

Conditional Flow Matching for Visually-Guided Acoustic Highlighting

Hugo Malard^{1*}, Gael Le Lan², Daniel Wong², David Lou Alon², Yi-Chiao Wu²,
and Sanjeel Parekh²

¹ LTCI, Télécom Paris, Institut Polytechnique de Paris

² Meta

Abstract. Visually-guided acoustic highlighting seeks to rebalance audio in alignment with the accompanying video, creating a coherent audio-visual experience. While visual saliency and enhancement have been widely studied, acoustic highlighting remains underexplored, often leading to misalignment between visual and auditory focus. Existing approaches use discriminative models, which struggle with the inherent ambiguity in audio remixing, where no natural one-to-one mapping exists between poorly-balanced and well-balanced audio mixes. To address this limitation, we reframe this task as a generative problem and introduce a Conditional Flow Matching (CFM) framework. A key challenge in iterative flow-based generation is that early prediction errors — in selecting the correct source to enhance — compound over steps and push trajectories off-manifold. To address this, we introduce a rollout loss that penalizes drift at the final step, encouraging self-correcting trajectories and stabilizing long-range flow integration. We further propose a conditioning module that fuses audio and visual cues before vector field regression, enabling explicit cross-modal source selection. Extensive quantitative and qualitative evaluations show that our method consistently surpasses the previous state-of-the-art discriminative approach, establishing that visually-guided audio remixing is best addressed through generative modeling. Qualitative samples are available at the project page.

Keywords: Audiovisual Learning · Smart Remixing · Flow Matching

1 Introduction

The recent proliferation of video content creation and consumption underscores the need for meticulous curation of both visual and audio elements to deliver an engaging user experience. While visual cue manipulation—through techniques such as optimal viewpoint selection or post-editing—has been a long-standing focus in media production [20, 23, 26], the audio domain has not seen equivalent advancements. This disparity frequently results in a noticeable disconnect between visual saliency and acoustic emphasis. For instance, when a person speaks on camera, viewers intuitively expect their voice to be clear and prominent. However,

* Work done during an internship at Meta.

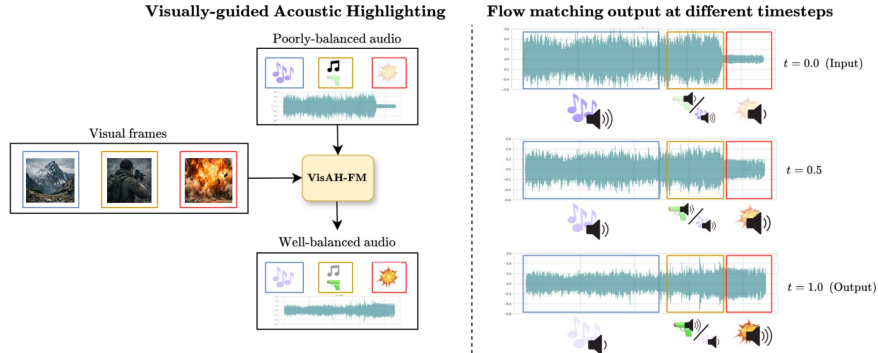


Fig. 1: VisAH-FM casts visually-guided acoustic highlighting as flow matching: sources are iteratively enhanced when audio-visual cues agree and subdued otherwise.

without deliberate mixing, their speech may be overwhelmed by background noise or music, breaking the perceptual alignment between what is seen and what is heard. Raw daily recordings, in particular, often contain poorly balanced audio due to recording device limitations where microphones attached to video cameras capture all sounds indiscriminately.

To bridge this crucial gap, the task of Visually-guided Acoustic Highlighting (VisAH) was introduced [13], aiming to automatically transform poorly balanced audio by using the accompanying video stream as guidance, ultimately creating a more harmonious audio-visual experience. This process involves rebalancing loudness of different audio sources, like speech, music, and sound effects, to reflect their relative prominence as implied by the visual context.

Previously, this task was approached using a discriminative paradigm [13], relying on sound source separation architectures, such as DEMUCS [7]. Thus, training for a one-to-one mapping. However, we observe that given the video stream, there are several ways of creating poor or well-balanced audio mixes. In other words, audio remixing (or rebalancing) is fundamentally a task of transporting from one distribution to another. Given its many-to-many nature we argue that a generative approach is better suited than a purely discriminative one. Indeed, prior work [29] shows that in many-to-many settings, where both input and output contain mutual uncertainty, generative models can exploit information from the marginal to capture correlations and latent structure that discriminative models ignore. Additionally, recent works have shown that, for multiple audio tasks where there is no clear one-to-one mapping, like speech enhancement or audio source separation, generative approaches significantly outperform their discriminative counterparts [1, 14, 27, 36].

Among generative modeling techniques, flow matching [24] provides a principled framework for learning a continuous transformation between input and target distributions, making it a natural fit for remixing as shown in Figure 1. In this formulation, a neural network learns the vector field that guides samples from

the source distribution towards the target through a sequence of intermediate states. For the visually guided acoustic highlighting task, the main difficulty lies in identifying the audio sources to enhance or suppress. Ideally, this crucial operation must occur from the very first step of the flow matching process, where the model begins transforming the input distribution towards the target one. As this step is inherently difficult, errors, even if small, can easily arise early on. In an iterative framework like flow matching, such early inaccuracies propagate through subsequent steps, compounding over time and driving the trajectory away from the data manifold. This accumulation of error substantially degrades model performance and motivates our proposal of the *rollout loss*, an additional loss to stabilize predictions across steps.

Furthermore, the source identification problem heavily relies on the conditioning of the model. In prior work, this was based solely on visual cues—either features extracted from an image encoder or textual captions describing the entire visual scene. These image-only features were passed through a small transformer and then fed to the main denoising model. Consequently, the latter had to implicitly learn the correspondence between auditory and visual signals—a task that goes beyond its primary goal of regressing the audio from a conditioning signal. We argue that this “division of responsibilities” can be made more effective. By integrating audio features directly into the visual encoder, we enable cross-modal filtering within the conditioning pathway itself. Such early fusion allows the conditioning network to preserve only the visual components relevant to the target sound sources, yielding a compact, audio-aware representation. The main model then receives conditioning features that are already semantically aligned with the audio content, allowing it to “concentrate” entirely on accurate audio regression rather than on discovering cross-modal correspondences.

