arXiv:2508.00782v1 [cs.GR] 1 Aug 2025

SpA2V: Harnessing Spatial Auditory Cues for Audio-driven Spatially-aware Video Generation

Kien T. Pham
Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong
tkpham@connect.ust.hk

Yingqing He
Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong
yhebm@connect.ust.hk

Yazhou Xing
Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong
yxingag@connect.ust.hk

Qifeng Chen
Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong
cqf@ust.hk

Long Chen*
Hong Kong University of Science and Technology
Clear Water Bay, Hong Kong
longchen@ust.hk

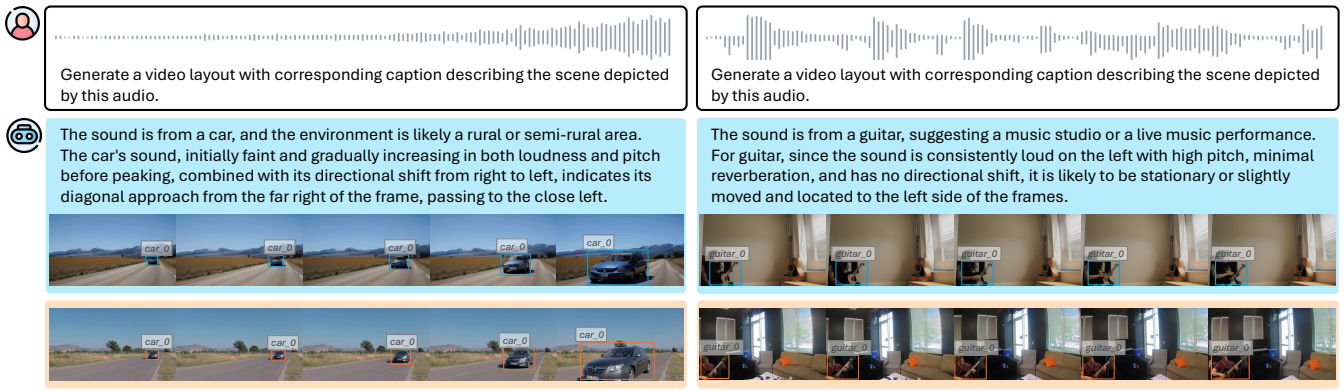


Figure 1: Audio-driven Spatially-aware Video Generation targets to synthesize realistic videos that are semantically and spatially aligned with input audio recordings. Our proposed SpA2V framework accomplishes this task by decomposing generation process into two stages: *Audio-guided Video Planning* and *Layout-grounded Video Generation*, achieving audio-video correspondence via leveraging VSLs as intermediate representation to capture auditory cues and guide the generation process respectively. Here ground-truth videos are for visual comparisons with generated videos only and are not inputted into our framework.

Abstract

Audio-driven video generation aims to synthesize realistic videos that align with input audio recordings, akin to the human ability to visualize scenes from auditory input. However, existing approaches predominantly focus on exploring semantic information, such as the classes of sounding sources present in the audio, limiting their ability to generate videos with accurate content and spatial composition. In contrast, we humans can not only naturally identify the semantic categories of sounding sources but also determine their deeply encoded spatial attributes, including locations and movement directions. This useful information can be elucidated by considering specific spatial indicators derived from the inherent physical properties of sound, such as loudness or frequency. As prior methods largely ignore this factor, we present **SpA2V**, the first framework explicitly exploits these spatial auditory cues from audios to generate videos with high semantic and spatial correspondence. SpA2V decomposes the generation process into two stages: 1) *Audio-guided Video Planning*: We meticulously adapt a state-of-the-art MLLM for a novel task of harnessing spatial and semantic cues from input audio to construct Video Scene Layouts (VSLs). This serves as an intermediate representation to bridge the gap between the audio and video modalities. 2) *Layout-grounded Video Generation*: We develop an efficient and effective approach to seamlessly integrate VSLs as conditional guidance into pre-trained diffusion models, enabling VSL-grounded video generation in a training-free manner. Extensive experiments demonstrate that SpA2V excels in

*Long Chen is the corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MM '25, Dublin, Ireland
© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3755705

generating realistic videos with semantic and spatial alignment to the input audios.

CCS Concepts

• **Computing methodologies** → **Computer vision tasks**; Image and video acquisition; **Animation**; **Spatial and physical reasoning**.

Keywords

Video Generation, Audio-driven, Spatially-aware, MLLM, Diffusion Models, Training-free

ACM Reference Format:

Kien T. Pham, Yingqing He, Yazhou Xing, Qifeng Chen, and Long Chen. 2025. SpA2V: Harnessing Spatial Auditory Cues for Audio-driven Spatially-aware Video Generation. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3746027.3755705

1 Introduction

Content creation has witnessed a significant transformation in recent years, leading to a proliferation of novel creative tasks that were previously unimaginable. This evolution is driven by the emergence of various powerful generative models capable of generating and manipulating content in different modalities, including text [11, 39, 53, 54, 63], image [12, 27, 33, 43, 44, 56], audio [8, 35, 36, 65], and video [1, 15, 19, 20, 29, 34, 61]. Particularly in the context of video generation, the advancement is becoming more elusive with many current works making progress in synthesizing video content based on text description [20, 61] and initial image [1, 61]. Despite showing impressive results, they often fall short in capturing the richness and temporal coherence of real-world events, because of the inherent ambiguity and static nature of their respective conditions.

Audio, in contrast, naturally grounds video in reality and encodes abundant temporal and contextual information on sound-emitting objects, their interactions, and the spatial arrangement of the soundscape. These intrinsic values provide unique advantages for generating more nuanced, immersive, and temporally consistent video content, leading to more realistic and engaging experiences. In addition, similar to the human ability to use auditory information to depict corresponding visual scenes and events, audio-to-video generation can be applied to diverse applications that span across industry verticals. Some of these include automated scene visualization in filmmaking, dynamic product creation in multimedia, engaging advertisements in marketing, and accessible learning materials in education. In light of these significant advantages and useful applications, it is imperative to explore the field of **audio-driven video generation**.

The prevailing audio-to-video generation methods typically rely on global semantic features extracted from audio tracks for synthesis. Although this approach can produce semantically aligned videos, it only works for specific simple soundscapes and often results in poor content quality and misaligned spatial composition with input audio in general scenarios. For example, existing works including [14, 38, 50, 51] can generate talking head videos conditioned on speech, yet are not applicable to other domains. Other methods such as [6, 47, 58, 62] can synthesize videos of different contexts that are globally aligned with the specific semantic categories (e.g., dancing, drumming, landscape, etc.) of the input audio recording. However, their results lack spatial coherence between visual and auditory elements, affecting realism and ultimately diminishing the immersive experience. Some current approaches [3, 64] alleviate such an issue by directly providing an initial frame or video segment which already establishes a spatial correspondence with audio as an additional visual input, but inevitably limits the diversity of the generated content.

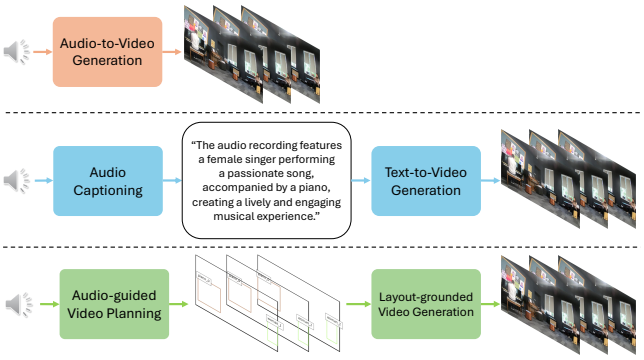


Figure 2: Different frameworks for audio-driven video generation. From top to bottom are the typical Audio → Video direct approach, two-stage Audio → Text → Video method, and our proposed novel Audio → Video Scene Layout → Video pipeline respectively.

Surprisingly, all the aforementioned works have largely overlooked the fact that sound inherently encompasses rich spatial information such as location and movement of sounding sources present in the scenes. Such information can be harnessed to generate according visual components with not only semantic but also spatial coherence to input audios. To this end, the first critical question that arises is: **Q1:** *Can we directly decode the spatial information embedded within audio to drive video generation?* We draw inspiration from the fact that humans spontaneously perform similar tasks to perceive and navigate the environment in our daily hearing. We intuitively utilize our multisensory and commonsense knowledge to exploit specific auditory cues from environmental sounds, then reason on them to derive necessary information. For instance, considering the top-left example in Fig. 1, we can instinctively imagine an approaching car when hearing its engine sound getting louder. This is because we know what a car generally sounds and looks like (semantic clue) and deduce that increasing in volume (spatial clue) implies approaching motion. By targeting these auditory cues, we contemplate that a strong foundational model with human-like multimodal understanding and reasoning capabilities like MLLM has the potential to adapt and replicate this human instinct, driving us to explore it extensively to address this challenge.

Once **Q1** is properly resolved, the important subsequent question that emerges is **Q2:** *How should these information be represented to bridge the gap between audio and video modalities and guide the generation process?* At first thought, text description seems like a viable option. However, it suffers from inherent ambiguity, leading to inconsistent results and a lack of precise spatial control over scene

composition in generation process. Video Scene Layout (VSL), on the other hand, offers a structured and unambiguous representation, enabling fine-grained manipulation of object placement and scene structure. Considering our concentration on spatial relationships between auditory and visual elements, VSL is intuitively advantageous compared to the textual counterpart. Therefore, we adopt it as our intermediate representation to capture the semantic and spatial attributes of the sounding sources extracted from input audio and then control the video generation process as shown in Fig. 2.

We propose a novel framework dubbed **SpA2V** which is the first attempt to explicitly exploit spatial auditory information for video generation conditioning solely on audio. SpA2V decomposes the generation process into two respective stages, namely Audio-guided Video Planning and Layout-grounded Video Generation. The first stage is responsible for identifying sounding objects occurring in an input audio and inferring their semantic and spatial attributes to construct a VSL as guidance for generation in the subsequent stage. We employ a state-of-the-art Multimodal Large Language Model (MLLM), such as Gemini 2.0 [52] or GPT4o [40], with demonstrated powerful understanding and reasoning capabilities across different modalities as the Video Planner for our SpA2V. We adapt them for our new task of audio-driven VSL generation via a meticulously designed prompting mechanism that leverages In-context Learning [2], allowing it to effectively and efficiently harness semantic and spatial cues presented in input audio.

