VAINPAINT: ZERO-SHOT VIDEO-AUDIO INPAINTING FRAMEWORK WITH LLMS-DRIVEN MODULE

Kam Man Wu   *Zeyue Tian*   *Liya Ji*   *Qifeng Chen*[†]

The Hong Kong University of Science and Technology

ABSTRACT

Video and audio inpainting for mixed audio-visual content has become a crucial task in multimedia editing recently. However, precisely removing an object and its corresponding audio from a video without affecting the rest of the scene remains a significant challenge. To address this, we propose VAInpaint, a novel pipeline that first utilizes a segmentation model to generate masks and guide a video inpainting model in removing objects. At the same time, an LLM then analyzes the scene globally, while a region-specific model provides localized descriptions. Both the overall and regional descriptions will be inputted into an LLM, which will refine the content and turn it into text queries for our text-driven audio separation model. Our audio separation model is fine-tuned on a customized dataset comprising segmented MUSIC instrument images and VGGSound backgrounds to enhance its generalization performance. Experiments show that our method achieves performance comparable to current benchmarks in both audio and video inpainting.

Index Terms— Audio-Video Inpainting, LLMs-driven alignment, LLMs-driven content refinement

1. INTRODUCTION

In the era of multimodal content creation, the ability to separate and edit mixed audio-visual (AV) elements is a crucial part of video editing. Traditional methods in AV combination content are often limited to the handling method within the speech domain or struggle with tangled audio-visual signals in static scenes [1]. Occasionally, when performing the visual-audio separation task in the outdoor scenes, issues with unwanted leftover audio or visual elements in the output can also lower perceived quality [2]. In this paper, we define a machine learning task as follows: Given a video with multiple objects, such as a scene where people are playing instruments, separate the instrument audio from the mixed soundtrack, and inpaint the video to remove the instrument and performer in a zero-shot manner. This paper addresses the challenge of isolating clean video and audio tracks from mixed content, using advanced segmentation, inpainting, and language-guided techniques.

Prior works have constructed important foundations for the components of audio-visual processing. However, they are usually limited by their focus on visual features or audio features individually, which have some potential alignments and connections to other modalities. In video inpainting, efforts have focused on temporal consistency and efficiency [3], with recent models like ProPainter [4] advancing transformer-based, mask-guided object removal. For segmentation, foundational models such as SAM2 [5] enable precise, prompt-driven object isolation. On the audio side, OmniSep, a query-based audio separation model [6], supports omni-modal inputs, including text. Meanwhile, LLM-guided methods [7] have emerged to align representations using human-like language. AV separation or inpainting pipelines are often built based on these techniques, using segmentation for initial source localization, as seen in The Sound-Of-Pixels [8], and SAVE [9]. However, a key limitation is that these methods often lack robust LLM alignment for high-level reasoning, leading to imprecise segmentation or correspondence and an inability to fully remove object-related sounds. Luckily, a new object-to-text pipeline named Describe Anything [10] provides valuable localized captions, which indicates the potential of the model for enabling precise audio separation and inpainting guidance.

As a result, to address the problem of the critical disconnection between high-precision visual inpainting and semantically accurate audio removal, we introduce VAInpaint, which unifies these elements for end-to-end AV inpainting. We first utilize SAM2 to segment out our needed object, then we generate a corresponding object mask to guide ProPainter in removing undesired visual elements. Next, we utilize an LLM to analyze the extracted frame for overall image understanding, while using Describe Anything to generate region-specific text descriptions using the previously generated masks. Combining the above outputs will lead to an LLM-refined text query for OmniSep as an input. For our training data, we developed a custom dataset generation pipeline: We first use SAM2 to segment instruments and performers from MUSIC videos [8], creating masks at any resolution. VGGSound videos are then resized to match the resolution, and we mix the segmented MUSIC elements into VGGSound backgrounds [11]. This results in a dataset of mixed audio-visual samples. To increase the separation effect, we fine-tune our audio separation model using our

customized dataset. To conclude, our paper's main contributions are:

- An integrated hybrid workflow combining video inpainting, LLMs-based comprehensive and regional Scene Understanding with text, and text-query based audio separation.
- A new dataset blending MUSIC[8] and VGGSound[11] with audio-visual content, supported by automated scripts to ensure scalability.
- A pipeline that assists LLMs in more effectively extracting and condensing content.