In summary, our contributions are threefold:

- We reformulate the visually guided acoustic highlighting problem as a *generative* task rather than a discriminative one, and demonstrate the advantages of this formulation both qualitatively and quantitatively.
- We introduce a *rollout loss* that mitigates the error accumulation inherent to standard flow matching for such a task, and analyze its impact in detail.
- We design an improved *conditioning module* that enables earlier fusion of audio and visual features before the velocity field estimation stage, improving performance and achieving state-of-the-art results.

2 Proposed Method

2.1 Task Formulation

Visually-guided Acoustic Highlighting. The visually-guided acoustic highlighting task was introduced to address the disconnect often observed between visual and acoustic saliency in video content. Its objective is to utilize the accompanying video stream as guidance to transform a poorly balanced audio track into a well-balanced one exhibiting appropriate highlighting effects. To facilitate training

models for this task, the Muddy Mix Dataset [13] was curated, leveraging the meticulous audio-visual crafting found in movies, providing a form of "free supervision". The input audio, representing poorly mixed content is generated using a pseudo-data generation process that simulates real-world mixing deficiencies. This process systematically disturbs the original high-quality movie audio through three key steps: (1) Separation, decomposing the audio into source components (speech, music, and sound effects); (2) Adjustment, altering the relative levels of these separated sources by applying suppression or emphasis using selected strength levels (high, moderate, low); and (3) Remixing, linearly combining the adjusted sources to form the ill-balanced input audio.

2.2 Our Approach

Conditional Flow Matching (CFM). Unlike discriminative models, which learn a point-wise direct one-to-one mapping between inputs and targets, flow matching methods focus on aligning probability distributions. They model a continuous transformation between a source and target distribution, enabling many-to-many correspondences rather than explicit pairwise matches. Given two distributions from which we can sample during training, $(x_0, x_1) \sim \pi_0 \times \pi_1$, where $x_0 \in \mathbb{R}^d$ is typically drawn from a base distribution (e.g., a standard Gaussian) and $x_1 \in \mathbb{R}^d$ represents a sample from the target data distribution, flow matching learns a time-dependent velocity field that continuously transports samples from π_0 to π_1 [24, 42]. An interpolant between the two samples is defined as

$$x_t = (1 - t)x_0 + tx_1, \quad (1)$$

whose evolution follows the conditional ODE

$$\frac{dx_t}{dt} = \frac{x_1 - x_t}{1 - t}. \quad (2)$$

In practice, a neural network $v_\theta(x_t, t)$ is trained to approximate this velocity field by minimizing

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t, x_0, x_1} \left[\left\| \frac{x_1 - x_t}{1 - t} - v_\theta(x_t, t) \right\|_2^2 \right]. \quad (3)$$

Conditional Flow Matching for Acoustic Highlighting. We propose to cast the acoustic highlighting problem as an instance of conditional flow matching, where the model has to learn the flow that goes from a poorly-balanced to a well-balanced audio distribution, while being conditioned on the visual cues. We argue that this formulation is particularly well-suited to the task, as a given video may admit multiple plausible high-quality mixes, as well as diverse degraded versions. Consequently, matching distributions rather than individual instances better captures the underlying variability. More formally, plugging into the conditional flow matching setup mentioned above, we are trying to minimize:

$$\mathcal{L}_{\text{CFM}}(\theta) = \mathbb{E}_{t, x_0, x_1} [\| (x_1 - x_t)/(1 - t) - v_\theta(x_t, t, c) \|_2^2] \quad (4)$$

Here x_0 and x_1 are audios from poorly-balanced and well-balanced audio distributions, respectively, c is the visual conditioning, and $t \sim \mathcal{U}(0,1)$ the sampled timestep. During training, we expose the model to multiple distinct inputs corresponding to the same target across epochs. We do so by randomly generating poorly balanced audio samples on-the-fly as opposed to a fixed dataset used in [13].

Also, motivated by prior work that shows flow matching benefits from pre-trained discriminative initialization [21,47], we adapt the VisAH architecture [13] and training objective to the generative setting while initializing with its pre-trained weights. Additional implementation details in Sec. 3.

CFM Training with Rollout Loss. In a standard flow matching setup, each step is trained independently, and the ground-truth sample is used as input at every time step (i.e., the real x_0 is used to obtain x_t).

However, flow matching can inherently generate drift along the predicted trajectory: small local inaccuracies in the velocity field may compound over successive steps, gradually deviating the model’s trajectory from the target distribution. To mitigate this drift, we propose using backpropagation through the flow, enabling the model not only to learn the correct instantaneous velocity at each step but also to self-correct so that the overall trajectory converges toward the target endpoint.

In the acoustic highlighting task, this effect is particularly pronounced. The key challenge lies in determining which sources should be enhanced or attenuated—a decision that must be made during the earliest steps. Later steps primarily refine these earlier decisions. This asymmetry in step difficulty means that early-step prediction errors can have a disproportionate effect, propagating and amplifying throughout the entire trajectory.

To make the model more robust to such drift, we introduce a rollout loss that supervises the entire generated trajectory. Unlike typical flow matching applications that use hundreds of integration steps, acoustic highlighting operates on relatively close input–output distributions, allowing us to use only a few steps (four in our experiments, following [3]). This setup makes end-to-end backpropagation through the full flow computationally feasible, enabling the model to jointly optimize intermediate updates and final consistency.