Following VSL generation, we synthesize the final video by conditioning on the VSL in the second stage. Our approach incorporates pre-trained diffusion models in an efficient and effective way, inspired by MIGC [67] and AnimateDiff [19]. These methods augment the pre-trained Stable Diffusion model with spatial grounding and motion modules for layout-to-image and text-to-video tasks. We exploit the fact that they train only these new modules while keeping the backbone intact. By directly integrating their learned modules into the same frozen backbone, we create a layout-to-video diffusion model capable of spatial grounding and motion modeling simultaneously without further training. We hereby employ it as our VSL-grounded video generator to complete this stage.

To assess the capability of our SpA2V framework, we introduce a new benchmark named AVLBench curated from real-world stereo audio-video recordings [13, 16, 69, 70] and repurposed for our specific use cases. It includes diverse test scenarios featuring different numbers of sounding objects with various spatial attributes. Results from our experiments on this benchmark demonstrate that SpA2V achieves a high degree of semantic and spatial correspondence between the generated VSLs, videos, and the input audios, marking the first successful attempt of spatially-aware audio-to-video generation.

Overall, our contributions are listed as follows:

- We propose a novel task of audio-driven spatially-aware video generation which aims at synthesizing videos with spatial correspondence to audio conditions.
- We present SpA2V, the first framework attempting to fulfill the task by decomposing the generation process into two stages Audio-guided Video Planning and Layout-to-Video Generation and leveraging powerful pre-trained MLLMs and diffusion models to accomplish each stage, respectively.
- We introduce AVLBench, a new benchmark for evaluating alignment between input audios and generated VSLs and videos.
- Extensive experiments on the benchmark highlight the capability of SpA2V in generating realistic VSLs and videos where visual elements correspond both semantically and spatially to the sound sources in input audio. The implementation will be released on GitHub[1].

2 Related Work

Audio-Visual Learning. Recent years have witnessed growing research efforts in audio-visual learning. Early studies primarily focused on cross-modal Audio-Visual Synchronisation [4, 24, 25], which employed self-supervised learning to align temporal relationships between audio and video, establishing foundational representations for downstream tasks. Despite resolving temporal alignment, these methods largely overlooked semantic and spatial correlations between audio and visual modalities. To provide more detailed audio-visual spatial alignment and support the audio as a guiding signal, [7, 68] explore the Audio-Visual Segmentation (AVS) task that pioneered the prediction of sounded object segmentation maps in video frames conditioned on audio input. However, they focused on perception rather than generation, limiting their applicability to generative tasks. Critically, they also neglected to model the spatial attributes of audio (e.g., sound source localization or motion trajectories), which are vital for grounding visual scenes in physical reality. Recent advances have started to explore *spatial audio* for audio-visual tasks. For example, BAT [66] leverages large language models (LLMs) for spatial sound reasoning, while ELSA [9] learns spatially-aware language-audio representations for fine-grained localization. Building on these insights, our work is the first to explicitly exploit spatial audio cues for audio-guided video generation, enabling the synthesis of videos where visual elements are both semantically and spatially coherent with sound sources.

Audio-to-Video (A2V) Generation. A2V generation focuses on producing visual content that aligns with given audio inputs. Several studies have explored this domain by leveraging audio to provide semantic cues and temporal dynamics for video generation. Sound2Sight [3] and CCVS [30] utilize audio alongside preceding video frames to forecast subsequent frames, capturing visual dynamics driven by the input audio. [31] employs StyleGAN, projecting audio into its latent space to navigate trajectories within this space, effectively aligning audio with visual content. Seeing and Hearing [58] introduces a diffusion latent aligner to synchronize audio with visual elements, enhancing the coherence between them. TempoTokens [62] adapts a pre-trained text-to-video diffusion model for A2V generation, aligning audio and visual components to improve synchronization. Although these approaches focus on semantic and temporal alignment, they often overlook the spatial aspect when processing input audio. Spatial information such as the location and distance of sounded objects can bring significant enhancement to the generated results. In this work, we pioneer the exploration of harnessing important spatial cues from input audio to guide video generation and fulfill this gap. We term the task as **audio-driven spatially-aware video generation**.

[1] https://github.com/tkpham3105/SpA2V

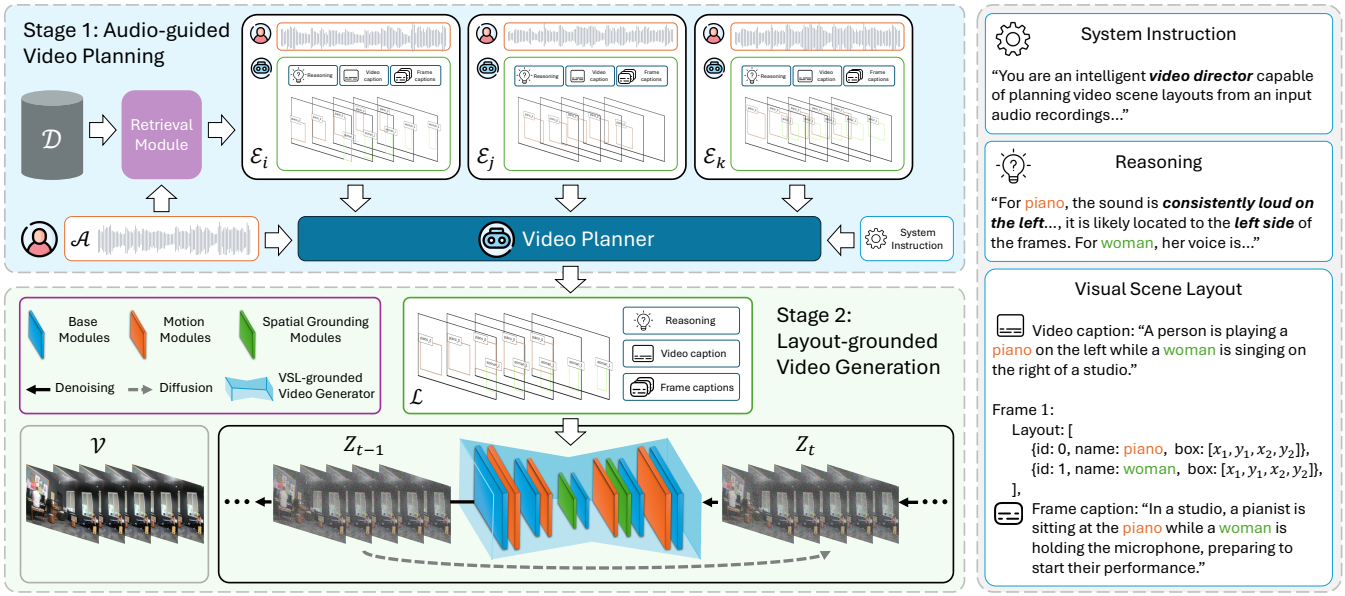


Figure 3: Illustration for the overall framework of SpA2V which is decomposed into two stages: Audio-guided Video Planning and Layout-grounded Video Generation. In the first stage (Section 3.1), given an input audio $\mathcal{A}$, we retrieve k example conversations $\{\mathcal{E}_1, \mathcal{E}_2, \dots \mathcal{E}_k\}$ from candidate database $\mathcal{D}$ via Retrieval Module and feed them together with a System Instruction and the audio itself into the MLLM Video Planner to perform reasoning and generate a desired VSL sequence $\mathcal{L}$ containing N consecutive keyframe layouts $\{\mathcal{L}_1, \mathcal{L}_2, \dots \mathcal{L}_N\}$ with respective global video caption and local frame captions. In the second stage (Section 3.2), the obtained VSL $\mathcal{L}$ and its captions are incorporated to guide a video diffusion model consisting of pretrained Base, Motion, and Spatial Grounding Modules to generate the final video $\mathcal{V}$ that is semantically and spatially coherent with the input $\mathcal{A}$.

3 Method

Although audio contains a significant amount of semantic and spatial information, effectively extracting and incorporating them for video generation are non-trivial and underexplored tasks. In this section, we describe how each stage of our proposed SpA2V framework tackles these challenges respectively.

3.1 Stage 1: Audio-guided Video Planning

Overview. In this stage, we introduce a novel task: generating video scene layouts (VSLs) depicting spatial arrangements of sounding objects presented in corresponding audio recordings. This task necessitates a model to first identify the categories of sounding sources (semantic component) and their respective locations and movements (spatial components) from the input audio. Then, the model must use this information to organize the objects into a coherent VSL, accurately reflecting their spatial correspondence with the audio and maintaining content consistency across the video sequence. Given these requirements, Multimodal Large Language Models (MLLMs) are particularly well-suited due to their strong multimodal understanding, reasoning abilities, and broad foundational knowledge. Consequently, we empirically investigate the potential of MLLMs to effectively address this challenging task.

Instruction Setup. To generate an audio-conditioned VSL using an MLLM, we query it with a prompt consisting of three components: a system instruction, a set of example conversations, and a user-specified audio recording. The system instruction includes task definition and guidance regarding the desired behavior and response for each request that the MLLM must follow. Specifically, we instruct the MLLM to act as a *video director* to plan VSLs that capture the content of the input audio recordings. We then outline the task requirements for the MLLM to fulfill, such as the expected layout format, coordinate system, canvas size, and number of frames. In complement, the example conversations provide the MLLM with reference query-response pairs, allowing it to efficiently learn and adapt to the given task. Finally, after supplying the above contextual information, we query the MLLM to perform completion on the user's input audio recording to generate the desired VSL.

