Listen and Count: Expanding the Frontier of Zero-shot Object Counting to Sound-centric Counting

Mingjie Wang[1*], **Zhuohang Li**[1], **Qi Zhang**[2], **Jun Zhou**[3], **Zili Yi**[4] and **Minglun Gong**[5]
[1]Department of Mathematics, Zhejiang Sci-Tech University, Hangzhou, China
[2]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
[3]School of Information Science and Technology, Dalian Maritime University, Dalian, China
[4]School of Intelligence Science and Technology, Nanjing University, Nanjing, China
[5]School of Computer Science, University of Guelph, Guelph, Canada

Abstract

While class-agnostic object counting has recently evolved from image-exemplar to language-guided paradigms, existing methods are limited by text polysemy and the lack of prompts in audio-sensing scenarios. To overcome these challenges, we introduce a sound-centric counting paradigm, enabling models to "*listen and count*" using the vivid and discriminative signatures of sound cues. We propose S2ICount, which achieves seamless, fine-grained audio-visual alignment for sound-guided counting. At its core, SoundMamba leverages linear-complexity long-range semantic modelling to align cross-modal features, while the Sound Calibration Module designs a circular scanning mechanism to progressively reconcile modality discrepancies using multi-scale cues from a Stable Diffusion backbone. To enhance saliency awareness, a Multi-tier Hybrid Optimization strategy enforces consistency across magnitude, semantics, and distribution. In addition, we introduce SoundCount, the first large-scale sound-oriented counting dataset, comprising over 2,000 scenes and 1,650 audio sources across 66 object categories. Extensive experiments demonstrate that S2ICount achieves state-of-the-art performance, validating the effectiveness of sound as a robust guidance modality for counting. The dataset, demo and implementation code are made publicly accessible at https://github.com/ZSTU-CV-Lab/S2ICount.

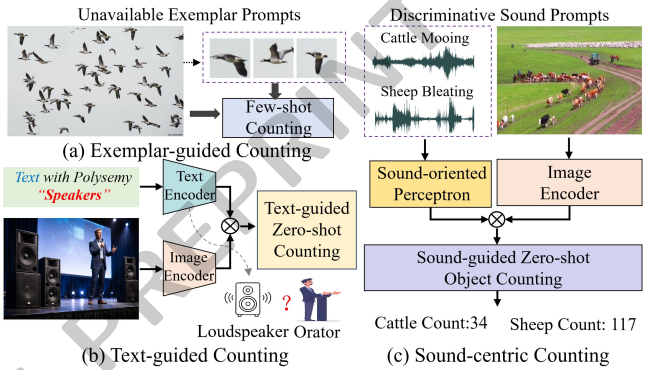

Figure 1: (a) Exemplar-based counting methods explicitly require user-provided exemplar prompts to specify the classes cues. Such image exemplar are often labour-intensive to acquire and unavailable in practical scenarios, which severely limits their robustness and generalization ability; (b) Text-guided variants introduce text prompt as a more flexible mode of interaction. While this improves practicality and effectiveness, the inherent semantic ambiguity arising from textual polysemy remains intractable; (c) To address the challenges of ambiguity and unavailability in audio-sensing applications, we pioneer the use of sound modality as a guiding cue for multimodal counting and propose a novel sound-centric paradigm.

1 Introduction

Driven by the success of deep learning, object counting has been extensively studied [Lempitsky and Zisserman, 2010; Sindagi and Patel, 2018; Wang *et al.*, 2022] and applied to diverse scenarios, including security monitoring, traffic management, and agricultural yield estimate. Early work primarily focuses on class-specific tasks, such as crowd [Meng *et al.*, 2025; Lin *et al.*, 2025; Zhang *et al.*, 2024] and vehicle counting [Mundhenk *et al.*, 2016; Marsden *et al.*, 2018; Zhang *et al.*, 2017]. Despite high accuracy, these methods rely on labour-intensive datasets with limited categories and often generalize poorly to unseen classes. To address this, recent approaches [Jiang *et al.*, 2023; Kang *et al.*, 2024] aim for class-agnostic counting, enabling arbitrary-category counting without extensive annotations. As intrinsic class cues are absent, class-agnostic algorithms necessitate auxiliary guidance prompts, typically either image exemplars [Ranjan *et al.*, 2021; Shi *et al.*, 2022] or linguistic names [Jiang *et al.*, 2023; Kang *et al.*, 2024]. Specifically, as shown in Fig. 1 (a), traditional exemplar-based methods input raw images with exemplars to count specific objects, whereas recent multimodal approaches leverage easily-acquired language prompts to specify target categories, see Fig. 1 (b).

Therefore, language-vision multimodal counting has gained increasing attention, fuelled by large-scale pre-trained text encoders (*e.g.*, CLIP [Radford *et al.*, 2021] and BERT [Devlin *et al.*, 2019]). Text-guided methods [Amini-Naieni *et al.*, 2023; Pelhan *et al.*, 2024; Qian *et al.*, 2025] achieve strong performance by mining the generalization

of these models and aligning text-image feature spaces via contrastive learning. Despite enabling compelling human-machine interaction, text-centric approaches are limited in scenarios where language prompts are unavailable, such as audio-sensing robotics, automatic movie subtitle generation, or assistive applications for visually impaired users, where only sound signals (*e.g.*, horse neighing or sheep bleating) are present. Moreover, sound conveys richer and more discriminative information than text, which often suffers from semantic ambiguity caused by polysemy (*e.g.*, "Speaker" as loudspeaker *v.s.* orator), whereas sound signals provide distinctive acoustic signatures. Motivated by these observations, it is natural to equip counting agents with the ability to *listen and count* in a human-like manner, enhancing their capacity to perceive and interpret real-world environments. For example, steering robotic agents to monitor wildlife density based solely on animal vocalizations. Motivated by recent advances in multimodal feature learning, particularly pre-trained audio-vision models (*e.g.*, SpeechCLIP [Shih *et al.*, 2023], AudioCLIP [Guzhov *et al.*, 2022]), audio-guided vision tasks [Yang *et al.*, 2025; Wang *et al.*, 2025; Sung-Bin *et al.*, 2025] have emerged as a promising direction in multimodal research. Building on these developments, we pioneer the use of sound as a guidance in object counting, introducing sound-centric class-agnostic counting, see Fig. 1 (c).