2. METHODOLOGY

2.1. Preliminaries

OmniSep performs audio source separation in the spectrogram domain [6]. For a mixture spectrogram $m \in \mathbf{R}^{B \times 1 \times F \times T}$ and N source embeddings $\{e_n\}_{n=1}^N \in \mathbf{R}^{B \times D}$ from multimodal inputs, the model predicts separation masks $\{p_n\}_{n=1}^N$ via:

$$
\begin{aligned}
f_s &= \text{U-Net}(\log(m + \epsilon)) \\
f_e^n &= \sigma(\mathbf{W}_e e_n + \mathbf{b}_e) \\
p_n &= \sigma(\langle \mathbf{s} \odot f_e^n, f_s \rangle + b)
\end{aligned}
\tag{1}
$$

where U-Net is an encoder-decoder network, $\mathbf{W}_e \in \mathbf{R}^{C \times D}$, $\mathbf{b}_e \in \mathbf{R}^C$, $\mathbf{s} \in \mathbf{R}^C$, $b \in \mathbf{R}$ are learnable parameters, σ is sigmoid activation, and $\langle \cdot, \cdot \rangle$ denotes inner product.

The training process minimizes the weighted BCE loss:

$$
\mathcal{L} = \frac{1}{N} \sum_{n=1}^N \mathbf{E}_{f,t} \left[\max(\log(1+m), 10^{-3}) \cdot \text{BCE}(p_n, t_n) \right]
\tag{2}
$$

with optional log-frequency warping. Separated sources are obtained as $\hat{s}_n = p_n \odot m$.
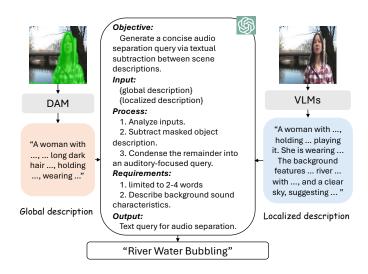
2.2. LLMs-Generated Text Queries

To convert visual-language descriptions into audio-language queries, we employ a two-step LLM process. First, a VLM analyzes an extracted frame for overall scene understanding: $d_v = \text{VLM}(f)$, where f is the extracted frame. Simultaneously, Describe Anything generates object-localized descriptions for the masked MUSIC region: $d_a = \text{DescribeAnything}(f, m)$, with mask m from SAM2. The LLM then performs textual subtraction and condensation:

$$
q = \text{LLM}(d_v - d_a)
\tag{3}
$$

condensing the difference between two descriptions into an audio-focused query. As we can see from the figure 2, this text query guides OmniSep, which is our sound separator, for separation tasks. Our workflow transforms visual cues into auditory prompts, enabling the precise isolation of VGGSound

audio [11]. For the details of the text query generation, we can refer to the content inside Figure 1. Inside, we describe how we utilize the Describe Anything Model (DAM) and VLM to generate regional descriptions and overall descriptions. Then we feed the descriptions into our used LLMs and generate a text query as requested.

Fig. 1: Pipeline of text query generation with LLMs.

2.3. Model Fine-Tuning

We fine-tune the OmniSep model [6] on our VAInpaint dataset using a supervised learning method. The model processes mixture spectrograms m and multimodal embeddings $\{e_n\}_{n=1}^N$ (extracted using ImageBind [12]) to predict separation masks $\{p_n\}_{n=1}^N$ through the architecture described in Section 2.1.

