ v,\ c)$ 和 $p^l(差|a^l,\ v,\ c)$ 则表示“差”的概率。

Flow-DPO。给定固定数据集 $\mathcal{D}=\{(v,\ c,\ a_0^w,\ a_0^l)\}$，其中每个样本包含一个视频 v、一个可选的文本提示 c 以及由参考模型 q_{ref}生成的一对音频，其中 a_0^w 优于a_0^l，用于优化策略模型 u_θ 的DPO奖励目标可以表示为

$$\mathcal{L}_{DPO}(\theta)=-\mathbb{E}_{(v,c,a_0^w,a_0^l)\sim\mathcal{D}}\Big[log\sigma\Big(\beta\,log\frac{q_\theta(a_0^w|v,c)}{q_{\text{ref}}(a_0^w|v,c)}-\beta\,log\frac{q_\theta(a_0^l|v,c)}{q_{\text{ref}}(a_0^l|v,c)}\Big)\Big] \tag{4}$$

其中，σ 是 Sigmoid 函数。a_0^* 指从 $\mathcal{D}$ 中采样得到的目标音频，该采样过程采用整流流匹配，且上标 $*$表示“w”（获胜样本）或“l”（落败样本）。

在将 DPO 适配到基于流的模型时，[38] 将对齐 v 解释为一个分类问题，并通过监督训练优化策略，以满足人类偏好。为简化起见，在下面的公式中，我们省略条件视频 v 和可选的文本提示 c。Flow-DPO 目标 $\mathcal{L}_{FD}(\theta)$定义为：

$$-\mathbb{E}\Big[log\sigma\Big(-\frac{\beta_t}{2}\Big(\|\mathbf{v}^w-\mathbf{u}_\theta(a_t^w,t)\|^2-\|\mathbf{v}^w-\mathbf{u}_{\text{ref}}(a_t^w,t)\|^2\Big)-\Big(\|\mathbf{v}^l-\mathbf{u}_\theta(a_t^l,t)\|^2-\|\mathbf{v}^l-\mathbf{u}_{\text{ref}}(a_t^l,t)\|^2\Big)\Big)\Big] \tag{5}$$

其中 $a_t^*=(1-t)a_0^*+t\epsilon,\ \epsilon\sim\mathcal{N}(0,\ I)$。$t\sim\mathcal{U}(0,\ 1)$ 是整流流匹配中的时间步。$\mathbf{u}_{\text{ref}}$ 和 $\mathbf{u}_\theta$ 分别表示策略模型 q_θ 和参考模型 q_{ref} 预测的向量场，而 $\mathbf{v}^w$ 和 $\mathbf{v}^l$ 分别表示胜出样本和落败样本的目标向量场。参数 β_t 控制 KL 散度约束的强度，并随 t 变化。在我们的实验中，我们采用常数 β 代替 β_t，遵循 [38]，中的方法，该方法已被证明能带来更好的性能。

直观而言，最小化 $\mathcal{L}_{FD}(\theta)$ 会引导预测向量场 $\mathbf{u}_\theta$ 更接近“偏好”样本的目标向量 $\mathbf{v}^w$，同时使其远离$\mathbf{v}^l$（“较不受偏好”的样本）。这种偏好信号的强度取决于预测误差与相应参考误差之间的差异，即$\|\mathbf{v}^w-\mathbf{u}_{\text{ref}}(a_t^w,\ t)\|^2$ 和 $\|\mathbf{v}^l-\mathbf{u}_{\text{ref}}(a_t^l,\ t)\|^2$。

3. 实验

3.1. 实验设置

数据集。如第 2.2 节所述，我们的所有实验均基于 VGGSound [22]数据集构建的全偏好对数据开展，该数据集为每个视频包含一个类别标签（总计 310 个类别）。为进行公平比较，鉴于数据污染，我们在 MMAudio [18] 的同一测试集上评估模型。

生成模型。我们使用两个预训练的基于流的视频到音频模型开展实验：MMAudio-L-44.1kHz [18] 和 Frieren [15]，其参数量分别为 1.03B 和 159M。

基线。我们将优化后的视频到音频模型与五种最先进的模型进行比较：基于扩散的方法 Seeing&Hearing [39], Fo- leyCrafter [17], V2A-Mapper [16],，以及基于自回归的方法 V- AURA [5], ThinkSound [7]。此外，我们还将直接偏好优化与另一种强化学习方法 DDPO [21], 进行比较，后者使用 AudioScore 预测的“优”概率作为奖励。