Accordingly, we propose S2ICount, a novel sound-guided counting protocol to explore sound-image collaboration. At a high level, S2ICount realizes an automatic pipeline of "*receiving*, *perceiving*, *specifying*, and *counting*" under sound guidance by bridging natural audio signals and visual scenes. At a finer scale, it comprises three key components: SoundMamba, Sound Calibration Module (SCM), and Multi-tier Hybrid Optimization scheme. SoundMamba aligns cross-modal sound-image semantics using textual prompts as a hub, leveraging Mamba's efficiency in modelling long-range dependencies. SCM progressively calibrates cross-modal discrepancies via a novel circular-scanning mamba module and multi-scale features from a Stable Diffusion backbone. Multi-tier Hybrid Optimization enforces comprehensive alignment in magnitude, semantics, and distribution during training. Beyond the protocol, we curate SoundCount, a large-scale sound-image dataset of over 2,351 scenes and 1,650 sound sources spanning 66 categories, addressing the scarcity of audio-informed counting data and providing a foundation for training, benchmarking, and deploying sound-guided counting. In summary, the contributions of this work are fourfold:

- **Modality Breakthrough**. This work expands the frontier of class-agnostic object counting for both the research or industry by pioneering sound guidance, a novel modality has thus far remained largely unexplored.
- **Novel Sound-centric Counting Protocol.** We propose S2ICount, a novel sound-centric counting protocol designed to allow models to *listen and count* in a human-like manner by estimating instance counts specified by sound stimulus in audio-sensing scenarios.
- **Fine-grained Feature Alignment.** To achieve seamless cross-modal alignment between sound and image inputs, we architecture three sophisticated components including a SoundMamba, a Sound Calibration Module (SCM), and a Multi-tier Hybrid Optimization scheme.
- **New Dataset and SOTA Performance.** We further curate a large-scale multimodal dataset (SoundCount) to catalyze studies and applications in novel sound-centric counting protocol. Extensive experiments illustrate the state-of-the-art performance of our S2ICount, proving the efficacy of sound-image synergy for object counting.

2 Related Work

2.1 Class-agnostic Multimodal Counting

Recently, research on class-agnostic multimodal object counting has shifted from exemplar-based, single-modal approaches to linguistic-guided multimodal methods, which take raw images together with class-specific text prompts as input and achieving highly interactive counting without requiring difficult-to-acquire exemplar prompts. Earlier method like Patch-selection [Xu *et al.*, 2023] utilizes VAEs to translate text prompts into visual exemplars, thereby recasting the text-guided task into a few-shot exemplar-based counting problem. To facilitate direct textual guidance independent of intermediate visual exemplars, a suite of recent approaches, such as CLIP-Count [Jiang *et al.*, 2023], VLCounter [Kang *et al.*, 2024], and CountX [Amini-Naieni *et al.*, 2023], harnesses the rich priors of the widely adopted pre-trained CLIP model. By directly aligning textual features with visual characteristics, these models achieve efficient text-guided counting with superior interactivity and generalization robustness. In addition to CLIP, other large-scale vision-language models have also been pivoted toward text-guided counting tasks. For instance, PseCo [Huang *et al.*, 2024] delves into the use of the Segment Anything Model (SAM) [Kirillov *et al.*, 2023] by automatically generating candidate dense regions, while VA-Count [Zhu *et al.*, 2024] capitalizes on the strong localization capabilities of Grounding DINO [Liu *et al.*, 2024b] to obtain high-quality visual exemplar patches via text-driven detection. More recently, the powerful Stable Diffusion models has been integrated into T2ICount [Qian *et al.*, 2025], providing fine-grained attention maps and multi-scale features to improve multimodal counting accuracy.

2.2 Audio-steering Multimodal Models

While existing text-guided protocols endeavour to mimic human logic of write and count, they are vulnerable to intractable polysemy of textual descriptions and fail in environments restricted to auditory stimuli. Recent breakthroughs in other audio-visual tasks have tried to address these challenges by leveraging the unique acoustic fingerprints inherent in sound signals. Specifically, AVLNet [Rouditchenko *et al.*, 2020] and SoundAdapter [Qin *et al.*, 2023] achieve audio-driven cross-modal retrieval and generation, respectively, by bridging the latent gap between audio and vision inputs. Similarly, foundational models like SpeechCLIP [Shih *et al.*, 2023], ImageBind [Girdhar *et al.*, 2023], CoDi [Tang *et al.*, 2023] and Sound2Scene [Sung-Bin *et al.*, 2023] generalize the CLIP principle to incorporate speech and environmental sounds via multi-way contrastive learning to align audio-text-image feature spaces. The maturation of these audio-centric

models catalyzes the transition toward sound-guided object counting, offering a pathway to resolve modality limitations in real-world multimodal counting agents.

3 Listen and Count via S2ICount

3.1 An Efficient Sound Adapter: SoundMamba

Training large-scale sound-vision counting models from scratch is extremely challenging due to the scarcity of well-paired data, especially across diverse audio types. Fortunately, recent large-scale pre-trained multimodal models, such as CLIP for text-image, AudioCLIP [Guzhov *et al.*, 2022] for audio-text-image, and CLAP [Wu *et al.*, 2023] for audio-text alignment, provide a practical, cost-effective means to extract rich cross-modal features. Notably, CLAP, trained on eight datasets, offers the largest audio-text corpus and broad sound coverage to date. Leveraging its strong generalization, we adopt a frozen CLAP as our sound encoder to generate coarse sound features for the counting task.

