How Far Can We Go With Synthetic Data for Audio-Visual Sound Source Localization?

Anonymous CVPR submission

Paper ID

Abstract

We present the first scalable framework for training sound source localization (SSL) models using synthetic data from text-to-X models. Although SSL has made notable progress, existing models remain constrained by limited-scale, uncurated real-world datasets that often suffer from semantic misalignment. Furthermore, the introduction of new SSL tasks and benchmarks has increased the need for more generalizable models. To address these challenges, we leverage synthetic data to create synthetic clones of the VGGSound dataset, enabling both fully synthetic and hybrid real–synthetic training. We demonstrate that synthetic data can effectively replace, refine, and scale real training datasets. Extensive experiments across multiple benchmarks show that synthetic data not only matches real data in performance but also enables significant improvements when combined with real samples. Our findings provide the first systematic evidence that synthetic data can serve as a scalable and effective approach for advancing SSL models.

1. Introduction

Perceiving, detecting, and recognizing are fundamental perceptual abilities, with humans primarily relying on visual and auditory senses for these tasks. In machine perception, these capabilities have been explored through sound source localization (SSL) models, which aim to identify sound-emitting objects or regions within a visual scene. Over time, SSL has evolved from a subtopic of audio-visual learning into an independent research field with its own objectives, methodological advancements, benchmarks and new tasks such as single sound source localization [1, 4, 34, 35], segmentation [13, 53–55], interactive localization [39], and audio-visual robustness [20, 26]. This has increased the need for more generalizable SSL models capable of handling all these diverse datasets and tasks at once.

While SSL has shown progress, existing models are model-centric in their development. However, most are trained on no more than 144K samples [17, 25, 26, 39, 45], leaving their scalability largely unexplored. Moreover, uncurated training data often suffer from misalignment between images and corresponding audio, as training images are typically extracted from the mid-frame of videos (see Figure 2) without considering semantic relevance [27, 36, 42]. These factors suggest that the first steps toward developing a more advanced SSL model may lie in addressing these challenges. Motivated by this, we suggest shifting the SSL paradigm from model-centric to data-centric learning.

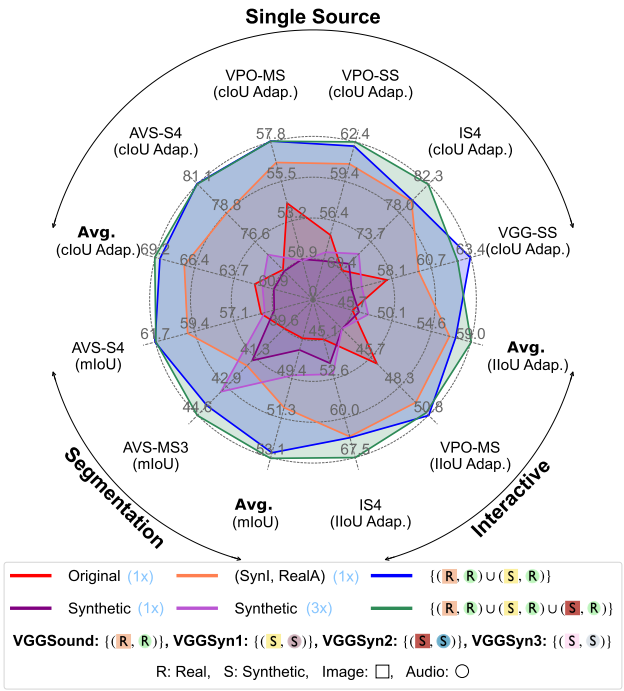

Figure 1. **Sound Source Localization Performance with Synthetic Data.** Synthetic data can replace, refine, and scale real datasets – matching real data and yielding substantial gains when combined with real data; 1.1x in localization/segmentation and 1.3x higher performance in interactive localization.

The aforementioned constraints may be solved at once through manual data collection and filtering, but this requires a substantial amount of human intervention. A more

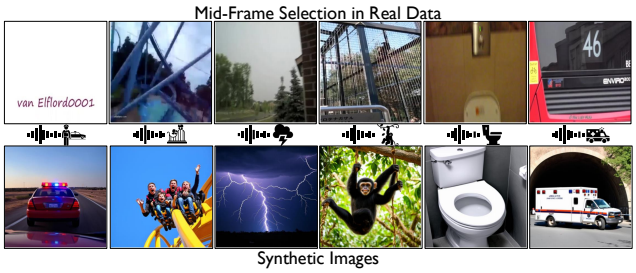

Figure 2. **Mid-frames in real data vs. synthetic images.** Mid-frame selection causes imperfections, misaligning audio and images, while synthetic images enhance semantic consistency.

practical alternative lies in the use of synthetic data, made feasible by recent advances in text-to-image/audio generative models. These models enable highly controllable data generation, allowing scalable dataset creation and the synthesis of samples for specific semantic concepts, which can help mitigate audio-visual misalignment issues. Although synthetic data has been explored in other modalities and research domains [6, 7, 18, 19, 24, 30, 32, 46, 50], its application to audio-visual SSL and more broadly to audio-visual learning has not. In this work, we leverage synthetic data as both a verification and investigation tool to explore *how far synthetic data can take us*. Specifically, we aim to (1) examine whether SSL models can effectively learn from synthetic data, (2) test the aforementioned existing bottlenecks, and (3) explore a viable path toward more advanced SSL models. We argue this path requires a specific recipe in which synthetic data plays a fundamental role.