VSL Structure. We ask the MLLM to generate VSLs according to a predefined template which is a connected sequence $\mathcal{L}$ of N consecutive keyframe layouts $\{\mathcal{L}_1, \mathcal{L}_2, \dots, \mathcal{L}_N\}$. Every layout $\mathcal{L}_i$ contains a set of N_i bounding boxes $\{\mathcal{B}_1, \mathcal{B}_2, \dots, \mathcal{B}_{N_i}\}$, each denotes a sounding object that occurs in the input audio. Each bounding box is represented by its location and size in numerical coordinates along with a labeling phrase that specifies the enclosed object. In addition, each box is assigned a unique numerical identifier, establishing and maintaining object correspondence across frames without the need for a dedicated box tracker. Finally, each VSL also entails a shared global video caption and a local frame caption for each keyframe describing the global content and local dynamic transition of the intended video creation. Note that in these captions, the MLLM has the freedom to bring about special information that cannot be inferred from input audio but is beneficial for video generation later such as visual appearance of sounding objects.

Spatial Reasoning. Spatial information can be inferred by reasoning on the fundamental spatial auditory cues, such as Interaural Time Difference (ITD), Interaural Level Difference (ILD), pitch and volume, and directional shift. ITD and ILD are typically used to infer the location of sounding objects, while pitch and volume often indicate their distance, and directional shift can imply their movement. To accurately deduce the corresponding spatial attributes and minimize spurious hallucinations, we explicitly instruct the MLLM in the system instruction to focus on analyzing these key indicators. Consequently, we ask the MLLM to output a brief statement summarizing its reasoning and the extracted spatial cues before generating the VSL to enhance the interpretability of the final response.

In-context Learning. Solely relying on the system instruction to provide task descriptions and reasoning guidance may fall short in allowing the MLLM to comprehend our need for precise real-world understanding and reasoning on the aforementioned physical sound properties, causing it to hallucinate incorrect spatial information with fuzzy or non-sensical reasoning, and eventually generate VSLs misaligned with the input audio recordings as shown in Fig. 4. Inspired by [2, 10, 37, 60] which show that In-context Learning can enhance the LLMs' task adaptability and compliance in various contexts, we employ it to further guide the behavior of the MLLM and mitigate mentioned problem. For each query, we provide the MLLM with example conversations, each including a reference prompt and a high-quality VSL with corresponding reasoning statement. Akin to [37], we hypothesize that the more semantically similar the audio recordings of the examples to that of the query, the more informative it can be for the MLLM. Therefore, we conduct *Top-k* Nearest Neighbor (*k*NN) search on CLAP [57] embedding space in our Retrieval Module to select k examples for each query.

3.2 Stage 2: Layout-grounded Video Generation

Overview. Leveraging the ability of MLLMs to generate semantically and spatially aligned VSLs and descriptive captions from auditory cues, we subsequently introduce an approach for video synthesis controlled by these VSLs in this stage. Our VSL-grounded Video Generator connects off-the-shelf layout-to-image and text-to-video diffusion models into a single pipeline. By combining their respective grounding and temporal modeling capabilities, our generator produces videos that adhere to the conditioned VSLs and entailed captions, thereby maintaining consistency with the input audio. Our method operates in a training-free manner that efficiently reduces computational cost and time, eliminates the need for extensive data annotation, and avoids potential catastrophic forgetting incurred by training.

Base Diffusion Model. We build our VSL-grounded Video Generator based on the pre-trained text-to-image LDM [45], *a.k.a* Stable Diffusion, of which the diffusion procedure follows the standard formulation in [22, 48, 49] that comprises a forward diffusion and a backward denoising process. Given a data sample $\mathbf{X} \sim \mathcal{P}(\mathbf{X})$, an autoencoder consisting of an encoder $\mathscr{E}$ and a decoder $\mathscr{D}$ will first project its latent correspondence $\mathbf{Z}_0 = \mathscr{E}(\mathbf{X})$. Subsequently, the diffusion and denoising processes are conducted in latent space. In one hand, the forward diffusion is essentially a fixed Markov process of T timesteps that gradually perturbs $\mathbf{Z}_0$ to yield $\mathbf{Z}_t$ via:

$$\mathbf{Z}_t = \sqrt{\bar{\alpha}_t}\mathbf{Z}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \epsilon \sim \mathcal{N}(0, \mathbf{I}), \tag{1}$$

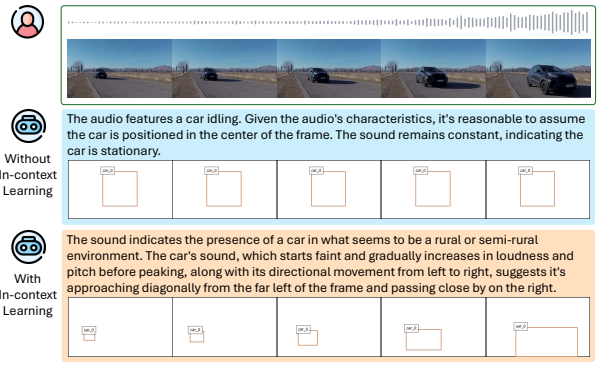


Figure 4: In-context Learning helps guide the MLLM to derive the correct spatial cues from the right physical sound properties and hence generate a highly-aligned VSL.

for $t = 1, 2, \ldots, T$. Here $\bar{\alpha}_t$ is pre-defined parameter which determines the noise strength at each timestep t. Eventually, $\mathbf{Z}_0$ turns into $\mathbf{Z}_T$ that is indistinguishable from a Gaussian noise. On the other hand, the backward process leverages a denoising network ϵ_θ with training objective of minimizing:

$$\mathbb{E}_{t,\mathbf{C},\mathbf{Z}_t,\epsilon\sim\mathcal{N}(0,\mathbf{I})}||\epsilon - \epsilon_\theta(\mathbf{Z}_t, t, \tau_\theta(\mathbf{C}))||_2^2, \tag{2}$$

where $\mathbf{C}$ is the condition and τ_θ represents its encoder, to iteratively denoise $\mathbf{Z}_t$. Once the denoising is finished and a final clean latent $\hat{\mathbf{Z}}_0$ is obtained, the generated sample can be decoded via $\hat{\mathbf{X}} = \mathscr{D}(\hat{\mathbf{Z}}_0)$. In Stable Diffusion, ϵ_θ adopts UNet [46] architecture comprising of down/middle/up blocks each consisting of ResNet [21], spatial self-attention layers, and cross-attention layers that incorporate text conditions. For conciseness, we call these blocks Base Modules in our VSL-grounded Video Generator, responsible for preserving pre-learned knowledge to generate diverse and high-fidelity samples guided by text prompts in image domain.

Integrating Grounding and Temporal Modeling. With Stable Diffusion as the base model, we respectively integrate pretrained Temporal Modules and Grounding Modules from AnimateDiff [19] and MIGC [67] into our VSL-grounded Video Generator, enabling spatial grounding and motion modeling capabilities to synthesize high-quality videos aligned with input VSLs. Specifically, AnimateDiff proposes learning meaningful motion priors by injecting temporal transformer blocks, namely Motion Modules, to inflate Stable Diffusion, allowing it to generate motion dynamics of visual content over time while alleviating quality degradation. Meanwhile, MIGC incorporates a set of articulated instance enhancement attention layers, which we call Spatial Grounding Modules, into Stable Diffusion to enable precise generations of multiple instances in the resulting image following a layout input. Since only these external modules are trained to learn their designated objectives while the same base modules are kept frozen, we hypothesize then empirically verify that directly combining them into a single end-to-end pipeline, *i.e.* our VSL-grounded Video Generator, can achieve both spatial grounding and motion modeling abilities.

Video Generation with VSL Guidance. Every VSL $\mathcal{L}$ comprises a sequence of N consecutive keyframe layouts, each containing a set of object bounding boxes, a shared global video caption, and a local