Specifically, given a raw input sound S, the CLAP encoder first extracts coarse time-frequency representations. However, since we employ a CLIP-based Stable Diffusion (SD) backbone to facilitate acoustic guidance, the inherent incompatibility between the CLAP and CLIP feature spaces prevents direct integration, as raw sound embeddings would lead to significant semantic misalignment. To address this, we design an efficient adapter, SoundMamba, to project and precisely align sound representations into the CLIP-friendly textual embedding space required for subsequent diffusion workflows. SoundMamba sequentially integrates layer normalization, 1D convolution, GELU activation, and linear projection, followed by several Mamba blocks. To stabilize feature and gradient propagation, we incorporate residual learning throughout the adapter. The resulting embeddings are fused with local features derived from a MLP to achieve a synergy of local and global semantic cues, ultimately yielding a 768-dimensional sound latent. The rationale for employing Mamba blocks over traditional Transformers lies in their ability to capture long-range temporal dependencies while maintaining linear computational complexity.

3.2 Sound Calibration Module

Diffusion models [Zhang *et al.*, 2025] capture high-dimensional data distributions through iterative denoising steps and have shown significant success across diverse generative benchmarks. Utilizing its superior image representation capability, the Stable Diffusion (SD) model [Rombach *et al.*, 2022] is adopted as the backbone for sound-guided visual feature extraction. Specifically, the input image is first encoded by a VAE into a compact latent manifold, after which a pre-defined UNet operates on this latent while being steered by the well-aligned sound embeddings from SoundMamba. To optimize a trade-off between computational efficiency and counting accuracy, we extract one-step diffusion features as the final image representations. To gain fine-grained alignment between the sound latent and visual features across multiple scales, this paper introduces a Sound Calibration Module (SCM). Specifically, visual feature maps at four different scales are captured from distinct hierarchical levels of the UNet within the SD backbone, which are denoted as:

$$F_i \in R^{\frac{H}{2^{i-1}} \times \frac{W}{2^{i-1}}}, \quad i \in \{1, 2, 3, 4\} \tag{1}$$

where H and W represent the spatial resolutions of the 2D visual feature maps.

The SCM recursively and hierarchically aggregates visual feature maps in a coarse-to-fine manner, progressively fusing information across increasing spatial resolution. This inter-scale transition is formally defined as:

$$F_i' = \texttt{Linear}(\texttt{Cat}(\texttt{Linear}(\texttt{Up}(P_{i+1})), F_i)), \quad i \in \{1, 2, 3, 4\} \tag{2}$$

where F_i' denotes the i_{th}-level fused characteristics, and P_{i+1} represents the feature map propagated from the preceding lower-resolution level (with the base case $P_4 = F_4$). The operator $\texttt{Up}(\cdot)$ performs spatial upsampling, $\texttt{Cat}(\cdot)$ signifies channel-wise concatenation, and $\texttt{Linear}(\cdot)$ is utilized for feature projection and dimensional reduction.

To better compute similarity metrics between the sound latent and visual pixels across multi-scale feature maps, a Cross-attention Mechanism (CM) imposed on the fused visual features F_i' and the sound latent s is a widely adopted approach. However, directly applying (CM) to locality-only F_i' and global s tends to overlook correlations between the global distributions of the visual image and the sound token, making it difficult to handle the density shift issues in sound-vision counting problems. Therefore, in this paper, we propose Global Cross-attention Mechanism (GCM) that explicitly models global vision instance distribution across the entire scene by introducing a novel circular-scanning Mamba module to transform local features F_i' into global representations G_i, while maintaining low computational overhead. The pipeline of GCM can be formulated as:

$$G_i = M(F_i'),\ P_i, S = CM(G_i, s), \tag{3}$$

where CM stands for the conventional cross-attention operation, and M denotes the circular-scanning Mamba module proposed in this paper. P_i and S represent the image and audio features obtained after processing by the GCM, respectively.

Specifically, the circular scanning strategy is designed to impose a circular traversal over pixels in two-dimensional images, in both clockwise and counterclockwise directions. This design is motivated by the observation that object instances in natural scenes often exhibit circular or radial distribution patterns inherently caused by radial camera perspective, as illustrated by the cattle examples in Fig. 2. Compared with naïve horizontal or vertical scanning, circular scanning is more effective at capturing global distribution characteristics and continuous instance density variations. First, the image center (x_c, y_c) is computed based on the row and column indices. Using a polar-coordinate transformation, each pixel (x, y) is then assigned to a discrete radial layer $R(x, y)$ according to its distance from the image center.

$$R(x, y) = \left\lfloor K \times \sqrt{(x - x_c)^2 + (y - y_c)^2} \right\rfloor, \tag{4}$$

where K denotes the radial discretization coefficient, whereas $\lfloor \cdot \rfloor$ means the floor operation. Meanwhile, the normalized angular position $\hat{\theta}(x, y)$ of each pixel within its corresponding