评估指标。我们使用与 [18]相同的指标评估生成质量。具体而言，按照 [41] 的方法，我们使用 PaSST [40]作为嵌入模型计算 Fr´echet 距离（FD_{PaSST}），并使用 PANNs 作为分类器计算库尔贝克-莱布勒距离（KL_{PANNs}），以及使用 PaSST 作为分类器计算库尔贝克-莱布勒距离（KL_{PaSST}），以评估真实音频与生成音频之间的特征分布相似度。此外，我们使用 Inception分数 [26], IB 分数 [5], 和 DeSync [18] 评估感知质量、语义一致性和时间对齐。

实现细节。优化模型的所有设置均与对应的预训练模型相同，即 MMAudio-L-44.1kHz 和 Frieren。我们使用 AdamW 优化器训练视频到音频模型 12K 步，全局批量大小为 8，学习率为 5e-6，线性预热步数为 1K。对于每个提示，

表1。VGGSound测试集上的视频转音频结果。 ⋄: [18] 中使用相同测试集的评估结果。 †: 使用官方评估代码复现。 ⋆: 测试时不使用文本提示。

Sys.	方法	参数	分布匹配			音频质量	语义对齐	时间对齐
			$\text{FD}_{\text{PaSST}}\downarrow$	$\text{KL}_{\text{PANNs}}\downarrow$	$\text{KL}_{\text{PaSST}}\downarrow$	IS↑	IB 分数↑	DeSync↓
1	Seeing&Hearing ⋄	415M	219.01	2.26	2.30	8.58	33.99	1.20
2	V-AURA ⋄⋆	695M	218.50	2.42	2.07	10.08	27.64	0.65
3	FoleyCrafter ⋄	1.22B	140.09	2.30	2.23	15.68	25.68	1.23
4	V2A-Mapper ⋄⋆	229M	84.57	2.69	2.56	12.47	22.58	1.23
5	ThinkSound †	1.30B	54.92	**1.32**	1.52	16.03	**34.13**	0.46
6	Frieren ⋆	159M	106.10	2.73	2.86	12.25	22.78	0.85
7	Frieren-DDPO ⋆	159M	75.41	2.58	2.61	13.12	23.19	0.65
8	Frieren-DPO ⋆	159M	69.98	2.55	2.63	13.98	24.11	0.62
9	MMAudio	1.03B	60.60	1.65	1.40	17.40	33.22	0.44
10	MMAudio-DDPO	1.03B	54.81	1.42	1.36	18.34	33.89	0.39
11	MMAudio-DPO	1.03B	**51.38**	1.38	**1.34**	**19.21**	34.08	**0.35**

表 2。关于 KL 散度约束参数 β 和预设阈值 $score_\Delta$ 对使用 V2A-DPO 优化的 MMAudio 性能影响的消融研究。

Sys.	β	$score_\Delta$	$\text{FD}_{\text{PaSST}}\downarrow$	IS↑	IB 分数↑	DeSync↓
0	600	0.7(81%)	**51.38**	19.21	**34.08**	**0.35**
1	1000	0.7(81%)	82.36	11.15	24.88	0.43
2	800	0.7(81%)	64.85	17.33	29.91	0.38
3	400	0.7(81%)	53.46	**19.46**	33.92	0.35
4	200	0.7(81%)	55.81	18.79	33.93	0.36
5	600	1.0(0%)	52.66	18.83	33.79	0.36
6	600	0.9(8%)	52.58	18.72	33.90	0.36
7	600	0.8(45%)	51.96	19.02	33.98	0.35
8	600	0.6(89%)	51.63	19.10	34.01	0.35

生成音频的数量 N 设为 5。此外，KL 散度约束参数 β 和预设阈值 $score_\Delta$ 分别设为 600 和 0.7。推理时，我们将无分类器引导（CFG） [42] 中的引导强度 γ 设为 4.5。所有实验均在 8 张 NVIDIA A100 GPU 上进行。