To achieve our objective, we propose a scalable pipeline that leverages off-the-shelf generative models to create synthetic *clone(s)* of VGGSound [3], the most widely used dataset for sound source localization training. Our system consists of four steps: (1) constructing concept dictionaries for each modality, where VGGSound classes are represented by sets of concepts; (2) using a large language model (LLM) to generate caption prompts from sampled concepts; (3) synthesizing data using text-to-X models, producing synthetic versions of VGGSound; (4) training a state-of-the-art SSL model with either fully synthetic data or a hybrid approach that pairs real and synthetic samples.

We evaluate model performance across multiple datasets and tasks (Figure 1). Models trained entirely on synthetic data perform on-par with fully real data. Replacing VGGSound images with synthetic counterparts (Figure 2) yields substantial improvements, hinting that synthetic data may mitigate imperfections in real datasets. Scaling up training data by combining real and synthetic samples gets state-of-the-art performance across all tasks. Our approach outperforms the real dataset baseline by *+8.62 cIoU and +7.00 cIoU Adaptive* in Single Sound Source Localization, *+5.58 mIoU and +4.22 mIoU Adap.* in Audio-Visual Segmentation, and *+12.16 IIoU and +13.35 IIoU Adap.* in Interactive Localization. The results show the potential of synthetic data for more advanced SSL models.

Our main contributions are as follows:

- We introduce the first scalable training framework for sound source localization using synthetic data from text-to-X models and provide synthetic clones of VGGSound.
- We shift the focus from model-centric learning to data-centric learning, demonstrating the impact of synthetic data on training.
- We provide empirical validation, showing that synthetic data can replace, refine, and scale real training datasets.
- We provide a recipe for an advanced SSL achieving state-of-the-art performance with hybrid real–synthetic training, surpassing real-data-only baselines.

All data and code will be released upon acceptance.

2. Related Work

Sound Source Localization. Audio-visual sound source localization identifies objects, events, or regions in a visual scene by leveraging semantic correspondences between audio and visual cues. Early works [1, 34, 35] explored this relationship through cross-modal attention and contrastive learning for audio-visual alignment. However, real-world videos may contain misaligned audio-visual pairs, where background noise, off-screen sounds, or silent objects introduce incorrect associations, hindering localization. To address these challenges, prior methods incorporate false negative-aware learning [41], negative-free predictive learning[40], and regularization techniques [26, 28]. Others leverage visual priors, such as object proposals and motion cues, to improve performance [12, 25, 49]. Additionally, ACL-SSL [29] integrates CLIP [31] into sound source localization without using text, utilizing its strong multimodal prior knowledge to improve localization accuracy. While these methods adopt a model-centric approach, our work shifts the focus toward a data-centric perspective by exploring the role of synthetic training data.

Another research direction in self-supervised contrastive learning aims to improve audio-visual alignment through better data utilization, including sample mining [4, 37], geometric equivalence learning [22], and multi-positive contrastive learning [38]. Unlike these approaches, which refine real data, our method explores learning from synthetic data to achieve superior benefits. As the field advances, new tasks such as Audio-Visual Segmentation [54] introduce additional model-centric solutions [23, 47, 53, 54]. However, our approach differs by tackling multiple tasks simultaneously with synthetic data, aiming to develop a more generalizable model. Additionally, a semi-synthetic dataset is recently proposed in [39], where a text-to-image model generates images paired with real audio. However, unlike our

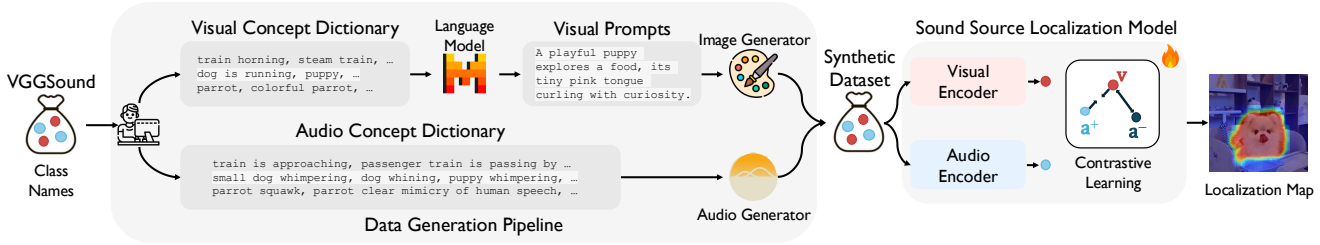

Figure 3. **The pipeline of our framework.** From VGGSound categories, we construct concept dictionaries with more descriptive terms for each class. An LLM generates visual prompts for image synthesis, while concepts are used directly for audio generation. The resulting synthetic data is then used to train a sound source localization model.

approach, their dataset is used solely for evaluation as a test set, not for training. To the best of our knowledge, this is the first systematic investigation of synthetic training data for sound source localization across multiple tasks and datasets.