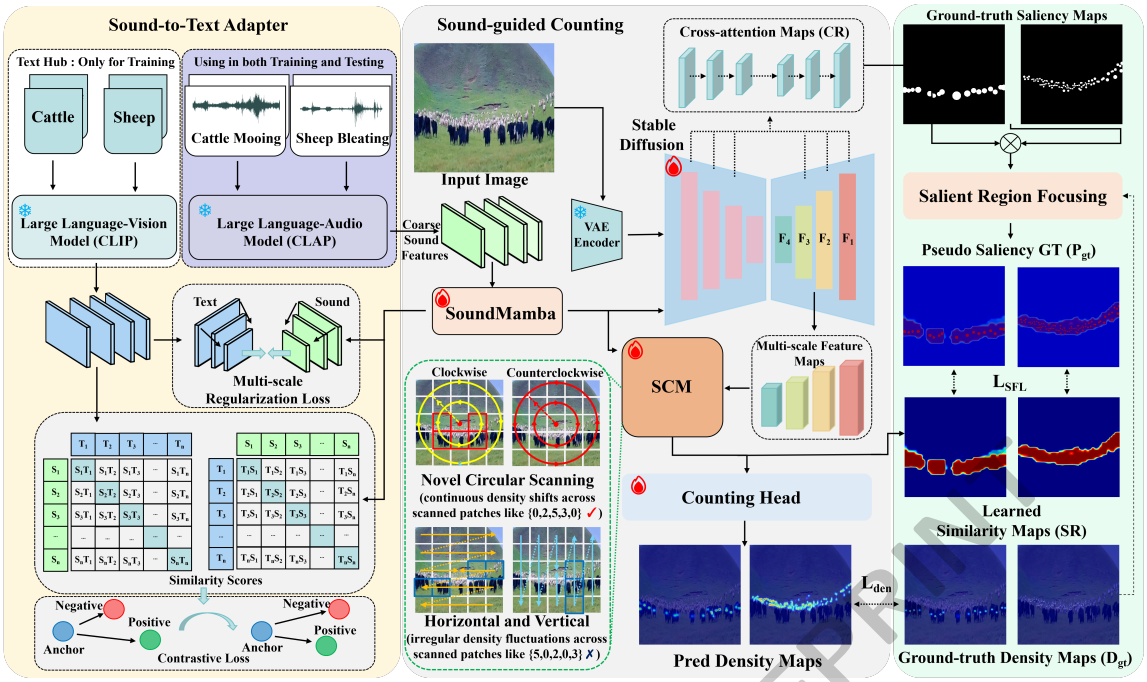

Figure 2: The proposed *S2ICount* mainly consists of: *i*) SoundMamba, an efficient adapter engineered to project coarse audio representations into fine-grained conditions compatible with diffusion model; *ii*) a Sound Calibration Module (SCM), which bridges cross-modal discrepancies by integrating a novel circular-scanning Mamba with multi-scale features from the diffusion backbone; and *iii*) a Multi-tier Hybrid Optimization scheme, designed to ensure comprehensive sound-text multimodal consistency in terms of magnitude, semantics, and distribution. By orchestrating these components, we establish a robust and effective paradigm for sound-centric object counting.

radial layer is computed to characterizes its relative polar angle with respect to the image center.

$$\hat{\theta}(x,y) = \frac{\arctan 2(y-y_c, x-x_c) + \pi}{2\pi}. \tag{5}$$

Based on the radial and angular components, we further define the pixel priority $O(x,y)$ in the proposed circular scanning scheme as follows:

$$O(x,y) = R(x,y) \cdot M + \hat{\theta}(x,y), \tag{6}$$

where M is a magnitude isolation constant satisfying $M \gg 1$, ensuring that pixels from different radial layers are strictly ordered while preserving the angular ordering within each layer. According to this priority value, a deterministic circular scanning order over the entire image can be obtained.

3.3 Multi-tier Hybrid Optimization

We introduce a hybrid supervision scheme that enables reliable, semantic-, and distribution-invariant projection. Specifically, the proposed hybrid optimization integrates a Multi-scale Regularization Loss to enforce magnitude consistency, a Contrastive Loss to preserve semantic coherence, and a Saliency-intent Filtering Loss to attain distribution alignment.

Multi-scale Regularization Loss

Our preliminary findings indicate that a strict single-scale MSE loss fails to reconcile the multi-granular discrepancies between sound and text features, resulting in brittle adaptation. Consequently, we propose a hierarchical regularization scheme for magnitude consistency that enforces alignment invariance of feature segments across multiple scales, leading to more semantically robust and well-calibrated cross-modal representation alignment. Specifically, we employ a bank of average pooling operators AVP with varying kernel sizes $K \in \{1,3,5,7,9,11,13\}$ and corresponding strides $S \in \{1,2,4,6,8,10,12\}$ to perform multi-scale temporal aggregation. This operation yields a hierarchical set of embeddings captured at diverse granularities. Hence, the multi-scale regularization loss is formally defined as follows:

$$\mathcal{L}_{\text{mse}} = \sum_{K} \left\| \text{AVP}_K(\text{Sound}) - \text{AVP}_K(\text{Text}) \right\|^2, \tag{7}$$

where Sound indicates the latent audio representations from SoundMamba, and Text signifies the target text embeddings from the frozen CLIP text encoder.

Positive-Negative Contrastive Loss

To preserve the semantic consistency across multimodal features and effectively model their intrinsic correlations, we incorporate a Positive-Negative Contrastive Loss to achieve holistic alignment between the audio and text modalities. Specifically, we adopt an InfoNCE [Radford *et al.*, 2021] based objective to promote high-level semantic coherence by pulling positive sample pairs closer while pushing negative pairs farther apart in the embedding space.

Given a batch of n feature pairs, positive and negative labels are constructed in a self-supervised in-batch manner. For each instance i, the sound-text and text-sound embeddings pairs $\{s_i, t_i\}$ and $\{t_i, s_i\}$ with matching batch indices

are designated as positive pairs. Conversely, all mismatched combinations $\{s_i, t_j\}$ and $\{t_i, s_j\}$ (where $j \neq i$) serve as negative samples. Therefore, a symmetric contrastive loss is formulated to optimize bidirectional similarity scores over all positive and negative sample pairs within the current training batch, defined as follows:

$$\mathcal{L}_{\text{sp}} = \frac{1}{2}\Bigg\{ -\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{e^{(Sim(s_i,t_i)\cdot e^{\tau})}}{\sum_{j=1}^{n} e^{(Sim(s_i,t_j)\cdot e^{\tau})}}\right) - \frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{e^{(Sim(t_i,s_i)\cdot e^{\tau})}}{\sum_{j=1}^{n} e^{(Sim(t_i,s_j)\cdot e^{\tau})}}\right)\Bigg\}, \tag{8}$$

where $Sim(\cdot,\cdot)$ represents the cosine similarity measure, and e^{τ} denotes a temperature parameter used to scale the similarity values. In summary, to ensure the embedding consistency in both magnitude and semantic, the final objective function for optimizing the SoundMamba during the training phase is formulated as $\mathcal{L}_{\text{adap}} = \mathcal{L}_{\text{mse}} + \mathcal{L}_{\text{sp}}$.