The model is optimized using the weighted binary cross-entropy loss:

$$
\mathcal{L} = \frac{1}{N} \sum_{n=1}^N \mathbf{E}_{f,t} \left[\max(\log(1+m), 10^{-3}) \cdot \text{BCE}(p_n, g_n) \right]
\tag{4}
$$

where g_n denotes the ground-truth mask for source n.
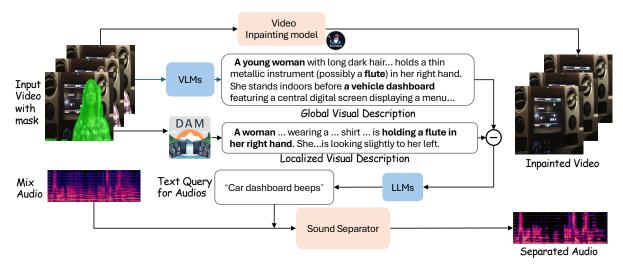
3. EXPERIMENTS

3.1. Objective Metrics

We evaluate our model using the following objective metrics:

3.1.1. Fréchet Distance (FD)

FD quantifies the similarity between two multivariate Gaussian distributions fitted to feature embeddings [13, 14]. The metric computes:

$$
\mathcal{FD} = \|\mu_g - \mu_t\|^2 + \text{Tr}\left(\Sigma_g + \Sigma_t - 2(\Sigma_g \Sigma_t)^{1/2}\right),
\tag{5}
$$

Fig. 2: **Overview of our video-audio inpainting pipeline**. Taking the input video with the removed region and the corresponding mixed audio as the inputs, our method can correctly generate the inpainted video and separate the audio. We use the text modality to fill the domain gap between videos and audio. We found that LLMs can effectively transfer language from the domain of Visual-language models to that of text-audio models.

where (μ_g, Σ_g) and (μ_t, Σ_t) represent the mean and covariance of embeddings from generated and target features, respectively. Lower FD indicates better distribution alignment.

3.1.2. Kullback-Leibler Divergence (KLD)

KLD measures the distributional discrepancy between classifier outputs for separated and target audio using the binary formulation [15, 16]:

$$\mathcal{D}_{KL}(P\|Q) = \sum_i \left[P_i \log \frac{P_i}{Q_i} + (1 - P_i) \log \frac{1 - P_i}{1 - Q_i} \right],$$
(6)

where P represents target class probabilities and Q represents generated audio probabilities from sigmoid-activated outputs. Lower KLD values indicate better distribution matching.

3.1.3. Scale-Invariant Signal-to-Distortion Ratio (SI-SDR)

SI-SDR evaluates separation quality while remaining invariant to amplitude scaling [17]:

$$\text{SI-SDR} = 10 \log_{10} \frac{\|\alpha \mathbf{s}_t\|^2}{\|\mathbf{e}\|^2},$$
(7)

$$\alpha = \frac{\mathbf{s}_g^\top \mathbf{s}_t}{\|\mathbf{s}_t\|^2}, \quad \mathbf{e} = \mathbf{s}_g - \alpha \mathbf{s}_t,$$
(8)

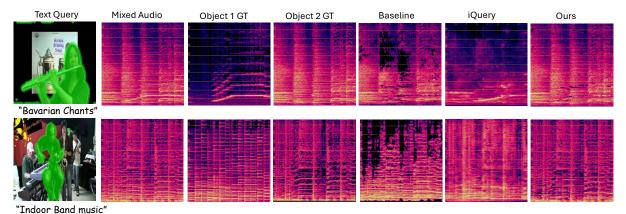where $\mathbf{s}_g$ is the separated signal and $\mathbf{s}_t$ is the target reference. Higher SI-SDR indicates superior separation performance.

3.2. Experimental Results

Method	Query Type	FD ↓	KID ↓	SI-SDR ↑	SDR ↑
Sound-of-Pixels [8]	Visual	95.32	11.16	-24.48	-1.66
iQuery [18]	Visual	47.95	6.16	-54.92	-2.98
Ours (VGGSound Label)*	Text	24.04	2.54	7.10	5.31
Ours (LLMs-Query)	Text	34.12	5.47	-3.66	2.86
Ours (LLMs-Query)*	**Text**	**25.77**	**3.50**	**4.07**	**4.59**

Table 1: Quantitative results on audio separation. Our LLMs text query approach outperforms visual query baselines, with further improvements after finetuning the model on our designed dataset. * indicates the usage of the fine-tuned models.