Specifically, we perform a full rollout during training and add an auxiliary mean-squared-error (MSE) loss between the final predicted output and the ground-truth target (see Figure 2). This encourages consistency across steps and stability under self-generated predictions, effectively reducing the impact of early-step errors. The overall training objective is:

$$\mathcal{L}(\theta) = \mathcal{L}_{\text{CFM}}(\theta) + \lambda \mathbb{E}_{x_0} [\|\hat{x}_T - x_T\|_2^2]. \quad (5)$$

Where $\hat{x}_T$ denotes the model prediction after T recurrent applications of the flow operator starting from x_0 , and x_T is the corresponding ground-truth target. Unless otherwise stated, we set $T = 4$ steps (matching the total number of inference steps).

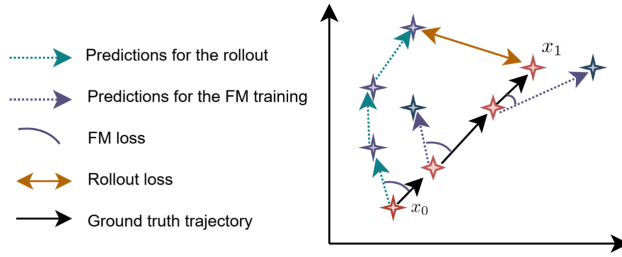


Fig. 2: Rollout loss supervises the output after all flow-matching steps, encouraging coherent trajectories and reducing compounding errors.

Relation to Consistency Models. Our rollout loss conceptually resembles strategies proposed in consistency models [38, 39] for diffusion or flow-based generation, which aim to stabilize iterative predictions by enforcing agreement between nearby timesteps. However, unlike consistency models that match predictions across infinitesimal noise scales or time intervals, our rollout loss supervises the final prediction after multiple recurrent steps against the ground truth trajectory. This makes it focus more on the exposure-bias mitigation [30, 35]: the network is explicitly trained to recover from its own intermediate errors rather than only ensuring local consistency. The optimization of this objective is performed via backpropagation through the flow, akin in mechanism to D-Flow [2]. Unlike the latter—which uses this process to optimize the model input—we use it to optimize the model weights. Empirically, we found that this end-to-end constraint yields more stable long-range trajectories and avoids the error-amplification issues observed when applying standard consistency loss (see Table 3).

Conditioning Module. As shown in Figure 3, we build upon the VisAH backbone [13], which consists of dual U-Net branches processing spectrogram and waveform representations, inspired by HybridDemucs [7]. Conditioning is performed in the latent space through a transformer-based module. In VisAH, frame-wise features extracted from a CLIP vision encoder [32] or from caption embeddings (T5 [33]) are processed by a temporal transformer and cross-attended into the U-Net latent space. However, the source to enhance is typically the dominant concept shared between the visual scene and the audio. Since the conditioning module in VisAH only accesses visual information, the burden of identifying the target source is largely shifted to the U-Net, which has access to the audio signal. This forces the latter to implicitly perform source selection in addition to the regression task. To better disentangle these roles, we aim for the conditioning module to determine the source to enhance, leaving the U-Net to specialize in vector field estimation.

Inspired by recent cross-modal adapter designs [6, 8], we propose an adapter layer that injects audio features from an additional audio encoder into intermediate layers of the CLIP model. The goal of this design is to let the CLIP representations dynamically attend to complementary audio cues, enriching the

visual embeddings with audio-specific context. Specifically, we compute cross-attention between a low-dimensional projection of both the CLIP and audio features, before mapping them back to the original embedding dimension:

$$m(F_k, E) = \text{Att}(EW_{\text{down}}^E, F_k W_{\text{down}}^F) W_{\text{up}}^E, \quad (6)$$

where F_k denotes the CLIP features at layer k , E the audio features, W_{down}^E and W_{down}^F the projection layers, W_{up}^E the up-projection layer, and Att denotes cross-attention. This operation effectively acts as a lightweight conditioning mechanism, enabling information exchange across modalities without retraining the CLIP backbone. Finally, we perform a weighted sum of the original CLIP features with

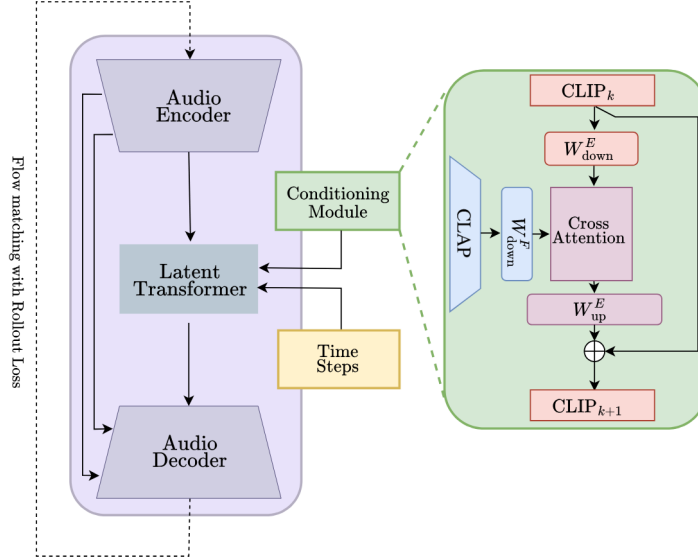


Fig. 3: VisAH-FM builds on VisAH [13] (purple), adding timestep conditioning, rollout loss, and an audio-aware cross-attention adapter inside CLIP (green).

the audio-aware adapted features using learnable weights and pass this onto the next layer $k + 1$. Formally, we define the output of the adapter layer as:

$$\text{adapter}(F_k, E, A) = F_k + \lambda_E m(F_k, E) \quad (7)$$

where λ_E is a learnable scalar controlling the contribution of the audio-aware features E . The overall conditioning mechanism is illustrated in Figure 3 (green box). Note that this conditioning can be applied to incorporate text features instead of, or along with, the audio. Importantly, initializing λ_E to zero ensures that the first forward pass matches the pretrained CLIP encoder, allowing us to fully leverage pretrained weights while gradually incorporating cross-modal cues

during training. In practice, we use E , the audio features of a CLAP [48] encoder as we expect that its alignment with text would simplify the incorporation to CLIP features.