Saliency-intent Filtering Loss

The fidelity of the similarity maps computed within the SCM is pivotal to the performance in prompt-guided counting pipelines. However, precisely isolating prompt-specific regions remains highly challenging in practice due to intensive background clutter and the semantic interference present in scenes with multiple object classes. To facilitate accurate sound-guided counting, we introduce a Salience-intent Filtering Loss (SFL) that enforces consistency between the diffusion-based attention maps and the learned similarity maps. This supervises both components to converge upon the correct sound-conditioned salient regions, effectively filtering out irrelevant background and inter-class distractors.

Specifically, cross-attention maps at multiple scales are first extracted from the diffusion model and then fused to form a unified attention map CR. The learned similarity map SR can be readily obtained from the SCM subnetwork. However, directly minimizing the distance between CR and SR is prone to training collapse due to the absence of explicit ground-truth supervision. To provide reliable supervision cues for learning saliency-aware maps, we use the ground-truth density map D_{gt} to construct an intermediate pseudo-saliency map P^{gt} as:

$$P^{\text{gt}}_{p,q} = \begin{cases} 1 & if \quad D^{gt}_{p,q} \geq \phi \\ 0 & if D^{gt}_{p,q} \leq 0.5 \times Mean(D^{gt}), CR_{p,q} \leq \xi, \end{cases} \tag{9}$$

where (p, q) is spatial position, and ϕ and ξ are heuristically-predefined thresholds. Then, the saliency-intent filtering loss is then written as $\mathcal{L}_{\text{SFL}} = \alpha\mathcal{L}_{\text{sal}} + \mathcal{L}_{\text{back}}$ ($\mathcal{L}_{\text{sal}} = \sum_{pq} 1 - SR_{p,q}$ if $P^{gt}_{p,q} = 1$, $\mathcal{L}_{\text{back}} = \sum_{pq} \max(0, SR_{p,q})$ if $P^{gt}_{p,q} = 0$) where α denotes the weight of the $\mathcal{L}_{\text{sal}}$. The overall hybrid optimization function for training S2ICount is given by $\mathcal{L} = \mathcal{L}_{\text{den}} + \mathcal{L}_{\text{adap}} + \lambda\mathcal{L}_{\text{SFL}}$, where λ denotes the weight of the $\mathcal{L}_{\text{SFL}}$.

4 SoundCount: A New Multimodal Dataset

4.1 Dataset Overview

Datasets	Image#	Class	Modality
ShanghaiTech [Zhang *et al.*, 2016]	1,198	1	Image
QNRF [Idrees *et al.*, 2018]	1,535	1	Image
NWPU [Wang *et al.*, 2020]	5,109	1	Image
JHU [Sindagi *et al.*, 2020]	4,372	1	Image
CARPK [Hsieh *et al.*, 2017]	1,448	1	Image
FSC-147 [Ranjan *et al.*, 2021]	6,135	147	Text-Image
SoundCount (ours)	**2,351**	**66**	**Audio-Image**

Table 1: Comparisons with prevailing object counting datasets.

Existing visual counting benchmarks are limited by narrow category coverage (*e.g.*, pedestrians, vehicles, cells) or single-class-dominated images, and critically, they lack sound annotations, hindering progress in audio-guided counting. For instance, FSC-147 [Ranjan *et al.*, 2021] includes multi-class scenes but no sound labels due to high annotation costs. To overcome these limitations, we introduce SoundCount, a large-scale audio-image paired dataset tailored for sound-guided counting. As shown in Fig. 3, SoundCount comprises 2,351 scenes across 66 categories, from wildlife and musical instruments to transportation. Of these, 735 are single-class images, while the rest depict complex multi-class scenes with an average of 58 instances per image, ranging from 1 to 2,092. The dataset's high density and background complexity make it a challenging benchmark. Comparisons with existing counting datasets are summarized in Tab. 1.

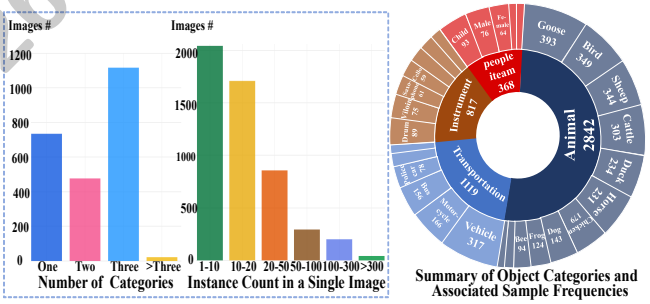

Figure 3: The statistical distribution of images across our SoundCount, categorized by class number, varying instance density intervals, and class cardinality.

Figure 4: Annotated "sound-vision" examples from SoundCount.

4.2 Image-Sound Data Collection and Curation

To circumvent the challenges inherent in scaling the acquisition of diverse multi-class scenes, we leverage two complementary data-collection strategies: generating synthetic images via off-the-shelf Multimodal Large Models (MLMs) and collecting real-world images from practical sources, culminating in a dataset of 2,351 images.

Synthetic Pipeline. Within the SoundCount, 1,328 images were synthesized using MLMs. Following a systematic comparative evaluation of leading foundational models, including *Doubao*, *DeepSeek* [Liu *et al.*, 2024a], *ChatGPT* [Welsby and Cheung, 2023], and *Qwen* [Bai *et al.*, 2023], we determined that consistently yields superior visual perceptual fidelity and structural controllability. The generation prompts were meticulously engineered to define complex scenes with multi-class coexistence and explicit random instance counts, often augmented with specific behavioural descriptions. For instance, a typical prompt specifies: "Morning scene at a suburban farm: 52 roosters crow at the rising sun; 48 cows low in the enclosure; 50 sheep bow their heads to graze and bleat". The synthetic samples are illustrated in the first row of Fig. 4.

Real-world Pipeline. Although synthetic data enables fast and efficient image generation, it often contains perceptual artifacts that can hinder real-world generalization. To bridge this gap, we augment our benchmark with authentic images collected via a retrieval-based strategy over a predefined set of object categories. Using major web search engines (*e.g.*, *Google*, *Bing*, *Baidu*) and video platforms like *TikTok*, we manually curated images and extracted frames that meet our class and density requirements. To improve retrieval efficacy and prioritize high-density scenes, search queries were augmented with quantitative modifiers such as "many", "a large number of", and "a pile of". The authentic examples are visualized in the second row of Fig. 4.