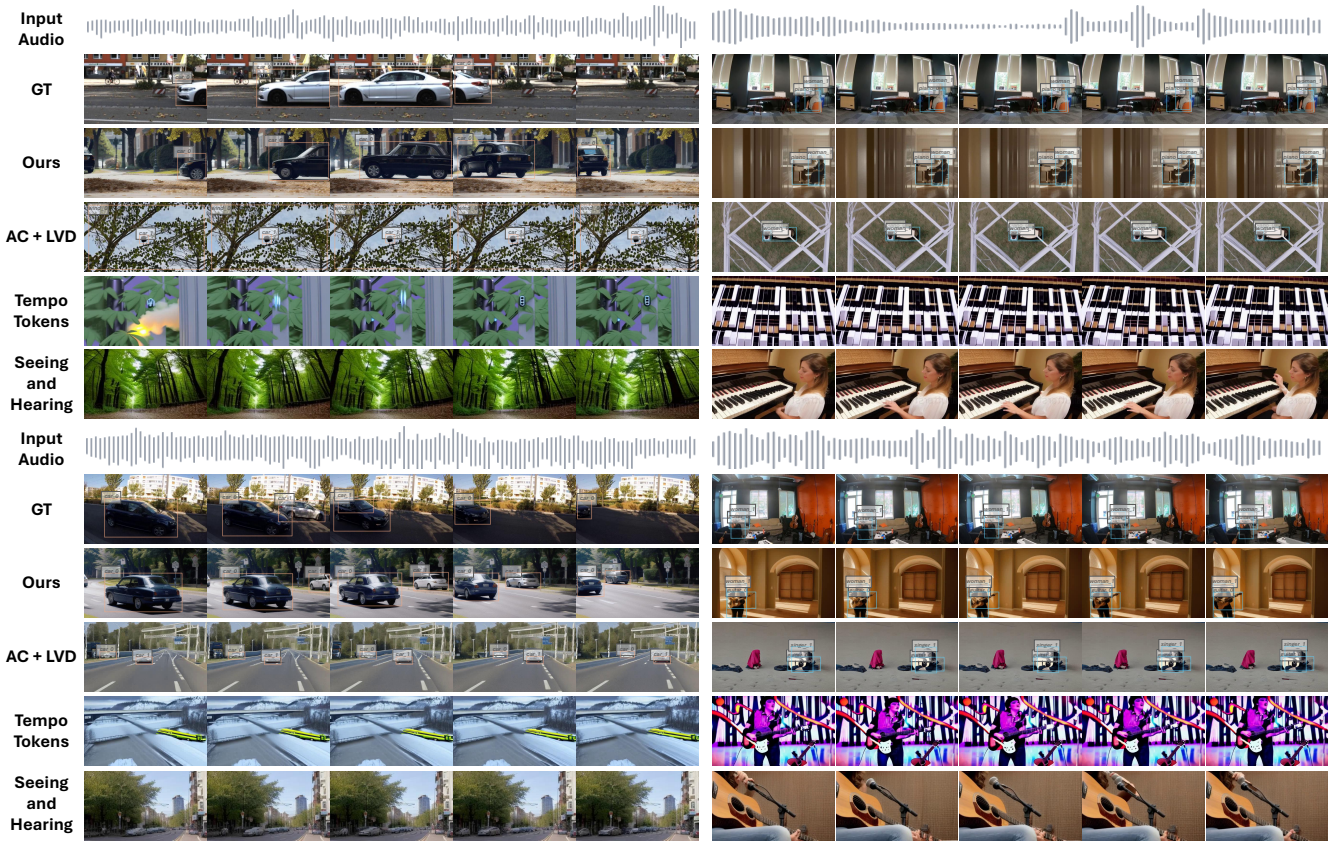


Figure 5: Qualitative comparisons of our SpA2V with prior SOTA works in audio-to-video generation. Here GT denotes ground-truth videos and VSLs for illustration of visual elements present in input audios. Zoom-in for details.

frame caption. To control our VSL-grounded Video Generator to synthesize a video of n frames, we first perform temporal-wise linear interpolation on the coordinates of the bounding boxes for each object to obtain a denser VSL with expanded length of n layouts. Each layout will then serve as a grounding signal for a corresponding frame. Since the base diffusion model uses text prompt as global condition for generation, we also input the global video caption of the VSL to preserve its pre-trained generative capability and maintain global consistency across generated frames. In addition, for the N keyframes, we use their local frame caption as alternative to global caption that empirically helps produce better frame transitions with more natural local dynamics.

4 Experiments

4.1 Setup

Benchmark. As our proposed two-stage Audio → VSL → Video pipeline is novel, there is no existing benchmark suitable for evaluating our SpA2V framework. Therefore, we created AVLBench, a new benchmark specifically designed for our use case, curated from real-world stereo audio-video recording datasets [13, 16, 69, 70] spanning a variety of sound sources, including instruments and moving vehicles in indoor and outdoor environments. We begin by manually selecting recordings for which the audio contains strong semantic and spatial signals clearly indicating the sounding sources and their attributes within the video. After filtering, we apply flip and reverse augmentations with quality control to increase the data diversity while maintaining strong correspondence between auditory and visual elements. Subsequently, we use Track Anything [59] to generate ground-truth VSLs by tracking the sounding objects in the videos. Since we need to provide SpA2V's Video Planner with example conversations for In-context Learning, we adopt LLaVA-OneVision [32] to generate global video caption and local frame captions for each video. We also include an accurate manually written reasoning statement for every sample. Eventually, AVLBench contains 7274 testing samples, of which 4702 samples are used to assess scenarios of single or multiple instruments playing while having *Stationary* motion in indoor settings, whereas the rest 2572 samples target cases of single or multiple vehicles with *Translational* movement in outdoor settings.

Implementation Details. The overall structure of our SpA2V framework is illustrated in Fig. 3. In Stage 1, we select Gemini 2.0 Flash [53] as our MLLM Video Planner to balance cost-effectiveness and performance. For each input audio, we provide it with $k = 3$ example conversations retrieved from the candidate database via Retrieval Module that performs a kNN search based on the similarity between CLAP embeddings of the input and the candidates.

Method	Combo	IL Setup	Eg. Sel.	(M)LLM	τ	Stationary									Translational								
						MaxIoU ↑			*LTSim* ↑			*DocSim* ↑			*MaxIoU* ↑			*LTSim* ↑			*DocSim* ↑		
						S	M	C	S	M	C	S	M	C	S	M	C	S	M	C	S	M	C
AC + LVD [34]		3-shot	Default	GPT4	0.5	0.92	0.96	0.94	40.48	46.32	42.97	4.40	4.91	4.61	1.79	1.51	1.77	47.51	45.17	47.35	3.69	3.98	3.71
SpA2V (Ours)	Full	3-shot	*k*NN	G2.0F	0.5	20.16	18.55	19.45	76.73	74.43	75.73	15.69	15.06	15.47	22.62	20.13	22.24	77.55	73.90	77.21	16.77	13.66	16.50
	w/o SR					14.57	8.75	12.03	74.90	69.24	72.41	14.39	13.10	13.90	17.10	15.43	16.87	75.09	72.27	74.88	17.03	12.99	16.74
	w/o IL	N/A	N/A			3.93	1.71	3.00	62.64	56.01	59.84	4.55	4.18	4.40	5.19	2.22	4.96	62.72	56.36	62.26	6.07	4.71	5.98
	Vanilla					4.63	1.77	3.42	66.56	56.37	62.23	5.47	4.61	5.10	6.58	2.58	6.32	67.37	60.52	66.91	6.05	4.05	5.93
	Full	3-shot	*k*NN	G2.0F	0.5	20.16	18.55	19.45	76.73	74.43	75.73	15.69	15.06	15.47	22.62	20.13	22.24	77.55	73.90	77.21	16.77	13.66	16.50
		2-shot				14.46	8.03	11.72	74.58	70.59	72.86	15.01	14.30	14.72	11.27	11.28	11.26	73.35	71.35	73.24	14.70	12.51	14.54
		1-shot				10.18	5.30	8.02	71.17	66.32	69.12	10.39	11.03	10.65	7.86	7.08	7.80	70.40	66.61	70.14	10.44	9.65	10.39
	Full	3-shot	*k*NN Random	G2.0F	0.5	20.16	18.55	19.45	76.73	74.43	75.73	15.69	15.06	15.47	22.62	20.13	22.24	77.55	73.90	77.21	16.77	13.66	16.50
						4.47	2.28	3.58	62.07	56.33	59.57	6.86	7.34	7.13	8.01	4.46	7.71	71.43	67.03	71.10	9.11	8.18	9.03
	Full	3-shot	*k*NN	G2.0F	0.5	20.16	18.55	19.45	76.73	74.43	75.73	15.69	15.06	15.47	22.62	20.13	22.24	77.55	73.90	77.21	16.77	13.66	16.50
				G1.5F		7.04	3.13	5.43	70.19	65.05	68.11	12.24	11.44	11.91	6.40	4.41	6.26	71.55	67.38	71.25	8.96	8.74	8.95
				G4oM		17.15	11.78	14.54	72.16	68.11	70.21	13.96	11.92	13.09	12.40	8.69	12.11	66.23	63.45	66.04	9.91	8.13	9.75
	Full	3-shot	*k*NN	G2.0F	0.5	20.16	18.55	19.45	76.73	74.43	75.73	15.69	15.06	15.47	22.62	20.13	22.24	77.55	73.90	77.21	16.77	13.66	16.50
					1.0	16.54	16.87	16.42	75.73	73.92	74.83	14.59	14.47	14.49	18.45	16.51	18.24	76.11	72.57	75.85	14.89	12.64	14.70
					1.5	14.09	14.31	14.22	74.76	73.04	74.02	13.76	13.63	13.68	15.87	16.74	15.84	75.39	72.19	75.11	14.05	12.33	13.91

Table 1: Quantitative results and ablation analysis conducted for Audio-driven Video Planning in Stage 1. Here AC, SR, IL, and Eg. Sel. are shortened for Audio Captioning, Spatial Reasoning, In-context Learning, and Example Selection, respectively. τ denotes the temperature value of (M)LLM and ↑ indicates higher values are better. S and M represent subsets of data samples having single or multiple sounding sources, while C represents combinations of all scenarios. G2.0F, G1.5F, and G4oM stands for Gemini 2.0 Flash, Gemini 1.5 Flash, and GPT4o Mini accordingly.

Subsequently, we prompt the Video Planner to generate VSL consisting of $N = 5$ keyframe layouts of resolution 454×256 with a temperature of $\tau = 0.5$ to control the randomness of its response. In Stage 2, Stable Diffusion 1.5 [44] is adopted as the base diffusion model of our VSL-grounded Video Generator. We then follow default settings in MIGC [67] and AnimateDiff [19] to accordingly deploy Spatial Grounding Modules and Motion Modules onto the base model. With this complete architecture, our Video Generator performs inference to synthesize video of $n = 16$ frames with resolution 512×320 conditioned on the VSL obtained from Stage 1. Unless otherwise specified, these settings are kept by default.

Metrics. In Stage 1, to measure the quality of the results in alignment with the input audio, we compute the similarity between the generated VSL and the ground-truth utilizing three metrics namely *LTSim* [41] , *MaxIoU* [28], *DocSim* [42]. These metrics are designed for image layouts that contain bounding boxes of close-set labels. To calculate the similarity between a pair of layouts $(\mathcal{L}, \mathcal{L}')$, they first match their enclosed set of bounding boxes $(\{\mathcal{B}_i\}, \{\mathcal{B}'_j\})$ then accumulate coordinate IoU scores of the matched boxes. Matching typically involves an indicator function $f_{abs}(\mathcal{B}_i, \mathcal{B}'_j) = \mathbb{I}_{\{c_i = c'_j\}}$ that fully ignores box pairs with different categories. Nevertheless, our method generates VSL which is essentially a sequence of image layouts consisting of bounding boxes with free labels. Therefore, we adjust this indicator function to a *soft* version $f_{soft}(\mathcal{B}_i, \mathcal{B}'_j) = cosine(P(c_i), P(c'_j))$ that measures the similarity of the two categories (c_i, c'_j) in the projector P's embedding space [5]. We then follow the rest of calculations for all metrics and average the score across frames for each VSL.