We evaluate our proposed OmniSep-LLMs method against the iQuery baseline [18] and the Sound-of-Pixels baseline [8] on our custom dataset. Our experimental results show that iQuery consistently fails to reproduce the correct audio pitch, producing outputs that are consistently lower than the ground truth. In contrast, our text-query pipeline accurately generates the pitch of separated audio sources. Meanwhile, the Sound-of-Pixels method frequently produces audio with significant data loss, which severely degrades both output audio quality and audience listening experience. Also, from the Table 1 we can see that although the original OmniSep checkpoint substantially outperforms these visual query baselines, it remains limited by its reliance on precise text descriptions and struggles to distinguish between sources from the same category (e.g., differentiating between two instruments).

To improve generalization, we fine-tuned the model on our custom dataset. As shown in Table 1, finetuning results in substantial gains across all metrics. Qualitatively, the separation is significantly improved, with reduced artifacts and clearer output, as illustrated in Figure 3. Although directly using VGGSound labels is infeasible in real-life applications

Fig. 3: **Qualitative results for our methods.** The spectrum indicates that our text-query pipeline (Original and Fine-tuned) produces cleaner source separation with fewer residual artifacts than the iQuery visual-query baseline.

[11], the proximity of our results to those obtained with ideal VGGSound labels indicates strong alignment between our LLMs-based visual-to-text pipeline and expert annotations. This demonstrates the effectiveness of our query generation approach. After finetuning, the model achieves more precise separation for both same-category and different-category audio mixtures.

3.3. Ablation Study

Model	Type	FD ↓	KLD ↓	SI-SDR ↑	SDR ↑
ChatGPT5	General	39.96	2.69	-0.29	3.49
ChatGPT5[*]	**General**	**19.02**	**0.91**	**6.74**	**5.29**
Gemini-2.5-Flash	Multimodal	84.92	3.62	-10.81	3.62
Gemini-2.5-Flash[*]	Multimodal	35.98	3.09	4.41	4.66
Grok4	General	44.62	2.71	0.63	4.39
Grok4[*]	General	25.56	2.01	6.01	5.04
Qwen3-Max-Preview	Multimodal	74.48	4.37	-11.64	3.08
Qwen3-Max-Preview[*]	Multimodal	35.83	3.25	2.66	4.14
DeepSeek-VL2-Small	Reasoning	45.71	3.78	-4.05	2.52
DeepSeek-VL2-Small[*]	Reasoning	30.34	2.19	4.23	4.37

Table 2: Objective metrics comparing audio separation performance, evaluated with synthetically generated text queries. [*] indicates the usage of a fine-tuned model.

Our ablation study assesses the image-to-text capabilities of various large language models on our custom dataset, utilizing our audio separation pipeline. We assess models on image understanding, regional description, and query refinement through our pipeline to determine their effectiveness in generating user-wanted text queries. Table 2 indicates that ChatGPT-5 and Grok-4 outperform other models in these tasks. While DeepSeek-VL2-Small is behind these leaders [19], it demonstrates stronger reasoning ability than Gemini-2.5-Flash and Qwen3-Max-Preview. Crucially, our fine-tuned model consistently surpasses the original OmniSep checkpoint. The above results confirm that our LLM-driven ap-

proach significantly enhances audio separation performance, acting as an important component of our video-audio inpainting pipeline.

3.4. Qualitative Results

Inside the content in Figure 3, Object 1 stands for the masked region; in the case of our qualitative results, Object 1 stands for the people and their instruments. Meanwhile, Object 2 represents the remaining elements within the overall scene. Based on the qualitative results from Figure 3, our fine-tuned model's predicted audio spectrum closely matches the ground truth audio spectrum, demonstrating clearer separation with fewer artifacts and more alignment compared to the baseline (LLMs-Query method tested on the original checkpoint) and the iQuery prediction method. The qualitative result visually confirms the superior performance of our pipeline in audio separation.