Selection and Validation Heuristics. All candidate images underwent rigorous manual inspection to ensure data integrity according to the following criteria: *i)* **Perceptual Fidelity**: Images must possess sufficient resolution and clarity to facilitate the unambiguous discrimination of individual instances; *ii)* **Instance Density**: We prioritized high-occupancy scenes, as the utility of automated counting is most significant in dense status where manual counting is cognitively demanding; *iii)* **Morphological Consistency**: Target instances in a single scene should exhibit a reasonable degree of uniformity in semantic appearance to ensure learning stability.

Audio Source Collection. We curated audio files of 1,650 for SoundCount by first identifying and downloading audio samples corresponding to countable object categories within the large-scale VGGSound [Chen *et al.*, 2020] repository. To ensure temporal consistency, each clip was segmented into a uniform 10-second window. Furthermore, all audio samples underwent manual screening to ensure category-specific clarity and signal integrity.

5 Experiments

5.1 Evaluation and Implementation Details

We evaluate the S2ICount on our self-curated SoundCount. Consistent with prior work in object counting, we use MAE and RMSE as evaluation metrics. The S2ICount is initialized with pre-trained SD v1.5 and CLAP models. During training, the parameters of the VAE encoder and the CLAP model are frozen, while the U-Net is further fine-tuned. The base learning rate is set to 5×10^{-5}. To preserve the pre-trained priors, the learning rate of the U-Net is reduced to one-tenth of this base rate. Training is performed for 200 epochs using the AdamW optimizer, with a weight decay of 1×10^{-4} and a batch size of 4. The hyper-parameters α, λ, K, and M are empirically set to 2, 0.01, 20, and 1000, respectively.

Prompt	Methods	Val Set		Test Set	
		MAE	RMSE	MAE	RMSE
Text	CLIPCount	19.38	64.32	*NA*	*NA*
	VLCounter	25.00	67.30	*NA*	*NA*
	T2ICount	7.55	21.20	*NA*	*NA*
Sound	Clipcount+SoundMamba	19.38	64.46	20.07	65.29
	VLCounter+SoundMamba	22.43	66.36	19.78	64.88
	T2ICount+SoundMamba	7.81	19.87	8.28	24.85
	S2ICount (ours)	**6.80**	**18.72**	**7.87**	**23.20**

Table 2: Comparisons of S2ICount and other state-of-the-art methods on the SoundCount dataset using text- and sound-guided schemes. *NA* means that text labels are unavailable in test set.

5.2 Comparisons with State of the Art

We carry out extensive experiments to compare the performance of our S2ICount with several state-of-the-art approaches. As shown in Tab. 2, the compared models can be broadly categorized into text-guided and sound-centric counting paradigms. Among text-guided methods, T2ICount achieves the best validation performance, with MAE of 7.55 and RMSE of 21.20. For sound-oriented approaches, we evaluate multiple variants, including existing SOTA multimodal counting models incorporating the SoundMamba module and our proposed S2ICount. Notably, S2ICount consistently outperforms all sound-based baselines and surpasses the strongest text-guided model, achieving a validation MAE of 6.80 and RMSE of 18.72. On the sound-guided test set, T2ICount with SoundMamba achieves a test MAE of 8.28 and a test RMSE of 24.85, whereas S2ICount further reduces the MAE to 7.87 and the RMSE to 23.20, yielding the best overall performance among all evaluated models. Several sound-guide counting results are visualized in Fig. 5.

5.3 Ablation Study

To validate the effectiveness of each core component in the proposed sound-oriented counting framework, we conduct ablation experiments on the SoundCount dataset. The quantitative results are reported in Tab. 3. Starting from a baseline configuration with only SoundMambaenabled, the model achieves a validation MAE of 7.19 and RMSE of 19.14, along with a test MAE of 8.45 and RMSE of 26.07. Incorporating SCM, Mamba, and L_{SFL} reduces the test RMSE to

Baseline	SoundMamba	SCM	Mamba	Circular Scanning	L_{SFL}	L_{adap}	Val Set		Test Set	
							MAE	RMSE	MAE	RMSE
✓	✓	✗	✗	✗	✗	✓	7.19	19.14	8.45	26.07
✓	✓	✓	✓	✗	✓	✓	7.21	18.57	7.98	21.93
✓	✓	✓	✗	✗	✓	✓	7.37	**18.24**	8.41	28.12
✓	✓	✓	✓	✓	✗	✓	7.52	19.89	7.88	24.31
✓	✓	✓	✓	✓	✓	✗	7.76	21.10	8.28	**21.85**
✓	✗	✓	✓	✓	✓	✗	9.09	20.91	9.27	21.25
✓	✗	✗	✗	✗	✗	✗	9.36	25.36	10.14	26.09
✓	✓	✓	✓	✓	✓	✓	**6.80**	18.72	**7.87**	23.20

Table 3: Extensive studies on the SoundCount dataset are carried out to ablate the impacts of individual components proposed in our S2ICount.

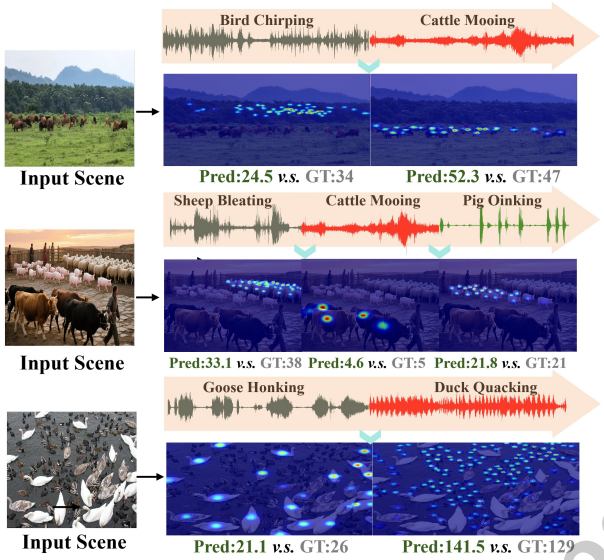

Figure 5: Qualitative examples provided by our S2ICount, demonstrating the flexible sound guidance and accurate results.