For Stage 2, we adopt the standard FVD [55] and AV-Align [62] to accordingly assess the overall content quality of generated videos and their temporal alignment with input audios. Especially, to evaluate spatial correspondence, we first utilize OV-AVSS [18] to localize the input audios' sounding objects within the synthesized videos to obtain respective VSLs. Subsequently, we compute the *LTSim* [41] scores between these and the ground-truth VSLs.

Baselines. Since there is no previous work explicitly explores audio-driven video planning, we choose a relevant baseline named LVD [34] for comparison in Stage 1. Note that this approach generates dynamic scene layout conditioning on text prompt, whereas our task requires audio as the sole input guidance. Therefore, we adopt an audio captioning (AC) model [17] to generate a textual description for each input audio and feed them into LVD respectively. For Stage 2, we additionally compare our SpA2V framework with TempoTokens [62], Seeing and Hearing [58], and LTX [20], alongside LVD for system-level evaluations of audio-to-video generation capabilities. TempoTokens follows the typical Audio → Video direct pipeline for generation. Meanwhile, since Seeing and Hearing and LTX use textual condition, we provide them with the same audio captions as LVD. Therefore, they can be categorized as two-stage Audio → Text → Video approaches, as shown in Fig. 2 respectively.

4.2 Evaluation of Audio-guided Video Planning

Overall Results. As demonstrated in Tab. 1 and Fig. 5, our SpA2V framework can generate VSLs with high similarity to the ground-truth VSLs which indicate strong alignments to the input audios. SpA2V significantly outperforms the baseline of combining audio captioning with LVD [34] in all metrics and test scenarios.

Component Ablation. We ablate each component of the MLLM Video Planner to evaluate their effectiveness accordingly. As indicated in Tab. 1, both In-context Learning and Spatial Reasoning are crucial for the planner to appropriately adapt to the instructed task and generate high-quality VSLs, omitting either one will lead to significant performance degradations. Interestingly, Spatial Reasoning needs to be accompanied by In-context Learning to synergistically help the planner achieve the best performance. Incorporating it alone may detrimentally confuse the planner and lead to subpar performance compared to not integrating both (Vanilla).

In-context Learning Setup. We assess the performance of the MLLM Video Planner when providing it with different numbers

Method	Cap. Sel.	VSL Sel.	Stationary									Translational								
			FVD ↓			*AV-Align* ↑			*LTSim* ↑			*FVD* ↓			*AV-Align* ↑			*LTSim* ↑		
			S	M	C	S	M	C	S	M	C	S	M	C	S	M	C	S	M	C
TempoTokens [62]			878.70	759.22	691.89	0.153	0.134	0.145	34.49	34.47	34.22	1549.51	1355.67	1462.94	0.179	0.171	0.179	29.61	28.90	29.56
Seeing and Hearing [58]			715.77	708.58	664.87	0.111	0.105	0.109	36.47	41.81	38.72	1144.97	979.19	1049.42	0.151	0.127	0.149	29.45	33.88	29.76
AC + LTX [20]			619.97	525.14	543.81	0.091	0.088	0.090	34.75	39.55	36.79	1094.19	1154.74	1022.49	0.156	0.130	0.154	32.57	37.61	32.92
AC + LVD [34]		Gen.	814.03	793.48	712.55	0.156	0.129	0.144	31.40	33.65	32.43	1306.68	793.48	1196.76	0.158	0.126	0.156	42.31	46.27	42.58
SpA2V (Ours)	Mix	Gen.	776.63	527.31	633.05	0.186	0.155	0.173	46.22	50.62	48.10	302.88	594.38	278.99	0.170	0.187	0.171	69.50	61.71	68.96
	Global		779.08	529.57	637.84	0.184	0.153	0.170	45.07	49.90	47.12	308.44	596.19	282.39	0.170	0.178	0.171	69.49	59.48	68.79
	Local		790.47	536.77	643.30	0.176	0.154	0.166	44.99	50.02	47.14	313.23	598.37	288.01	0.159	0.184	0.161	69.44	62.24	68.94
	Mix	Gen.	776.63	527.31	633.05	0.186	0.155	0.173	46.22	50.62	48.10	302.88	594.38	278.99	0.170	0.187	0.171	69.50	61.71	68.96
		GT	744.79	515.20	619.05	0.185	0.158	0.173	49.83	52.05	50.77	244.55	576.18	231.84	0.170	0.180	0.171	78.67	65.13	77.72

Table 2: Quantitative results and ablation analysis conducted for Layout-grounded Video Generation in Stage 2 and system-wise comparisons. Here Cap. Sel. and VSL Sel. denotes Caption Selection and Video Scene Layout Selection, ↑ and ↓ indicate higher or lower values are better, and Gen. and GT are shortened for Generated and Ground-truth VSL. Besides, S and M represent subsets of data samples having single or multiple sounding sources, while C represents combinations of all scenarios.

of example conversations. Compared to the zero-shot Vanilla, In-context Learning consistently brings improvements to the planner by delivering more context information via selective examples.

Example Selection. We aim to empirically verify our assumption in Section 3.1 that the more reference audio recordings in the example conversations are semantically similar to that of the query, the better information it can bring to the MLLM Video Planner to generate higher quality VSLs. We replace the kNN Searching strategy in the Retrieval Module with a simple random selection while keeping other settings as default. As shown in Tab. 1, this adjustment severely harms the overall performance, highlighting the advantages of the kNN Searching strategy we adopt.

Choices of MLLM. Since the design of our SpA2V is flexible, it allows better models selected as its components to attain better performance. Here we try to employ different state-of-the-art MLLMs as the Video Planner for SpA2V. Specifically, we conduct the same experiments but switch from the default Gemini 2.0 Flash to its predecessor Gemini 1.5 Flash and GPT4o Mini [40]. The results in Tab. 1 demonstrate that the default option significantly exceeds these alternatives, making it the best choice to accomplish this task.

Temperature. We test the performance of our SpA2V's MLLM Video Planner with different values for temperature which controls the randomness of its response with higher values being more creative while lower ones being more deterministic. Apparently, a low value of 0.5 best suits our need for the task, as shown in Tab. 1.

4.3 Evaluation of Layout-to-Video Generation

Overall Results. As shown in Tab. 3.2 and Fig. 5, our SpA2V framework can generate high-quality videos with compelling semantic and spatial correspondence to input audios across various scenarios. Meanwhile, the synthesized videos of prior works are prone to having limited semantic coherence and inconsistent spatial composition with input audios. Additionally, these methods tend to create videos with minimal dynamics and struggle with cases where sounding objects have large movements. These results highlight the superiority of our proposed SpA2V framework and its two-stage Audio → VSL → Video pipeline in harnessing informative auditory cues from input audios for video generation objectives. Besides, SpA2V also achieve competitive *AV-Align* scores which imply strong temporal alignment between generated videos and input audios. We attribute this to the MLLM Video Planner which has the innate potential to capture temporal features in complement of semantic and spatial cues from input audios. These information will then be harmoniously organized into according VSLs and propagated to the subsequent video generation.

Caption Selection. As indicated in Tab. 2, simultaneously utilizing shared global and local keyframe captions as text conditions alongside VSL is empirically effective in enhancing SpA2V's performance. While the former helps preserve the pre-trained generative capability of employed diffusion model and maintain global consistency, the latter encourages better frame transitions with more natural local dynamics across generated frame.

Impact of VSL Quality. To evaluate the importance of VSL quality in generating high-fidelity videos aligned with input audios, we skip the video planning steps in Stage 1 and directly use ground-truth VSLs as alternative control signals to guide the video synthesis process in Stage 2. As demonstrated in Tab. 2, such an adjustment substantially enhances overall performance, indicating that the better the quality of VSLs, the better results we can achieve. Since our two-stage pipeline is implementation-agnostic, this observation further implies that our SpA2V framework can continue to improve generation quality and audio-video alignment by adopting more capable MLLMs and video diffusion models in flexible manner.

5 Conclusion

We have presented **SpA2V**, the first framework capable of harnessing spatial auditory cues for audio-driven spatially-aware video synthesis. SpA2V decomposes the generation process into two stages: *Audio-guided Video Planning* and *Layout-grounded Video Generation*. In Stage 1, we adopt a SOTA MLLM as the Video Planner and instruct it to generate VSLs from input audios through a diligently designed prompting mechanism. In Stage 2, we propose an effective Video Generator which efficiently incorporates off-the-shelf diffusion models to synthesize videos grounded by the VSLs obtained from previous stage. The experimental results on our newly introduced AVLBench benchmark highlight the superiority of our SpA2V in producing videos with high semantic and spatial consistency to the input audios, outperforming previous methods by large margins. We hope that our pipeline will encourage further exploration into related areas of study in the future.

Acknowledgement. This research was supported by the Innovation and Technology Fund of HKSAR (GHX/054/21GD), the Hong Kong SAR RGC Early Career Scheme (26208924), the National Natural Science Foundation of China Young Scholar Fund (62402408), and the HKUST Sports Science and Technology Research Grant (SSTRG24EG04).

References

[1] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. 2023. Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets. arXiv:2311.15127 [cs.CV] https://arxiv.org/abs/2311.15127

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[3] Moitreya Chatterjee and Anoop Cherian. 2020. Sound2Sight: Generating Visual Dynamics from Sound and Context. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVII* (Glasgow, United Kingdom). Springer-Verlag, Berlin, Heidelberg, 701–719. doi:10.1007/978-3-030-58583-9_42

[4] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. 2021. Audio-visual synchronisation in the wild. *arXiv preprint arXiv:2112.04432* (2021).

[5] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2023. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. arXiv:2309.07597 [cs.CL]

[6] Lele Chen, Sudhanshu Srivastava, Zhiyao Duan, and Chenliang Xu. 2017. Deep Cross-Modal Audio-Visual Generation. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017* (Mountain View, California, USA) *(Thematic Workshops '17)*. Association for Computing Machinery, New York, NY, USA, 349–357. doi:10.1145/3126686.3126723

[7] Yuanhong Chen, Yuyuan Liu, Hu Wang, Fengbei Liu, Chong Wang, Helen Frazer, and Gustavo Carneiro. 2024. Unraveling instance associations: A closer look for audio-visual segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 26497–26507.