Time (CFM) Conditioning. We encode the flow matching timestep using sinusoidal embeddings, relying on the positional encoding scheme as adopted for diffusion timesteps in [12], and append this embedding to the conditioning tokens provided to the latent transformer block via cross-attention. We ablate this choice in Appendix.

3 Experiments

Implementation Details. To optimally leverage the pretrained VisAH model, which was trained to directly predict well-balanced audio rather than a vector field, we systematically incorporate the source sample (x_0) into the model output to define the estimated vector field as $v_\theta(x_t, t, c) = x_0 - u_\theta(x_t, t, c)$, where $u_\theta(x_t, t, c)$ denotes the actual model. We trained the model for 50 epochs ($\sim 23,500$ iterations) using a learning rate of 1×10^{-4} , a cosine annealing scheduler, and a batch size of 32.

Following VisAH, the input waveform was sampled at 44.1 kHz and converted to mono by averaging the stereo channels. We observed an issue with audio saving quality, and therefore resampled the test set using the official code from [13]. We also trained another model on the exact same dataset as [13], without on-the-fly remixing, and present the results in Appendix. Similar to [3], we discretized time into four timesteps, sampled them uniformly during training, and used fixed-step Euler integration at inference. Although our model performs several recurrent passes of the vector-field estimator at inference, the resulting overhead is minor. Most computation is dominated by the conditioning encoders — CLIP (~ 300 M parameters), CLAP (~ 31 M), and optionally T5-XXL (11B) or InternVL (8B) - which are run once per video segment, whereas the repeated U-Net vector-field estimator is comparatively lightweight (~ 60 M). The adapter module is incorporated at the 18th layer of the CLIP model, and training-time VRAM measurements are reported in Appendix. Except for the changes mentioned in Sec. 2, we used the exact VisAH architecture. All experiments were performed on the Muddy Mix Dataset. Additional details are given in Appendix.

Metrics. We evaluate our models using the same metrics as VisAH, along with the introduction of an additional remixing metric:

- **Signal metrics:** We measure the magnitude distance (Mag) [49], which evaluates audio quality in the time–frequency domain, and envelope distance (Env) [22], which assesses quality in the time domain. We also measure temporal alignment using the Wasserstein distance (Was).

- **Semantic alignment:** We compute the KL divergence (KLD) between the predicted logits of a PASST [18] model for the predicted and ground-truth audio. Additionally, we report the ImageBind score (IB Score) [11] to quantify the similarity between predicted audio and image frames. Unlike VisAH, we report the IB Score directly instead of ΔIB , as we find it more interpretable.
- **Remixing metric:** Since the goal is to adjust the relative loudness across three categories—human speech, music, and sound effects—we employ a sound source separation model [45] to extract stems from both the predicted and ground-truth audio, and measure the loudness difference (LDif) between corresponding sources as $\frac{1}{K} \sum_k |\text{loud}(s_k(\hat{x})) - \text{loud}(s_k(x))|$, where $\hat{x}$ and x denote the predicted and ground-truth audio, respectively, $\text{loud}(\cdot)$ computes the loudness, and $s_k(\cdot)$ the separator of the k -th source from the audio.

3.1 Quantitative Analysis

Table 1: Main results. Best results are highlighted in **bold**, and all values except the LDif are multiplied by 100.

Model	Conditioning	IB Score $\uparrow$	KLD $\downarrow$	LDif $\downarrow$	Mag $\downarrow$	Env $\downarrow$	Was $\downarrow$
Input	-	28.14	20.74	18.36	22.69	6.29	1.96
VisAH	CLIP	28.84	11.37	9.66	9.99	3.38	0.84
VisAH	T5	28.92	11.71	9.63	10.22	3.44	0.88
VisAH	CLIP-CLAP	28.82	11.28	9.60	10.27	3.56	0.80
VisAH-FM (Ours)	CLIP-CLAP	29.12	9.70	7.77	8.28	2.74	0.63

Main Results. Table 1 compares our proposed model VisAH-FM against the VisAH baseline using image (CLIP), text (T5-encoded), or our audio-adapted CLIP-CLAP conditioning. Adding the adapter conditioning to the discriminative VisAH baseline slightly improves performance, showing that the conditioning module is independently useful. However, VisAH-FM still clearly outperforms all VisAH variants, indicating that the gains come from both the improved conditioning and the flow matching formulation.

Ablation on Conditioning Module. Table 2 presents an ablation study of the different conditioning modalities and methods for the flow matching model. Without adapters, vision-based conditioning performs slightly better than text-based conditioning. Incorporating additional modalities into the image encoder consistently improves performance across all metrics. Notably, adding audio features encoded with CLAP [48] yields larger gains than adding textual features. We attribute this improvement to the stronger audio–visual interaction enabled by the conditioning module: the module determines which source to enhance, allowing the velocity-field estimator to focus only on regression. As text captions originate

Table 2: Conditioning ablation. We compare different conditioning methods (with and without adapter) and different modalities: audio (A), vision (V), and text (T).

Adapter	Modality	IB Score $\uparrow$	KLD $\downarrow$	LDif $\downarrow$
✓	V+A	29.12	9.70	7.77
✓	V + T	29.08	9.76	7.85
✓	V+T+A	29.10	9.73	7.90
✗	V	29.09	9.79	7.87
✗	T	28.88	10.50	8.13

from a VLM, they mainly contain information similar to image features (i.e., visual scene semantics), whereas audio features encode the auditory scene. Since the target source typically corresponds to elements common to both the visual and auditory domains, integrating audio features inside the conditioning proves more beneficial than adding the textual counterpart. Additionally, CLAP extracted audio representations are complementary to those from the DEMUCS encoder, further enriching the model’s understanding of the auditory scene. Interestingly, incorporating both textual and audio features performs slightly worse than using only visual and audio modalities, indicating that textual features are not needed when the audio is directly incorporated into the image encoder. This finding holds practical importance for real-world applications as the extraction of text features is very costly, requiring a forward pass through a VLM [4] and a large text encoder [33].

Ablation of the Rollout Loss. Table 3 ablates the rollout loss and compares it with several standard methods designed to improve consistency across the flow matching steps. The addition of the rollout loss proves crucial: removing it leads to a marked performance drop *i.e.* approximately +1.2 in KLD and +1.6 in loudness difference.