21.93, indicating a clear improvement in generalization performance. Removing Mamba yields the lowest validation RMSE of 18.24; however, this comes at the cost of degraded test performance, with the RMSE increasing to 28.12. Introducing circular scanning while excluding L_{SFL} further deteriorates validation performance, with the MAE and RMSE increasing to 7.52 and 19.89, respectively. Disabling the loss term L_{adap} leads to the lowest test RMSE of 21.85, but simultaneously worsens validation results, suggesting potential overfitting. In contrast, removing SoundMamba causes a substantial performance drop, with the validation MAE rising to 9.09 and the test MAE increasing to 9.27. With all components fully integrated, the proposed model achieves the best overall performance, attaining a validation MAE of 6.80 and a test MAE of 7.87. Conversely, removing all components produces the poorest performance across all evaluation metrics in both validation and test sets.

5.4 How do Audio Extractors Impact the Results?

To evaluate the effectiveness of varied pre-trained audio models for our method, comparative experiments are conducted on SoundCount dataset in Tab. 4. Two typical audio pre-trained models (AudioCLIP and SpeechCLIP) are selected for comparison, while our method adopts a CLAP-based audio feature extraction strategy. On the validation set, Audio Clip achieves a mean absolute error of 12.61 and a root mean square error of 29.15, and Speech Clip yields a mean absolute error of 11.64 and a root mean square error of 32.22. On the test set, the two models obtain mean absolute errors of 14.01 and 13.51 as well as root mean square errors of 38.62 and 39.35 respectively. In contrast, our CLAP-based method outperforms both counterparts by a substantial margin, delivering the optimal validation mean absolute error of 6.80 and root mean square error of 18.72, along with the best test mean absolute error of 7.87 and root mean square error of 23.20.

Methods	Val Set		Test Set	
	MAE	RMSE	MAE	RMSE
S2ICount (AudioCLIP)	12.61	29.15	14.01	38.62
S2ICount (SpeechCLIP)	11.64	32.22	13.51	39.35
S2ICount (**CLAP**)	**6.80**	**18.72**	**7.87**	**23.20**

Table 4: Comparisons of using different audio pre-trained models including AudioCLIP, SpeechCLIP, and CLAP.

6 Conclusion

In summary, this work introduces sound-guided counting via the novel S2ICount and the SoundCount dataset. S2ICount enables models to "listen and count", bridging audio-visual features through SoundMamba, a Sound Calibration Module, and Multi-tier Hybrid Optimization. SoundCount provides 2,351 scenes and 1,650 audio sources across 66 categories, addressing the scarcity of audio-informed counting data. Experiments demonstrate state-of-the-art performance, validating sound-image synergy and opening new directions for sound-centric multimodal object counting.

Acknowledgements

We thanks all reviewers and this work is supported by the National Natural Science Foundation of China (NSFC) under grant No.12526646, the Zhejiang Provincial Natural Science Foundation of China under Grant No.LQN25F020004, the Science Foundation of Zhejiang Sci-Tech University (ZSTU) under Grant No.22062338-Y.

References

[Amini-Naieni *et al.*, 2023] Niki Amini-Naieni, Kiana Amini-Naieni, Tengda Han, and Andrew Zisserman. Open-world text-specifed object counting. In *BMVC*, 2023.

[Bai *et al.*, 2023] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[Chen *et al.*, 2020] 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)*, pages 721–725. IEEE, 2020.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[Girdhar *et al.*, 2023] 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*, pages 15180–15190, 2023.

[Guzhov *et al.*, 2022] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE, 2022.

[Hsieh *et al.*, 2017] Meng-Ru Hsieh, Yen-Liang Lin, and Winston H Hsu. Drone-based object counting by spatially regularized regional proposal network. 2017.

[Huang *et al.*, 2024] Zhizhong Huang, Mingliang Dai, Yi Zhang, Junping Zhang, and Hongming Shan. Point segment and count: A generalized framework for object counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17067–17076, 2024.

[Idrees *et al.*, 2018] Haroon Idrees, Muhmmad Tayyab, Kishan Athrey, Dong Zhang, Somaya Al-Maadeed, Nasir Rajpoot, and Mubarak Shah. Composition loss for counting, density map estimation and localization in dense crowds. 2018.

[Jiang *et al.*, 2023] Ruixiang Jiang, Lingbo Liu, and Changwen Chen. Clip-count: Towards text-guided zero-shot object counting. In *ACM MM*, pages 4535–4545, 2023.

[Kang *et al.*, 2024] Seunggu Kang, WonJun Moon, Euiyeon Kim, and Jae-Pil Heo. Vlcounter: Text-aware visual representation for zero-shot object counting. In *AAAI*, volume 38, pages 2714–2722, 2024.

[Kirillov *et al.*, 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023.

[Lempitsky and Zisserman, 2010] Victor Lempitsky and Andrew Zisserman. Learning to count objects in images. In *NeurIPS*, volume 23, 2010.

[Lin *et al.*, 2025] Wei Lin, Chenyang Zhao, and Antoni B Chan. Point-to-region loss for semi-supervised point-based crowd counting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29363–29373, 2025.

[Liu *et al.*, 2024a] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[Liu *et al.*, 2024b] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024.

[Marsden *et al.*, 2018] Mark Marsden, Kevin McGuinness, Suzanne Little, Ciara E Keogh, and Noel E O'Connor. People, penguins and petri dishes: Adapting object counting models to new visual domains and object types without forgetting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8070–8079, 2018.

[Meng *et al.*, 2025] Haoliang Meng, Xiaopeng Hong, Zhengqin Lai, and Miao Shang. Free lunch enhancements for multi-modal crowd counting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14013–14023, 2025.