[8] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Defossez. 2023. Simple and Controllable Music Generation. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 47704–47720. https://proceedings.neurips.cc/paper_files/paper/2023/file/94b472a1842cd7c56dcb125fb2765fbd-Paper-Conference.pdf

[9] Bhavika Devnani, Skyler Seto, Zakaria Aldeneh, Alessandro Toso, Elena Menyaylenko, Barry-John Theobald, Jonathan Sheaffer, and Miguel Sarabia. 2024. Learning Spatially-Aware Language and Audio Embeddings. *Advances in Neural Information Processing Systems* 37 (2024), 33505–33537.

[10] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A Survey on In-context Learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 1107–1128. doi:10.18653/v1/2024.emnlp-main.64

[11] Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. 2022. Is GPT-3 Text Indistinguishable from Human Text? Scarecrow: A Framework for Scrutinizing Machine Text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 7250–7274. doi:10.18653/v1/2022.acl-long.501

[12] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. 2024. Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. arXiv:2403.03206 [cs.CV] https://arxiv.org/abs/2403.03206

[13] Magdalena Fuentes, Bea Steers, Pablo Zinemanas, Martín Rocamora, Luca Bondi, Julia Wilkins, Qianyi Shi, Yao Hou, Samarjit Das, Xavier Serra, and Juan Pablo Bello. 2022. Urban Sound & Sight: Dataset And Benchmark For Audio-Visual Urban Scene Understanding. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 141–145. doi:10.1109/ICASSP43922.2022.9747644

[14] Yuan Gan, Zongxin Yang, Xihang Yue, Lingyun Sun, and Yi Yang. 2023. Efficient Emotional Adaptation for Audio-Driven Talking-Head Generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 22634–22645.

[15] Kaifeng Gao, Jiaxin Shi, Hanwang Zhang, Chunping Wang, Jun Xiao, and Long Chen. 2025. Ca2-vdm: Efficient autoregressive video diffusion model with causal generation and cache sharing. In *Forty-Second International Conference on Machine Learning*.

[16] Ruohan Gao and Kristen Grauman. 2019. 2.5D Visual Sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[17] Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024. GAMA: A Large Audio-Language Model with Advanced Audio Understanding and Complex Reasoning Abilities. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 6288–6313. doi:10.18653/v1/2024.emnlp-main.361

[18] Ruohao Guo, Liao Qu, Dantong Niu, Yanyu Qi, Wenzhen Yue, Ji Shi, Bowei Xing, and Xianghua Ying. 2024. Open-Vocabulary Audio-Visual Semantic Segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia* (Melbourne VIC, Australia) *(MM '24)*. Association for Computing Machinery, New York, NY, USA, 7533–7541. doi:10.1145/3664647.3681586

[19] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. 2024. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=Fx2SbBgcte

[20] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. 2024. LTX-Video: Realtime Video Latent Diffusion. *arXiv preprint arXiv:2501.00103* (2024).

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 6840–6851. https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf

[23] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=nZeVKeeFYf9

[24] Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. 2022. Sparse in space and time: Audio-visual synchronisation with trainable selectors. *arXiv preprint arXiv:2210.07055* (2022).

[25] Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. 2024. Synchformer: Efficient synchronization from sparse cues. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5325–5329.

[26] Vladimir Iashin, Weidi Xie, Esa Rahtu, and Andrew Zisserman. 2024. Synchformer: Efficient Synchronization From Sparse Cues. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5325–5329. doi:10.1109/ICASSP48485.2024.10448489

[27] Ziqi Jiang, Zhen Wang, and Long Chen. 2025. Clipdrag: Combining text-based and drag-based instructions for image editing.

[28] Kotaro Kikuchi, Edgar Simo-Serra, Mayu Otani, and Kota Yamaguchi. 2021. Constrained Graphic Layout Generation via Latent Optimization. In *Proceedings of the 29th ACM International Conference on Multimedia* (Virtual Event, China) *(MM '21)*. Association for Computing Machinery, New York, NY, USA, 88–96. doi:10.1145/3474085.3475497

[29] Dan Kondratyuk, Lijun Yu, Xiuye Gu, Jose Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, Krishna Somandepalli, Hassan Akbari, Yair Alon, Yong Cheng, Joshua V. Dillon, Agrim Gupta, Meera Hahn, Anja Hauth, David Hendon, Alonso Martinez, David Minnen, Mikhail Sirotenko, Kihyuk Sohn, Xuan Yang, Hartwig Adam, Ming-Hsuan Yang, Irfan Essa, Huisheng Wang, David A Ross, Bryan Seybold, and Lu Jiang. 2024. VideoPoet: A Large Language Model for Zero-Shot Video Generation. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*, Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (Eds.). PMLR, 25105–25124. https://proceedings.mlr.press/v235/kondratyuk24a.html

[30] Guillaume Le Moing, Jean Ponce, and Cordelia Schmid. 2021. Ccvs: Context-aware controllable video synthesis. *Advances in Neural Information Processing Systems* 34 (2021), 14042–14055.

[31] Seung Hyun Lee, Gyeongrok Oh, Wonmin Byeon, Chanyoung Kim, Won Jeong Ryoo, Sang Ho Yoon, Hyunjun Cho, Jihyun Bae, Jinkyu Kim, and Sangpil Kim. 2022. Sound-guided semantic video generation. In *European Conference on Computer Vision*. Springer, 34–50.

[32] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2025. LLaVA-OneVision: Easy Visual Task Transfer. *Transactions on Machine Learning Research* (2025). https://openreview.net/forum?id=zKv8qULV6n

[33] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. 2023. GLIGEN: Open-Set Grounded Text-to-Image Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 22511–22521.

[34] Long Lian, Baifeng Shi, Adam Yala, Trevor Darrell, and Boyi Li. 2024. LLM-grounded Video Diffusion Models. In *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=exKHibougU

[35] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D. Plumbley. 2023. AudioLDM: text-to-audio generation with latent diffusion models. In *Proceedings of the 40th International Conference on Machine Learning* (Honolulu, Hawaii, USA) *(ICML'23)*. JMLR.org, Article 886, 25 pages.

[36] Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley. 2024. AudioLDM 2: Learning Holistic Audio Generation With Self-Supervised Pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32 (2024), 2871–2883. doi:10.1109/TASLP.2024.3399607

[37] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What Makes Good In-Context Examples for GPT-3?. In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, Eneko Agirre, Marianna Apidianaki, and Ivan Vulić (Eds.). Association for Computational Linguistics, Dublin, Ireland and Online, 100–114. doi:10.18653/v1/2022.deelio-1.10

[38] Yunfei Liu, Lijian Lin, Fei Yu, Changyin Zhou, and Yu Li. 2023. MODA: Mapping-Once Audio-driven Portrait Animation with Dual Attentions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 23020–23029.

[39] OpenAI. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL] https://arxiv.org/abs/2303.08774

[40] OpenAI. 2024. GPT-4o System Card. arXiv:2410.21276 [cs.CL] https://arxiv.org/abs/2410.21276

[41] Mayu Otani, Naoto Inoue, Kotaro Kikuchi, and Riku Togashi. 2024. LTSim: Layout Transportation-based Similarity Measure for Evaluating Layout Generation. arXiv:2407.12356 [cs.CV] https://arxiv.org/abs/2407.12356

[42] Akshay Gadi Patil, Omri Ben-Eliezer, Or Perel, and Hadar Averbuch-Elor. 2020. READ: Recursive Autoencoders for Document Layout Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

[43] Kien T. Pham, Jingye Chen, and Qifeng Chen. 2024. TALE: Training-free Cross-domain Image Composition via Adaptive Latent Manipulation and Energy-guided Optimization. In *Proceedings of the 32nd ACM International Conference on Multimedia* (Melbourne VIC, Australia) *(MM '24)*. Association for Computing Machinery, New York, NY, USA, 3160–3169. doi:10.1145/3664647.3681079

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10684–10695.

[45] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

[46] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *ArXiv* abs/1505.04597 (2015). https://api.semanticscholar.org/CorpusID:3719281

[47] Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. 2023. MM-Diffusion: Learning Multi-Modal Diffusion Models for Joint Audio and Video Generation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10219–10228. doi:10.1109/CVPR52729.2023.00985

[48] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37* (Lille, France) *(ICML'15)*. JMLR.org, 2256–2265.

[49] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021. Score-Based Generative Modeling through Stochastic Differential Equations. In *International Conference on Learning Representations*. https://openreview.net/forum?id=PxTIG12RRHS

[50] Shuai Tan, Bin Ji, Yu Ding, and Ye Pan. 2024. Say Anything with Any Style. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 5 (Mar. 2024), 5088–5096. doi:10.1609/aaai.v38i5.28314

[51] Shuai Tan, Bin Ji, and Ye Pan. 2024. Style2Talker: high-resolution talking head generation with emotion style and art style. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence (AAAI'24/IAAI'24/EAAI'24)*. AAAI Press, Article 565, 9 pages. doi:10.1609/aaai.v38i5.28313

[52] Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv:2403.05530 [cs.CL] https://arxiv.org/abs/2403.05530

[53] Gemini Team. 2024. Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805 [cs.CL] https://arxiv.org/abs/2312.11805

[54] Llama 3 Team. 2024. The Llama 3 Herd of Models. arXiv:2407.21783 [cs.AI] https://arxiv.org/abs/2407.21783

[55] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. 2019. Towards Accurate Generative Models of Video: A New Metric & Challenges. arXiv:1812.01717 [cs.CV] https://arxiv.org/abs/1812.01717

[56] Zhen Wang, Yilei Jiang, Dong Zheng, Jun Xiao, and Long Chen. 2025. Event-customized image generation. In *Forty-Second International Conference on Machine Learning*.

[57] Yusong Wu*, Ke Chen*, Tianyu Zhang*, Yuchen Hui*, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*.