The **consistency loss** encourages the vector field predicted at time t to remain consistent with the one predicted at a slightly later time $t + \Delta$, when the input is advanced along the flow using the earlier prediction:

$$\mathcal{L}_{\text{cons}} = \|v_{\theta}(x_t, t) - \text{StopGrad}(v_{\theta}(x_t - \Delta_t v_{\theta}(x_t, t), t + \Delta_t))\| \quad (8)$$

Interestingly, adding this loss substantially degrades performance. We hypothesize that this occurs because if the first estimate is noisy, the propagated estimate becomes even less accurate, effectively guiding the model in an incorrect direction. In contrast, the rollout loss anchors predictions to the ground-truth trajectory, preventing such drift and stabilizing training.

Recent works have shown that injecting Gaussian noise into the input trajectories can improve robustness to out-of-distribution trajectories at inference time [3, 25]. The row “*Bridge Matching*” reports the results of a model trained with this strategy (using Gaussian noise centered on zero with standard deviation of $1e-5$). While it slightly outperforms the standard flow matching baseline, it

Table 3: Rollout component ablation.

Objective	IB Score $\uparrow$	KLD $\downarrow$	LDif $\downarrow$
FM + Rollout	29.09	9.79	7.87
FM	28.92	10.99	9.48
Rollout	28.94	9.92	7.71
FM + Consistency	28.40	14.58	10.30
Bridge Matching	29.04	10.85	9.62
FM Weighted	28.86	11.16	9.44

Table 4: Hyperparameter sensitivity.

λ	Horizon	Steps	IB Score $\uparrow$	KLD $\downarrow$	LDif $\downarrow$
0.1	4	4	29.12	9.85	8.02
0.3	4	4	29.09	9.79	7.87
0.5	4	4	29.12	9.93	7.80
1.0	4	4	29.04	9.82	7.84
0.1	2	4	29.08	9.68	7.94
0.3	4	10	29.10	9.79	8.35

remains notably inferior to the version trained with rollout loss. In VAE-based setups, the latent space is assumed Gaussian, hence, added noise can mimic realistic errors. However, in our case, the input space corresponds to raw audio, and the injected noise effectively acts as white noise—significantly different from the model’s own prediction errors. Consequently, these perturbations fail to simulate realistic inference conditions, and robustness improvements remain limited. In contrast, the rollout loss directly reuses the model outputs as inputs, accurately reproducing the inference scenario and yielding significantly better performance.

The row “*FM Weighted*” corresponds to a flow matching model trained with a loss weighted inversely proportional to the time variable, emphasizing samples closer to the input distribution. This baseline performs slightly worse than standard flow matching, indicating that simply prioritizing early steps is insufficient. Instead, enabling the model to iteratively correct its errors across time proves essential—highlighting the effectiveness of the rollout loss. The row “*Rollout*” corresponds to a model trained without the standard flow matching loss, relying solely on the rollout loss. Remarkably, this variant noticeably outperforms vanilla flow matching and only slightly underperforms the full model that combines flow matching and rollout loss. This highlights the central role of the rollout mechanism: even without the intermediate supervision and linearity prior imposed by flow matching, the model achieves strong performance. Moreover, this variant significantly surpasses the original VisAH model, suggesting that allowing the model to repeatedly refine its predictions through multiple forward passes over the same architecture yields meaningful improvements. This iterative refinement behavior is evocative of phenomena observed in large language models, where multi-step reasoning improves output quality (e.g., chain-of-thought prompting [44, 46]). We further analyze the behavior of this model and its learned trajectories in Section 3.1.

Hyperparameter Sensitivity. Table 4 reports the sensitivity of our method to hyperparameters of the rollout loss. The first block of the table varies the rollout coefficient λ (Eq. 5). Performance remains stable across a broad range, highlighting the robustness of the method to this hyperparameter. We also vary the rollout horizon, defined as the number of steps before applying the rollout loss. When the horizon is shorter than the full inference trajectory, we randomly choose the starting step to ensure coverage of different temporal segments; otherwise, rollout begins from the first step. The last two rows correspond to settings where the

rollout horizon is smaller than the total number of inference steps. We observe no significant performance degradation, suggesting that rollout over only a subset of steps is sufficient to teach the model to recover from its own intermediate errors.

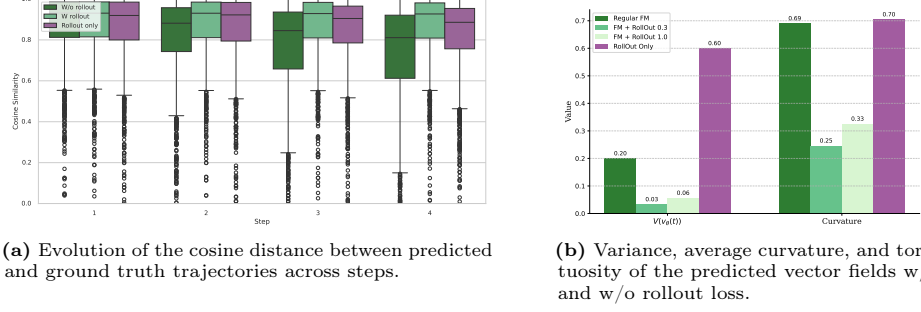


Fig. 4: Impact of rollout loss: with flow matching it stabilizes trajectories and limits error accumulation; alone it learns nonlinear paths.

Model Behavior. Here we analyze how the rollout loss shapes the learned trajectories, and how these differ from those obtained with standard flow matching.

Error Accumulation Analysis. Figure 4a reports the cosine similarity between the predicted and ground-truth vector fields over inference steps on the test set. Consistent with our hypothesis, the model trained without rollout loss progressively diverges from the ground-truth trajectory: initial errors compound and drive the model toward out-of-distribution regions, leading to worsening predictions over time. In contrast, the model trained with both rollout and flow matching losses maintains high alignment throughout the trajectory, demonstrating that the rollout objective effectively constrains error accumulation and stabilizes long-range trajectories. Interestingly, the model trained with rollout loss alone exhibits small drift during the last steps, although it performs well overall (see Table 3), indicating that the learned trajectories differ from the linear one imposed in flow matching.