3.2. 实验结果

直接偏好优化模型、DDPO 优化模型与基础模型的性能比较。如表 1 所示，使用 V2A-DPO 实现人类偏好对齐的 MMAudio（系统 11 ）在音频质量和语义对齐方面始终优于经 DDPO 优化的模型（系统 10）和预训练模型（系统 9），IS 与 IB 分数的绝对增幅分别最高达 1.81 和 0.86（相对增幅为 10.4% 和 2.6%；系统 11 对比系统 9）。此外，优化后的 MMAudio 的时间对齐性能显著提升，DeSync 的绝对降幅达 0.09（相对降幅为 20.5%；系统 11 对比系统 9）。如图 2 直观所示，经直接偏好优化的 MMAudio 能够同时与演示中“轻柔拨弦”（蓝色虚线）和“快速反复拨弦”（绿色虚线）的手部动作对齐，而经 DDPO 优化的 MMAudio 与预训练 MMAudio 模型则未能做到。在 Frieren 优化过程中也可观察到相同趋势（系统 8 对比系统 6、7）。

DPO 优化模型与已发表的最先进视频到音频模型之间的性能比较。采用 V2A-DPO 优化后的 MMAudio 在表 1 的多项指标上表现优于其他模型（系统 11 对比系统 1–5、6、9），但 KL_{PANNs} 和 IB 分数除外。我们推测，这种差异可能

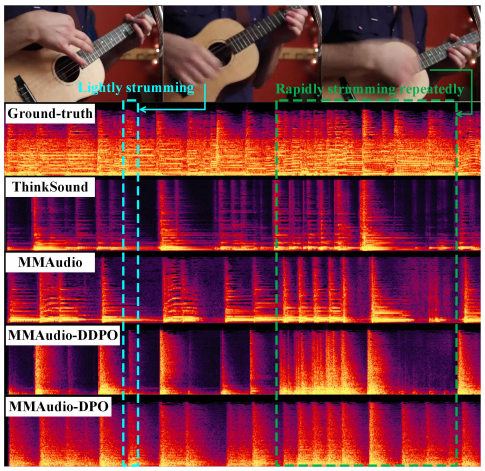


图2。视频到音频模型生成性能示意图。

归因于 ThinkSound 使用思维链分析视觉线索并提升语义对齐。更多演示可在我们的网站上查看。**消融研究**。表 2 展示了 KL 散度约束参数 β 和预设阈值 $score_\Delta$ 对 DPO 优化后 MMAudio 性能的影响。我们观察到，当 β 等于 600 时，优化后的模型能够生成语义一致性更高、时间对齐效果更好的音频。此外，DPO 优化模型的性能会随 $score_\Delta$ 变化，因为在第一阶段使用的简单配对比例不同。请注意，借助课程学习的 DPO 会退化为常规 DPO，模型性能也会显著下降。

4. 结论

我们提出了 V2A-DPO，这是一种专为基于流的 V2A 模型量身打造的新型 DPO框架，旨在有效地使生成音频符合人类偏好。我们的方法包含三项关键改进：（1）AudioScore；（2）自动化偏好对数据生成流程；（3）由课程学习赋能的 DPO 优化策略。在 VGGSound数据集上的实验表明，采用 V2A-DPO 优化的 Frieren 和 MMAudio 优于经 DDPO 优化的模型和预训练基线。此外，经 DPO 优化的 MMAudio 在多项指标上达到了最先进性能，超越了已发表的视频到音频模型。