4. CONCLUSION

We present VAInpaint, a novel audio-visual inpainting pipeline integrating segmentation (SAM2), video inpainting (ProPainter), LLMs, and a query-based audio separation model (OmniSep). Our key innovation is an LLM-driven "textual subtraction" method that generates precise separation queries by contrasting global and regional image descriptions. Supported by a custom MUSIC-VGGSound dataset, our fine-tuned model demonstrates competitive performance against current benchmarks. Ablation studies confirm the superiority of high-performance LLMs and significant gains from the fine-tuning process on our custom dataset. Although the complexity of our pipeline presents challenges, this work still advances the field of automated audio-visual editing. Our future work will focus on dynamic multi-object scenes and extend the content of text queries for better Video-Audio inpainting control.

5. REFERENCES

[1] Jie Pu, Yannis Panagakis, Stavros Petridis, and Maja Pantic, "Audio-visual object localization and separation using low-rank and sparsity," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2901–2905.

[2] Efthymios Tzinis, Scott Wisdom, Aren Jansen, Shawn Hershey, Tal Remez, Daniel PW Ellis, and John R Hershey, "Into the wild with audioscope: Unsupervised audio-visual separation of on-screen sounds," *arXiv preprint arXiv:2011.01143*, 2020.

[3] Chuan Wang, Haibin Huang, Xiaoguang Han, and Jue Wang, "Video inpainting by jointly learning temporal structure and spatial details," in *Proceedings of the AAAI conference on artificial intelligence*, 2019, vol. 33, pp. 5232–5239.

[4] Shangchen Zhou, Chongyi Li, Kelvin CK Chan, and Chen Change Loy, "Propainter: Improving propagation and transformer for video inpainting," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 10477–10486.

[5] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al., "Sam 2: Segment anything in images and videos," *arXiv preprint arXiv:2408.00714*, 2024.

[6] Xize Cheng, Siqi Zheng, Zehan Wang, Minghui Fang, Ziang Zhang, Rongjie Huang, Ziyang Ma, Shengpeng Ji, Jialong Zuo, Tao Jin, et al., "Omnisep: Unified omni-modality sound separation with query-mixup," *arXiv preprint arXiv:2410.21269*, 2024.

[7] Shentong Mo and Yibing Song, "Aligning audio-visual joint representations with an agentic workflow," *Advances in Neural Information Processing Systems*, vol. 37, pp. 58841–58867, 2024.

[8] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba, "The sound of pixels," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 570–586.

[9] Khanh-Binh Nguyen and Chae Jung Park, "Save: Segment audio-visual easy way using the segment anything model," *Computer Vision and Image Understanding*, p. 104460, 2025.

[10] Long Lian, Yifan Ding, Yunhao Ge, Sifei Liu, Hanzi Mao, Boyi Li, Marco Pavone, Ming-Yu Liu, Trevor Darrell, Adam Yala, et al., "Describe anything: Detailed localized image and video captioning," *arXiv preprint arXiv:2504.16072*, 2025.

[11] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman, "Vggsound: A large-scale audio-visual dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 721–725.

[12] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra, "Imagebind: One embedding space to bind them all," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 15180–15190.

[13] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi, "Fr\'echet audio distance: A metric for evaluating music enhancement algorithms," *arXiv preprint arXiv:1812.08466*, 2018.

[14] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al., "Cnn architectures for large-scale audio classification," in *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2017, pp. 131–135.

[15] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez, "Simple and controllable music generation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 47704–47720, 2023.

[16] Muhammad Taimoor Haseeb, Ahmad Hammoudeh, and Gus Xia, "Gpt-4 driven cinematic music generation through text processing," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 6995–6999.

[17] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey, "Sdr–half-baked or well done?," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 626–630.

[18] Jiaben Chen, Renrui Zhang, Dongze Lian, Jiaqi Yang, Ziyao Zeng, and Jianbo Shi, "iquery: Instruments as queries for audio-visual sound separation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14675–14686.

[19] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al., "Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding," *arXiv preprint arXiv:2412.10302*, 2024.