Trajectory Linearity. Next, we study how the rollout loss affects the geometry of the learned trajectories. In flow matching, the optimal trajectory is theoretically linear in the data space; however, finite-capacity models and error accumulation during inference can induce nonlinear deviations. To quantify this effect, we compute both the variance of the predicted vector field across inference steps and the average discrete curvature of the trajectory on the test set. Formally, we define the discrete curvature at step t as

$$\Theta_t = \arccos\left(\frac{\hat{v}_{t+1} \cdot \hat{v}_t}{\|\hat{v}_{t+1}\| \|\hat{v}_t\| + \epsilon}\right), \quad (9)$$

where $\hat{v}_t$ denotes the predicted vector field at time t . Θ_t measures the angular deviation between successive vector field directions; higher values indicate stronger trajectory bending. Similarly, higher variance across steps reflects instability in the predicted vector fields. Figure 4b reports average curvature and variance for models trained with (i) flow matching only, (ii) flow matching with rollout loss ($\lambda = 0.3$ and 1.0), (iii) rollout loss only. The flow matching-only model exhibits strong nonlinear deviations after the first step, consistent with error compounding and off-manifold drift. Introducing the rollout loss markedly stabilizes the trajectory, yielding substantially lower curvature and variance. Interestingly, when the rollout term increases, trajectories become less linear again. Consequently, when using rollout loss alone, the model learns highly non-linear trajectories, explaining the behavior observed in Figure 4a: predicted vector fields are not aligned with the ground-truth, as the model takes a completely different path to reach the target.

3.2 Qualitative Analysis

Visualization of the Rollout Loss. Figure 5 shows outputs across flow-matching steps. The rollout-trained model reliably enhances speech (red boxes), whereas the regular flow model does not. Since each step remains realistic, the step count can act as a knob for highlighting strength. Additionally, as the flow maps poorly-balanced audio to well-balanced audio (as opposed to starting from noise), the output of each step is incrementally transformed realistic audio. This unlocks the possibility for a user to use the number of steps as a knob to specify the degree of desired highlighting.

Visualization of VisAH-FM against VisAH. Figure 6 shows two cases where VisAH-FM does consistently better than VisAH when attenuating overly loud sources. Occasional artifacts in VisAH-FM output leave room for further improvement. More visualizations are available in the Appendix. Demo examples are provided in the supplements.

Subjective Test. We conduct a subjective test to compare VisAH and VisAH-FM. Specifically, we ask participants to “choose the audio that best aligns with the visual scene in terms of loudness balance and overall quality”. Thirteen participants evaluated the models on a set of nine Muddy-Mix and MovieGen [31] generated videos. VisAH-FM achieves a $45.55 \pm 9.46\%$ win rate versus $17.17 \pm 6.17\%$ for VisAH, with the remaining 37.28% of comparisons judged as draws. These results underscore our method’s effectiveness.

4 Related work

Audio Remixing. Highlighting a mixed audio signal can be viewed as rebalancing its constituent sources—effectively translating one mixing style into another. Prior work in music production has examined this extensively [17, 28, 34, 43], exploring both traditional and creative strategies for shaping a track’s emotional

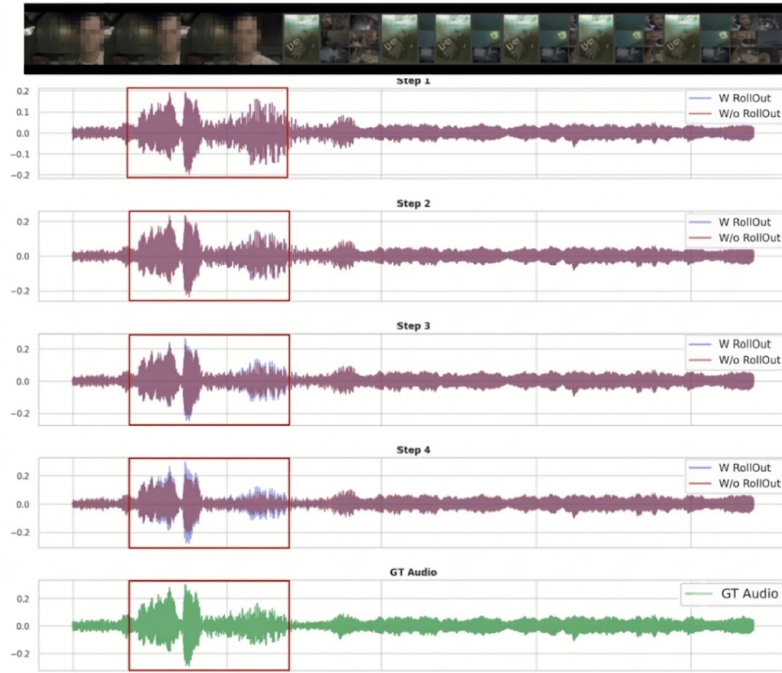


Fig. 5: Illustration of the behavior of the model trained with and without rollout loss. Wave-forms are overlapped for readability. The first step is similar, but the difference appears in the latter steps when the rollout-trained model enhances speech, while the other does not.

and acoustic character. Reproducing a target mix often requires controlling source energy and applying audio effects to achieve stylistic consistency, whether through expert-driven pipelines [37] or modern learning-based approaches [28, 40]. While most existing efforts emphasize music and instrument stems, these techniques do not generalize naturally to the diverse and highly dynamic sound compositions in real-world media. In contrast, VisAH [13] extends the mixing paradigm beyond music to speech and cinematic sound effects, and leverages visual cues from video to guide source highlighting. While effective, this method relies on a discriminative mapping from input to output mixes, implicitly assuming a one-to-one correspondence between degraded audio mixture and professionally mixed audio—an assumption that may not hold in ambiguous and diverse audio contexts like movies. Concurrently, SemMix [16] studies LVLM-derived semantic prompts for visually-guided acoustic highlighting, which is complementary to our focus on flow matching.