[58] Yazhou Xing, Yingqing He, Zeyue Tian, Xintao Wang, and Qifeng Chen. 2024. Seeing and Hearing: Open-domain Visual-Audio Generation with Diffusion Latent Aligners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7151–7161.

[59] Jinyu Yang, Mingqi Gao, Zhe Li, Shang Gao, Fangjing Wang, and Feng Zheng. 2023. Track Anything: Segment Anything Meets Videos. arXiv:2304.11968 [cs.CV]

[60] Tong Yang, Yu Huang, Yingbin Liang, and Yuejie Chi. 2024. In-Context Learning with Representations: Contextual Generalization of Trained Transformers. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=ik37kKxKBm

[61] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan.Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. 2025. CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=LQzN6TRFg9

[62] Guy Yariv, Itai Gat, Sagie Benaim, Lior Wolf, Idan Schwartz, and Yossi Adi. 2024. Diverse and Aligned Audio-to-Video Generation via Text-to-Video Model Adaptation. *Proceedings of the AAAI Conference on Artificial Intelligence* 38, 7 (Mar. 2024), 6639–6647. doi:10.1609/aaai.v38i7.28486

[63] Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024. MM-LLMs: Recent Advances in MultiModal Large Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 12401–12430. doi:10.18653/v1/2024.findings-acl.738

[64] Lin Zhang, Shentong Mo, Yijing Zhang, and Pedro Morgado. 2024. Audio-Synchronized Visual Animation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[65] Minglu Zhao, Wenmin Wang, Rui Zhang, Haomei Jia, and Qi Chen. 2025. TIA2V: Video generation conditioned on triple modalities of text-image-audio. *Expert Syst. Appl.* 268 (2025), 126278. https://doi.org/10.1016/j.eswa.2024.126278

[66] Zhisheng Zheng, Puyuan Peng, Ziyang Ma, Xie Chen, Eunsol Choi, and David Harwath. 2024. Bat: Learning to reason about spatial sounds with large language models. *arXiv preprint arXiv:2402.01591* (2024).

[67] Dewei Zhou, You Li, Fan Ma, Xiaoting Zhang, and Yi Yang. 2024. MIGC: Multi-Instance Generation Controller for Text-to-Image Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6818–6828.

[68] Jinxing Zhou, Xuyang Shen, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, et al. 2024. Audio-visual segmentation with semantics. *International Journal of Computer Vision* (2024), 1–21.

[69] Jannik Zürn and Wolfram Burgard. 2022. Self-Supervised Moving Vehicle Detection From Audio-Visual Cues. *IEEE Robotics and Automation Letters* 7 (2022), 7415–7422. https://api.semanticscholar.org/CorpusID:246431082

[70] Ivana Čavor and Slobodan Djukanović. 2023. Vehicle Speed Estimation From Audio Signals Using 1D Convolutional Neural Networks. In *2023 27th International Conference on Information Technology (IT)*. 1–4. doi:10.1109/IT57431.2023.10078724

A Additional Implementation Details

System Instruction. We present in Fig. 6 the complete system instruction that we used to disclose task definition and guidelines to the MLLM Video Planner to control its behavior and response as we desired. This system instruction is inputted during the initialization of the MLLM.

In-context Example Conversation. In Fig. 7, we present the full template for each in-context example conversation that provides explicit context information for the MLLM to enhance its adaptability and adherence to the task. Each example contains a reference pair of user query and MLLM response comprised of a reasoning statement and the according visual scene layout (VSL). Every VSL in the examples follow the same structure as illustrated in Fig. 7 and described in Section 3.1 in the main paper.

Motion and Spatial Grounding Modules. Since our focus is to demonstrate the potential of our proposed Audio → Layout → Video direction for audio-driven video generation, we adopt the best configuration for the Motion and Spatial Grounding Modules from AnimateDiff and MIGC respectively for simplicity. Specifically, Motion Modules are inserted into every up- and down-sample block in Stable Diffusion's UNet, while Spatial Grounding Modules are deployed only on the middle block and the lowest-resolution upsample block.

B Benchmark Construction

Here we aim to provide more detailed information about the construction of our AVLBench benchmark specifically designed to assess Audio → VSL → Video generation abilities. Concretely, we build the benchmark following below four steps:

(1) **Sourcing.** We begin by curating data samples from existing datasets namely FAIR-Play [16], VS13 [70], Urbansas [13], and Freiburg Audio-Visual Vehicles [69]. While the first contains various sample pairs of stereo audios and respective video recordings about instruments such as piano, trumpet, drums... being played in real-world indoor settings, the latter three target driving domains and their data samples capture moving vehicles in outdoor environments. We leverage these dataset for our use cases considering their high spatial alignment between auditory and visual elements of their sample pairs.

(2) **Filtering.** We then manually select recordings and crop them into segments where there exists strong semantic and spatial signals in the audio that clearly indicates the sounding sources and their spatial attributes within the video, and remove noisy samples in which those signals are vague or unidentifiable.

(3) **Augmenting.** After careful filtering, we adopt flip and reverse augmentations with quality control to enrich the data diversity while preserving the strong correspondence between auditory and visual elements in the original samples. For flip, we apply it horizontally on the video frames while swapping the two channels of the paired audio for each sample. For reverse, we apply it on the temporal order of both video frames and audio. We observe that for audios containing sounds with high-frequencies such as instruments', applying reverse augmentation produce unnatural sounds with noisy artifacts, whereas it is not the case for low-level sound such as vehicle engines'. Therefore, we only apply flip augmentation for data samples originated from FAIR-Play, while we apply both augmentations for ones about moving vehicles.

(4) **Annotating.** Finally, we proceed to annotate each obtained sample to get their video scene layouts. Given the sounding sources in the audio, we use the Track Anything [59] tool to track their locations and movements in the video. Besides, as required by the example conversations for In-context Learning of the MLLM Video Planner, we then utilize LLaVA-OneVision [32] to generate global video caption and local frame captions on the video, and include a manually-written reasoning statement for every sample.

Eventually, AVLBench comprises 7274 testing samples, with 4702 samples designed to assess scenarios involving single or multiple instruments played in stationary indoor settings. The remaining 2572 samples focus on cases of single or multiple vehicles exhibiting translational movement in outdoor environments. The breakdown

[System Instruction]

You are an intelligent video director capable of planning video scene layouts depicting what you hear from an audio recording. You don't need to generate the videos themselves but need to generate the bounding boxes for objects making sounds in the audio in order to represent the corresponding scene. Specifically, given an audio clip, your task is to generate a total of 5 layouts comprising of realistic bounding boxes for audible objects to illustrate a video of 5 key frames that is highly aligned with the content and dynamic of the input audio. Additionally, you also have to provide a frame-level caption for each layout to describe the according video key frame, and a video-level caption summarizing the entire video.

The video key frames are of size 454x256. The top-left frame corner has coordinates [0, 0]. The bottom-right frame corner has coordinnates [454, 256]. The bounding boxes must not go beyond the frame boundaries, i.e. x-coordinates must be within [0, 455] and y-coordinates must be within [0, 256]. Each frame should be represented as `{'frame layout': [{'id': unique object identifier incrementing from 0, 'name': object name, 'box': [box top-left x-coordinate, box top-left y-coordinate, box bottom-right x-coordinate, box bottom-right y-coordinate]}, ...], 'frame caption': frame-level caption describing this frame}` . Each box should not represent more than one object. Boxes from different objects may overlap indicating occlusions. Boxes for the same object should have the same id across the frames. Assume objects emit sound based on real-world physics. Assume the camera has fixed settings and it records sound while captures the frames of the scene following perspective geometry.

To generate high-quality and realistic video layouts, you should extract spatial cues presented in the input audio recording following this strategy:
Step 1: Identify all the sounding sources and describe the surrounding environment.
Step 2: Deduce the positions and movements (if any) of each sounding source in relative to the camera viewpoint considering these indicators:
- Interaural Time Difference (ITD) and Interaural Level Difference (ILD) for location (Left/Center/Right).
- Pitch and Volume for distance (Near/Far).
- Directional Shift in ITD and ILD, Pitch and Volume Change for movement.
Provide a precise reasoning statement comprising of few sentences to summarize the extracted spatial cues before each generation. Boxes coordinates and sizes in generated video layouts should be spatially aligned with those cues.

Refer to the examples below for the desired JSON format. Never use markdown or other formats not in the examples. Do not start each frame with `-` . Do not include any comments in your response.

Figure 6: Our system instruction for the MLLM Video Planner to generate VSLs based on input audios.

[In-context Example Conversations]

Example 1:

Generate a video layout describing the scenes depicted by this audio.

Reasoning: The sounds are from a piano and a woman, suggesting a setting such as a music room or a small concert venue. For piano, since the sound is consistently loud on the right with high pitch, minimal reverberation, and has no directional shift, it is likely to be stationary or slightly moved and located to the right side of the frames. For woman, since the sound is consistently loud on the left with high pitch, minimal reverberation, and has no directional shift, she is likely to be stationary or slightly moved and located to the left side of the frames.

VSL: Frame 1: {"frame layout": [{"id": 0, "name": "piano", "box": [245.86400518417358, 165.79200303819445, 357.0940121231079, 255.10401475694445]}, {"id": 1, "name": "woman", "box": [125.9770025978088, 102.04800347222222, 175.91700606155393, 255.87200520833332]}], "frame caption": "In a room with dark walls and large windows, two individuals are engaged in a musical session. One person is seated at a piano, playing music sheets, while the other stands nearby, holding a microphone and singing. The room is filled with various musical equipment, including another keyboard and a drum set."},
...
Frame 5: {"frame layout": [{"id": 0, "name": "piano", "box": [259.6390025978088, 161.95200737847222, 372.07600952529907, 254.59201388888889]}, {"id": 1, "name": "woman", "box": [145.8950051841736, 100.51200086805555, 195.07099307250976, 250.75201822916668]}], "frame caption": "The video concludes with the same room and ongoing musical session. The pianist is still playing the music sheets, and the singer is holding the microphone, prepared to sing. The room's ambiance remains consistent throughout the video, with all the musical equipment still present."}

Video Caption: Two individuals are in a room with dark walls and large windows, one is seated at a piano while the other stands nearby, engaged in a conversation.