5. 参考文献

[1] Manjie Xu 等人，“带隐藏对齐的视频转音频生成”，arXiv 预印本 arXiv:2407.07464，2024年。[2] Peihao Chen 等人，“从视频生成视觉对齐的声音”，IEEE 图像处理汇刊，2020年。[3] Kun Su 等人，“Audeo：为无声表演视频生成音频”，NeurIPS，2020年。[4] Sanchita Ghose 等人，“Foleygan：在无声视频中实现视觉引导的生成对抗网络式同步声音生成”，IEEE 多媒体汇刊，2022年。[5] Ilpo Viertola 等人，“利用自回归为视频生成时间对齐音频”，载于 ICASSP，2025年。[6] Chuang Gan 等人，“拟音音乐：学习从视频生成音乐”，载于 ECCV，2020年。[7] Huadai Liu 等人，“ThinkSound：用于音频生成与编辑的多模态大语言模型中的思维链推理”，arXiv 预印本 arXiv:2506.21448，2025年。[8] Roy Sheffer 等人，“我听见你真实的色彩：图像引导的音频生成”，载于 ICASSP，2023年。[9] Xinhao Mei 等人，“Foleygen：视觉引导的音频生成”，载于 MLSP，2024年。[10] Yunming Liang 等人，“DeepSound-V1：开始逐步思考如何从视频生成音频”，arXiv 预印本 arXiv:2503.22208，2025年。[11] Ziyang Chen 等人，“利用多模态控制实现视频引导的拟音生成”，载于 CVPR，2025年。[12] Shentong Mo 等人，“Foley-flow：利用掩码音视频对齐和动态条件流实现协同视频转音频生成”，载于 CVPR，2025年。[13] Si mian Luo 等人，“Diff-foley：利用潜在扩散模型实现同步视频转音频合成”，NeurIPS，2023年。[14] Changan Chen 等人，“Action2sound：从第一人称视角视频生成具备环境感知能力的动作声音”，载于 ECCV，2024年。[15] Yongqi Wang 等人，“Frieren：利用整流流匹配实现高效视频转音频生成”，NeurIPS，2024年。[16] Heng Wang 等人，“V2A-Mapper：通过连接基础模型实现视觉转音频生成的轻量级方案”，载于 AAAI，2024年。[17] Yiming Zhang 等人，“FoleyCrafter：用逼真且同步的声音让无声视频栩栩如生”，arXiv 预印本 arXiv:2407.01494，2024年。[18] Ho Kei Cheng 等人，“MMAudio：通过调控多模态联合训练，实现高质量视频转音频合成”，载于 CVPR，2025年。[19] Rafailov Rafailov 等人，“直接偏好优化：你的语言模型实际上是一个奖励模型”，NeurIPS，2023年。[20] Yoshua Bengio 等人，“课程学习”，载于 ICML，2009年。[21] Kevin Black 等人，“利用强化学习训练扩散模型”，arXiv 预印本 arXiv:2305.13301，2023年。[22] Honglie Chen 等人，“VGGSound：大规模音视频数据集”，载于 ICASSP，2020年。[23] Rohit Girdhar 等人，“ImageBind：一个将万物绑定在一起的嵌入空间”，载于 CVPR，2023年。

[24] Benjamin Elizalde 等人，“通过自然语言监督学习音频概念”，发表于*ICASSP*，2016年。[25] Vladimir Iashin 等人，“Synchformer：利用稀疏线索实现高效同步”，发表于*ICASSP*，2024年。[26] Tim Salimans 等人，“改进生成对抗网络的训练技术”，发表于*NeurIPS*，2016年。[27] Xiaoxue Gao 等人，“Emo-DPO：通过直接偏好优化实现可控的情感语音合成”，发表于*ICASSP*，2025年。[28] Dong Zhang 等人，“SpeechAlign：使语音生成与人类偏好保持一致”，发表于*NeurIPS*，2024年。[29] Jinchuan Tian 等人，“偏好对齐提升基于语言模型的文本转语音”，发表于*ICASSP*，2025年。[30] Navonil Majumder 等人，“Tango 2：通过直接偏好优化使基于扩散模型的文本到音频生成保持一致”，发表于*ACMMM*，2024年。

[31] Huakang Chen 等人，《Diffrhythm+：通过偏好优化实现可控且灵活的全长歌曲生成》，*arXiv* 预印本 *arXiv:2507.12890*，2025年。

[32] R. Liu 等人，《Videodpo：面向视频扩散生成的全模态偏好对齐》，载于*CVPR*，2025年。

[33] Bram Wallace 等人，《使用直接偏好优化进行扩散模型对齐》，发表于*CVPR*，2024年。

[34] Qihan Huang 等人，《Dpo-kid：用于微调免费个性化图像生成的补丁级直接偏好优化》，发表于*CVPR*，2025年。

[35] Zhendong Wang 等，《DesignDiffusion：使用扩散模型进行高质量文本到设计图像生成》，发表于*CVPR*，2025年。[36] Sanghyeon Na 等，《通过直接偏好优化提升人物图像生成模型》，发表于*CVPR*，2025年。[37] Florinel-Alin Croitoru 等，《面向扩散模型和一致性模型的课程式直接偏好优化》，发表于*CVPR*，2025年。[38] Jie Liu 等，《利用人类反馈改进视频生成》，*arXiv* 预印本 *arXiv:2501.13918*，2025年。[39] Yazhou Xing 等，《视听共赏：通过扩散潜在对齐器实现开放域视觉-音频生成》，发表于*CVPR*，2024年。[40] Khaled Koutini 等，《使用 Patchout 高效训练音频变换器》，国际语音通信协会，2022年。[41] Haohe Liu 等，《AudioLDM：基于潜在扩散模型的文本到音频生成》，发表于*ICML*，2023年。[42] Jonathan Ho 等，《无分类器扩散引导》，*arXiv* 预印本 *arXiv:2207.12598*，2022年。