Visually-conditioned Audio Generation. With the rapid progress in audio synthesis, several works have explored conditioning audio generation on video frames [5, 31, 41]. This setting is challenging due to the tight temporal cou-

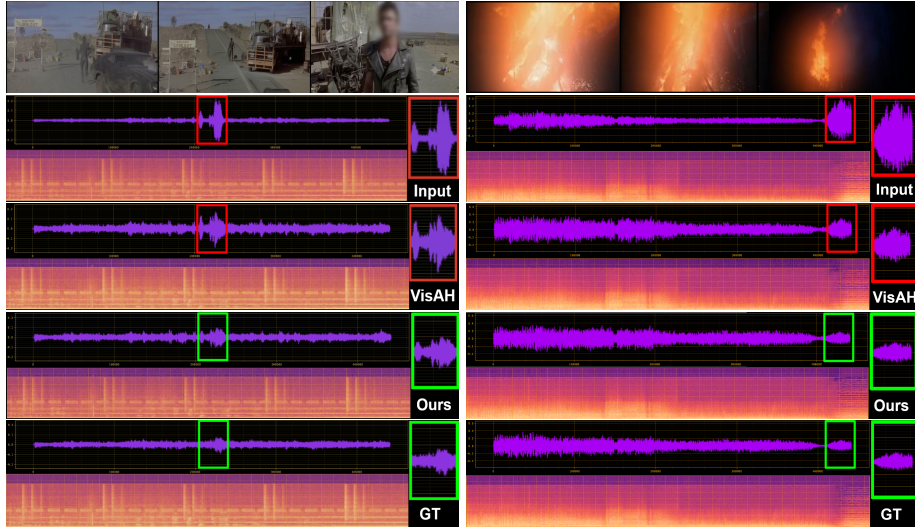


Fig. 6: Qualitative comparison Ours (VisAH-FM) vs VisAH. We achieve better balance by appropriately attenuating overly loud sources (SFX Column 1, Music Column 2).

pling between visual events and sounds, as well as the high diversity of real-world audio scenes. Current methods typically specialize in either speech or non-speech sounds (e.g., music, sound effects), limiting their generality. Consequently, these models are not directly suited for acoustic highlighting, which requires jointly modeling speech, music, and effects.

Audio-Visual Source Separation. Early works in audio-visual source separation [9, 10, 50] demonstrated that visual cues provide powerful guidance for isolating sound sources in videos. More recent studies [14, 15, 19] tackle this task using generative modeling, such as diffusion-based and flow-based frameworks, benefiting from their strong ability to model complex multimodal distributions and generate coherent audio conditioned on visual context. Our work builds upon these trends by applying conditional flow matching to the audio remixing problem, focusing specifically on maintaining trajectory stability via a rollout loss and enabling early cross-modal fusion through adapter layers, which distinguishes our approach from prior diffusion-based conditioning pipelines.

5 Conclusion

We showed that generative modeling offers a more effective solution than discriminative approaches for visually informed acoustic highlighting. With a rollout loss and cross-modal conditioning module, our flow matching model outperforms prior works on the Muddy Mix dataset.

References

1. Ayilo, J.E., Sadeghi, M., Serizel, R., Alameda-Pineda, X.: Diffusion-based unsupervised audio-visual speech enhancement. In: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2025)
2. Ben-Hamu, H., Puny, O., Gat, I., Karrer, B., Singer, U., Lipman, Y.: D-flow: Differentiating through flows for controlled generation. arXiv preprint arXiv:2402.14017 (2024)
3. Chadebec, C., Tasar, O., Sreetharan, S., Aubin, B.: Lbm: Latent bridge matching for fast image-to-image translation. arXiv preprint arXiv:2503.07535 (2025)
4. Chen, Z., Wang, W., Tian, H., Ye, S., Gao, Z., Cui, E., Tong, W., Hu, K., Luo, J., Ma, Z., et al.: How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences* **67**(12), 220101 (2024)
5. Cheng, H.K., Ishii, M., Hayakawa, A., Shibuya, T., Schwing, A., Mitsufuji, Y.: Mmaudio: Taming multimodal joint training for high-quality video-to-audio synthesis. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28901–28911 (2025)
6. Cuttano, C., Trivigno, G., Rosi, G., Masone, C., Averta, G.: Samwise: Infusing wisdom in sam2 for text-driven video segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3395–3405 (2025)
7. Défossez, A.: Hybrid spectrogram and waveform source separation. In: Proceedings of the ISMIR Workshop on Music Source Separation (2021)
8. Ebrahimi, S., Arik, S.O., Nama, T., Pfister, T.: Crome: cross-modal adapters for efficient multimodal llm. arXiv preprint arXiv:2408.06610 (2024)
9. Ephrat, A., Mosseri, I., Lang, O., Dekel, T., Wilson, K., Hassidim, A., Freeman, W.T., Rubinstein, M.: Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. arXiv preprint arXiv:1804.03619 (2018)
10. Gao, R., Grauman, K.: Visualvoice: Audio-visual speech separation with cross-modal consistency. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 15490–15500. IEEE (2021)
11. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15180–15190 (2023)
12. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. *Advances in neural information processing systems* pp. 6840–6851 (2020)
13. Huang, C., Gao, R., Tsang, J., Kurcius, J., Bilén, C., Xu, C., Kumar, A., Parekh, S.: Learning to highlight audio by watching movies. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23925–23935 (2025)
14. Huang, C., Liang, S., Tian, Y., Kumar, A., Xu, C.: Davis: High-quality audio-visual separation with generative diffusion models (2023)
15. Huang, C., Liang, S., Tian, Y., Kumar, A., Xu, C.: High-quality sound separation across diverse categories via visually-guided generative modeling. arXiv preprint arXiv:2509.22063 (2025)
16. Huang, J., Huang, C., Xu, C.: Semantic visually-guided acoustic highlighting with large vision-language models. arXiv preprint arXiv:2601.08871 (2026)
17. Koo, J., Martínez-Ramírez, M.A., Liao, W.H., Uhlich, S., Lee, K., Mitsufuji, Y.: Music mixing style transfer: A contrastive learning approach to disentangle audio effects. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)