[Mundhenk *et al.*, 2016] T Nathan Mundhenk, Goran Konjevod, Wesam A Sakla, and Kofi Boakye. A large contextual dataset for classification, detection and counting of cars with deep learning. In *European conference on computer vision*, pages 785–800. Springer, 2016.

[Pelhan *et al.*, 2024] Jer Pelhan, Alan Lukezic, Vitjan Zavrtanik, and Matej Kristan. A novel unified architecture for low-shot counting by detection and segmentation. *Advances in Neural Information Processing Systems*, 37:66260–66282, 2024.

[Qian *et al.*, 2025] Yifei Qian, Zhongliang Guo, Bowen Deng, Chun Tong Lei, Shuai Zhao, Chun Pong Lau, Xiaopeng Hong, and Michael P Pound. T2icount: Enhancing cross-modal understanding for zero-shot counting. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25336–25345, 2025.

[Qin *et al.*, 2023] Can Qin, Ning Yu, Chen Xing, Shu Zhang, Zeyuan Chen, Stefano Ermon, Yun Fu, Caiming Xiong, and Ran Xu. Gluegen: Plug and play multi-modal encoders for x-to-image generation. In *ICCV*, pages 23085–23096, October 2023.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[Ranjan *et al.*, 2021] Viresh Ranjan, Udbhav Sharma, Thu Nguyen, and Minh Hoai. Learning to count everything. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3394–3403, 2021.

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

[Rouditchenko *et al.*, 2020] Andrew Rouditchenko, Angie Boggust, David Harwath, Brian Chen, Dhiraj Joshi, Samuel Thomas, Kartik Audhkhasi, Hilde Kuehne, Rameswar Panda, Rogerio Feris, et al. Avlnet: Learning audio-visual language representations from instructional videos. *arXiv preprint arXiv:2006.09199*, 2020.

[Shi *et al.*, 2022] Min Shi, Hao Lu, Chen Feng, Chengxin Liu, and Zhiguo Cao. Represent, compare, and learn: A similarity-aware framework for class-agnostic counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9529–9538, 2022.

[Shih *et al.*, 2023] Yi-Jen Shih, Hsuan-Fu Wang, Heng-Jui Chang, Layne Berry, Hung-yi Lee, and David Harwath. Speechclip: Integrating speech with pre-trained vision and language model. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 715–722. IEEE, 2023.

[Sindagi and Patel, 2018] Vishwanath A Sindagi and Vishal M Patel. A survey of recent advances in cnn-based single image crowd counting and density estimation. *Pattern Recognition Letters*, 107:3–16, 2018.

[Sindagi *et al.*, 2020] Vishwanath A Sindagi, Rajeev Yasarla, and Vishal M Patel. Jhu-crowd++: Large-scale crowd counting dataset and a benchmark method. *arXiv preprint arXiv:2004.03597*, 2020.

[Sung-Bin *et al.*, 2023] Kim Sung-Bin, Arda Senocak, Hyunwoo Ha, Andrew Owens, and Tae-Hyun Oh. Sound to visual scene generation by audio-to-visual latent alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6430–6440, 2023.

[Sung-Bin *et al.*, 2025] Kim Sung-Bin, Kim Jun-Seong, Junseok Ko, Yewon Kim, and Tae-Hyun Oh. Soundbrush: Sound as a brush for visual scene editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7167–7175, 2025.

[Tang *et al.*, 2023] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. *Advances in Neural Information Processing Systems*, 36:16083–16099, 2023.

[Wang *et al.*, 2020] Qi Wang, Junyu Gao, Wei Lin, and Xuelong Li. Nwpu-crowd: A large-scale benchmark for crowd counting. *arXiv preprint arXiv:2001.03360*, 2020.

[Wang *et al.*, 2022] Mingjie Wang, Hao Cai, Xian-Feng Han, Jun Zhou, and Minglun Gong. Stnet: Scale tree network with multi-level auxiliator for crowd counting. *TMM*, 25:2074–2084, 2022.

[Wang *et al.*, 2025] Mingjie Wang, Song Yuan, Xian-Feng Han, and Zili Yi. Draw what you hear: High-fidelity image generation and manipulation via soundadapter. *IEEE Transactions on Neural Networks and Learning Systems*, 2025.

[Welsby and Cheung, 2023] Philip Welsby and Bernard MY Cheung. Chatgpt, 2023.

[Wu *et al.*, 2023] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[Xu *et al.*, 2023] Jingyi Xu, Hieu Le, Vu Nguyen, Viresh Ranjan, and Dimitris Samaras. Zero-shot object counting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15548–15557, 2023.

[Yang *et al.*, 2025] Wenhao Yang, Jianguo Wei, Wenhuan Lu, and Lei Li. You only speak once to see. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2025.

[Zhang *et al.*, 2016] Yingying Zhang, Desen Zhou, Siqin Chen, Shenghua Gao, and Yi Ma. Single-image crowd counting via multi-column convolutional neural network. 2016.

[Zhang *et al.*, 2017] Shanghang Zhang, Guanhang Wu, Joao P Costeira, and José MF Moura. Fcn-rlstm: Deep spatio-temporal neural networks for vehicle counting in city cameras. In *Proceedings of the IEEE international conference on computer vision*, pages 3667–3676, 2017.

[Zhang *et al.*, 2024] Shiwei Zhang, Wei Ke, Shuai Liu, Xiaopeng Hong, and Tong Zhang. Boosting semi-supervised crowd counting with scale-based active learning. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 8681–8690, 2024.

[Zhang *et al.*, 2025] Zhenfei Zhang, Ming-Ching Chang, and Xin Li. Training-free image manipulation localization using diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 10376–10384, 2025.

[Zhu *et al.*, 2024] Huilin Zhu, Jingling Yuan, Zhengwei Yang, Yu Guo, Zheng Wang, Xian Zhong, and Shengfeng He. Zero-shot object counting with good exemplars. In *European Conference on Computer Vision*, pages 368–385. Springer, 2024.