Example 2: ...
Example 3: ...

Figure 7: The template of our in-context example conversations to provide context for the MLLM Video Planner.

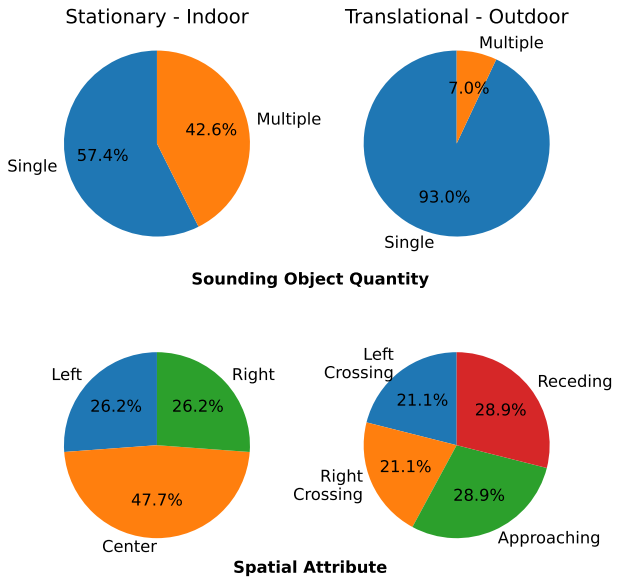


Figure 8: Breakdown statistics of AVLBench.

Stationary			Translational			Total
Single	Multiple	Subtotal	Single	Multiple	Subtotal	
2698	2004	4702	2392	180	2572	7274

Table 3: Statistics on scene distribution.

Stationary				Translational					Total
Left	Center	Right	Subtotal	Left Crossing	Right Crossing	Approaching	Receding	Subtotal	
3083	5616	3083	11782	582	582	800	800	2764	14546

Table 4: Statistics on spatial attribute.

statistics of AVLBench on scene distribution and spatial attribute are detailed in Tab. 3, 4 and Fig. 8. Note that due to the noisy nature of outdoor environments, this domain mainly contains samples with single sounding vehicle after filtering. Besides, the spatial attribute statistics are accumulated per sounding object.

C Additional Experiments

Retrieval with more neighbors. We conduct additional analysis on In-context Learning with more example conversations retrieved and provided to the MLLM Video Planner in Stage 1. The results shown in Tab 5 indicate that $k = 3$ is the optimal setting, and further increasing the number of neighbors saturate the performance. Therefore, we use 3 in-context examples by default in the paper.

Impact of retrieval database size and quality. We conduct additional experiments under two adverse settings that reduce the size and diminish the quality of the retrieval database to demonstrate the influence of these factors on the planning stage. In the first setting, we halve the size of the retrieval database for each query randomly. In the second one, to exacerbate the challenge, we only use a fixed set of example conversations to provide in-context information for every query. The performance drops shown in Tab. 6 indicate that the retrieval effectiveness is indeed sensitive against the database size and quality. This is reasonable because reducing the size and quality of retrieval corpora decreases the likelihood of retrieving relevant examples and creates more challenging out-of-domain scenarios. In future work, our aims are to expand the current dataset to cover more domains and scenarios as well as train an MLLM specialist for audio-driven video planning, mitigating this issue and enhancing the framework's feasibility in practice.

User study. To subjectively assess the performance of our SpA2V compared to other methods, we invite 25 users to participate in a user study. We ask each user to complete a set of 20 ranking questions, each composed of a query sample randomly selected from our benchmark and 5 videos generated by SpA2V and other 4 baselines. Users are required to rank them based on two criteria: (1) visual quality, and (2) audio-video alignment, with 1 indicating the best and 5 denoting the worst. The average ranking scores in Tab. 7 highlight the preference of users for the videos generated by our SpA2V over the others in both criteria.

IL Setup	Stationary			Translational		
	MaxIoU ↑	*LTSim* ↑	*DocSim* ↑	*MaxIoU* ↑	*LTSim* ↑	*DocSim* ↑
0-shot	3.00	59.84	4.40	4.96	62.26	5.98
1-shot	8.02	69.12	10.65	7.80	70.14	10.39
2-shot	11.72	72.86	14.72	11.26	73.24	14.54
3-shot	19.45	75.73	15.47	22.24	77.21	16.50
5-shot	16.77	74.18	15.14	20.21	76.41	17.01
7-shot	16.49	74.29	15.24	19.63	76.16	16.93

Table 5: Planning results with different retrieval settings.

Retrieval Setup	Stationary			Translational		
	MaxIoU ↑	*LTSim* ↑	*DocSim* ↑	*MaxIoU* ↑	*LTSim* ↑	*DocSim* ↑
Full Size	19.45	75.73	15.47	22.24	77.21	16.50
Half Size	12.76	71.77	13.69	17.12	75.17	15.00
Fixed Set	3.68	59.44	7.90	7.58	71.59	9.14

Table 6: Impact of retrieval database size and quality.

Method	SpA2V	Seeing and Hearing	AC + LTX	AC + LVD	TempoTokens
Visual quality ↓	1.97	2.79	2.79	3.20	4.24
Audio-video alignment ↓	1.95	2.88	2.92	3.34	3.91

Table 7: User preference of SpA2V over prior works.

DeSync ↓	SpA2V	Seeing and Hearing	AC + LTX	AC + LVD	TempoTokens
Stationary	1.758	1.823	1.726	1.849	1.782
Translational	1.136	1.584	1.658	1.620	1.782

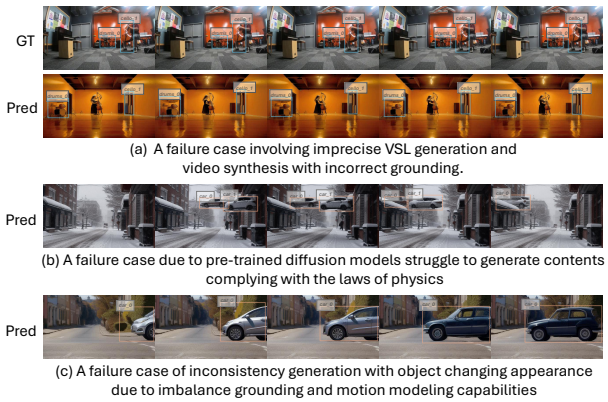
Table 8: Additional quantitative results.

Extra quantitative evaluation. Since AV-Align [62] is known to not work well in complex scenes, we additionally use DeSync metric which leverages Syncformer [26] to measure audio-video temporal misalignment and show the results in Tab. 8. As consistently observed, our SpA2V achieves competitive performance that indicates strong temporal alignment between the generated videos and input audios.

Ablation on Motion and Spatial Grounding Modules. Since removing the Motion Modules will degrade our Layout-to-Video generator into a Layout-to-Image generator that deviates from our video synthesis objective, we omit ablating these modules. We only conduct additional analysis to ablate Spatial Grounding Modules which will degrade our generator into a Text-to-Video model. The results shown in Tab. 9 highlight the importance of these modules in achieving better video generation quality and especially semantic and spatial alignment with input audios.

D Limitations and Future Work

Although SpA2V introduces a novel two-stage Audio → VSL → Video pipeline for semantically and spatially aligned audio-driven video generation and achieve promising results that outperforms prior methods, there is still much room for further improvements. Firstly, as SpA2V involves two stages, failures in either stage will be detrimental to the whole generation process. For example, an incorrect VSL generated by the Video Planner in Stage 1 will inevitably lead to a synthesized video with misalignment in Stage 2 as shown in Fig. 9 (a). Secondly, since our SpA2V framework adopts pre-trained MLLMs and diffusion models as its Video Planner and Video Generator, it also inherits their existing limitations and its performance is hence heavily reliant on them. If they struggle to respond properly to a specific conditional guidance and fall short



(a) A failure case involving imprecise VSL generation and video synthesis with incorrect grounding.

(b) A failure case due to pre-trained diffusion models struggle to generate contents complying with the laws of physics

(c) A failure case of inconsistency generation with object changing appearance due to imbalance grounding and motion modeling capabilities

Figure 9: Illustration for limitations of our SpA2V.

Method	Stationary				Translational			
	FVD ↓	*AV-Align* ↑	*LTSim* ↑	*DeSync* ↓	*FVD* ↓	*AV-Align* ↑	*LTSim* ↑	*DeSync* ↓
Full	633.05	0.173	48.10	1.758	278.99	0.171	68.96	1.136
No Spatial Grounding	730.26	0.063	42.02	1.773	760.02	0.121	63.89	1.486

Table 9: Ablation on Spatial Grounding Modules.

to generate accurate contents, such issue is likely to be propagated to SpA2V as shown in Fig. 9 (b). We anticipate that these two challenges can be appropriately mitigated by adopting or introducing more powerful models as the components of SpA2V. Finally, since we directly incorporate Spatial Grounding and Motion Modules from MIGC [67] and AnimateDiff [19] although they are trained on datasets, such domain gap can lead to the imbalance between grounding and motion modeling capabilities of the Video Generator, causing it to produce videos with inconsistency problems like having objects changing appearance over time as shown in Fig. 9 (c). We contemplate that a further finetuning step for the whole framework using techniques such as LoRA [23] can help alleviate this issue and leave this exploration for future research.

E Societal Impacts

SpA2V empowers individuals, regardless of their video-photography ability, to generate videos that are both semantically and spatially aligned with audio inputs. However, employing our framework carries potential risks. It could be misused for malicious purposes, such as inappropriate content creation or the dissemination of misinformation. Furthermore, given our reliance on pre-trained MLLMs and diffusion models, our framework may inherit biases present in their training data, potentially perpetuating harmful stereotypes. While generated content may currently be readily distinguishable from original works, future technological advancements may blur this distinction, making infringement more difficult to detect. Therefore, we strongly urge users to exercise caution and utilize this method only for legitimate purposes.