18. Koutini, K., Schlüter, J., Eghbal-Zadeh, H., Widmer, G.: Efficient training of audio transformers with patchout. *arXiv preprint arXiv:2110.05069* (2021)
19. Lee, S., Jung, C., Jang, Y., Kim, J., Chung, J.S.: Seeing through the conversation: Audio-visual speech separation based on diffusion model. In: *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 12632–12636. IEEE (2024)
20. Lei, J., Berg, T.L., Bansal, M.: Detecting moments and highlights in videos via natural language queries. *Advances in Neural Information Processing Systems* **34**, 11846–11858 (2021)
21. Lemerrier, J.M., Richter, J., Welker, S., Gerkmann, T.: Storm: A diffusion-based stochastic regeneration model for speech enhancement and dereverberation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **31**, 2724–2737 (2023)
22. Liang, S., Huang, C., Tian, Y., Kumar, A., Xu, C.: Av-nerf: Learning neural fields for real-world audio-visual scene synthesis. *Advances in Neural Information Processing Systems* **36**, 37472–37490 (2023)
23. Lin, K.Q., Zhang, P., Chen, J., Pramanick, S., Gao, D., Wang, A.J., Yan, R., Shou, M.Z.: Univtg: Towards unified video-language temporal grounding. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 2794–2804 (2023)
24. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. *ICLR* (2022)
25. Liu, X., Wu, L., Ye, M., Liu, Q.: Let us build bridges: Understanding and extending diffusion generative models. *arXiv preprint arXiv:2208.14699* (2022)
26. Liu, Y., Li, S., Wu, Y., Chen, C.W., Shan, Y., Qie, X.: Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 3042–3051 (2022)
27. Mariani, G., Tallini, I., Postolache, E., Mancusi, M., Cosmo, L., Rodolà, E.: Multi-source diffusion models for simultaneous music generation and separation. *arXiv preprint arXiv:2302.02257* (2023)
28. Martínez Ramírez, M.A., Benetos, E., Reiss, J.D.: Deep learning for black-box modeling of audio effects. *Applied Sciences* **10**(2), 638 (2020)
29. Ng, A., Jordan, M.: On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems* **14** (2001)
30. Ning, M., Li, M., Su, J., Salah, A.A., Ertugrul, I.O.: Elucidating the exposure bias in diffusion models. *arXiv preprint arXiv:2308.15321* (2023)
31. Polyak, A., Zohar, A., Brown, A., Tjandra, A., Sinha, A., Lee, A., Vyas, A., Shi, B., Ma, C.Y., Chuang, C.Y., et al.: Movie gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720* (2024)
32. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: *International conference on machine learning*. pp. 8748–8763. PmLR (2021)
33. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research* **21**(140), 1–67 (2020)
34. Ramírez, M.A.M., Reiss, J.D.: Modeling nonlinear audio effects with end-to-end deep neural networks. In: *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pp. 171–175. IEEE (2019)

35. Ranzato, M., Chopra, S., Auli, M., Zaremba, W.: Sequence level training with recurrent neural networks. In: ICLR (2016)
36. Scheibler, R., Hershey, J.R., Doucet, A., Li, H.: Source separation by flow matching. arXiv preprint arXiv:2505.16119 (2025)
37. Schlecht, S., Parker, J., Schäfer, M., Rabenstein, R.: Physical modeling using recurrent neural networks with fast convolutional layers. In: International Conference on Digital Audio Effects. pp. 138–145. DAFx (2022)
38. Song, Y., Dhariwal, P.: Improved techniques for training consistency models. ICLR (2024)
39. Song, Y., Dhariwal, P., Chen, M., Sutskever, I.: Consistency models. ICLR (2023)
40. Steinmetz, C.J., Bryan, N.J., Reiss, J.D.: Style transfer of audio effects with differentiable signal processing. arXiv preprint arXiv:2207.08759 (2022)
41. Tian, Z., Jin, Y., Liu, Z., Yuan, R., Tan, X., Chen, Q., Xue, W., Guo, Y.: Audiox: Diffusion transformer for anything-to-audio generation. arXiv preprint arXiv:2503.10522 (2025)
42. Tong, A., Fatras, K., Malkin, N., Huguet, G., Zhang, Y., Rector-Brooks, J., Wolf, G., Bengio, Y.: Improving and generalizing flow-based generative models with minibatch optimal transport. arXiv preprint arXiv:2302.00482 (2023)
43. Vanka, S.S., Steinmetz, C., Rolland, J.B., Reiss, J., Fazekas, G.: Diff-mst: Differentiable mixing style transfer. arXiv preprint arXiv:2407.08889 (2024)
44. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D.: Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171 (2022)
45. Watcharasupat, K.N., Wu, C.W., Ding, Y., Orife, I., Hipple, A.J., Williams, P.A., Kramer, S., Lerch, A., Wolcott, W.: A generalized bandsplit neural network for cinematic audio source separation. IEEE Open Journal of Signal Processing **5**, 73–81 (2023)
46. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems **35**, 24824–24837 (2022)
47. Welker, S., Le, M., Chen, R.T., Hsu, W.N., Gerkmann, T., Richard, A., Wu, Y.C.: Flowdec: A flow-based full-band general audio codec with high perceptual quality. arXiv preprint arXiv:2503.01485 (2025)
48. Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., Dubnov, S.: Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
49. Xu, X., Zhou, H., Liu, Z., Dai, B., Wang, X., Lin, D.: Visually informed binaural audio generation without binaural audios. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15485–15494 (2021)
50. Zhao, H., Gan, C., Rouditchenko, A., Vondrick, C., McDermott, J., Torralba, A.: The sound of pixels. In: Proceedings of the European conference on computer vision (ECCV). pp. 570–586 (2018)