Representation Learning with Synthetic Data. The use of synthetic data for training computer vision and machine learning models is well established [6, 7, 18, 19, 24, 30, 32, 46, 50]. Recent advancements in text-to-image models have further expanded its role, particularly in representation learning for both vision-only and multi-modal tasks, especially in the vision-text domain. By generating synthetic images, the recent studies [2, 11, 15, 33, 52] train image classifiers in a supervised setting, while StableRep [43] applies synthetic images to self-supervised contrastive learning. In vision-text learning, prior works have adapted CLIP-style training using fully synthetic data [14, 44] or hybrid approaches that keep one modality real [10, 11, 51]. Additionally, Tian *et al.* [44] explores multi-modal training by incorporating synthetic data and optimizing contrastive learning objectives to maximize its benefits. Beyond conventional vision-text applications, Liu *et al.* [21] explores synthetically generated radiology reports and Chest X-ray images for contrastive learning, using either fully synthetic or hybrid paired data. In contrast, the potential of synthetic data in audio and vision remains unexplored. We focus on contrastive representation learning from the perspective of audio-visual SSL.

3. Methodology

Our goal is to systematically investigate synthetic training data for sound source localization. To this end, we propose a scalable framework that leverages text-to-X generative models for synthetic data generation and training. Given a category name, concept dictionaries are constructed for each modality. Captions are then generated based on the selected concept – optionally with LLM assistance – and used as prompts to synthesize data in both modalities via text-to-X models. Finally, a sound source localization model is trained on this synthetic data, either fully or in a hybrid manner. Our pipeline is illustrated in Figure 3.

3.1. Concept Dictionaries

The first stage of our pipeline involves constructing concept dictionaries for each VGGSound category to enhance diversity in image and audio generation. These dictionaries expand beyond simple class names by providing list of more descriptive terms [14, 48]. Since concepts in one modality may not be equally descriptive in the other, we maintain separate dictionaries for audio and visual information. For example, "*thunder in the night*" provides useful context for a visual description by emphasizing "*night*", but "*night*" does not add significant information when describing audio. This stage is carried out by human annotators and is the only part of the pipeline requiring human intervention. Ten annotators, each assigned 31 categories, are asked to generate descriptive concepts for each given category, addressing audio and visual perspectives separately. On average, each class is associated with only **3.35** and **7.80** concepts for the visual and audio modalities, respectively. These enriched concepts are then used in prompts to generate synthetic data. We emphasize that this minimal manual process is the only human intervention in our pipeline, used solely to enhance the diversity of generated samples. It is far more practical than manual large-scale data collection or filtering.

3.2. Prompt Generation

In our pipeline, we use text-to-image (T2I) and text-to-audio (T2A) models to generate synthetic data, requiring text captions as prompts. Since prompt quality is crucial for text-to-X generative models, and a broader range of captions enhances data diversity, we incorporate a large language model (LLM) to generate prompts, following previous works [10, 14, 44].

For each sample in the VGGSound, a random concept is selected from the concept dictionary based on its class label and used as input to the LLM to generate caption prompts. However, different text-to-X generative models have varying sensitivities to input prompts. Our attempts show that while T2I model effectively generates images from descriptive and diverse captions, the T2A model struggles with full

sentences. Therefore, we use LLM-generated captions for image generation but rely on randomly selected concepts as direct prompts for the T2A model, as illustrated in Figure 3. Similar to [14], we design our prompt for caption generation with LLM as follows for a given concept $\{\mathbf{k}\}$: "*Your task is to write an image caption that includes and visually describes a scene around a concept. You can make this scene in unusual places. Your concept is:*$\{\boldsymbol{k}\}$*. Output one single caption that is no longer than 15 words. Don't make too artistic and emotional sentences. Focus on the concept.*"

To further enhance diversity, we randomly enable or disable the part of the prompt related to scenes in unusual locations. Formally, we can formulate entire prompt generation stage of our pipeline as follows: $\boldsymbol{\mathcal{T}}_{\text{image}} = \{t_i = G(p, k_i) \mid k_i \sim \text{Uniform}(D(c(\mathbf{x}_i))), \forall \mathbf{x}_i \in \boldsymbol{\mathcal{D}}\}$ and $\boldsymbol{\mathcal{T}}_{\text{audio}} = \{k_i \sim \text{Uniform}(D(c(\mathbf{x}_i))) \mid \forall \mathbf{x}_i \in \boldsymbol{\mathcal{D}}\}$, where $\boldsymbol{\mathcal{D}} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ is the dataset with N samples, $c(\mathbf{x}_i)$ gives the class label of the sample $\mathbf{x}_i$, and $D(c)$ returns the set of concepts $\{k_1, k_2, \ldots, k_m\}$ associated with the class label. $G(p, k_i)$ represents the LLM that takes the prompt p and the randomly selected concept k_i, returning a prompt caption for generation.

3.3. Synthetic Image and Audio Generation

Using the caption prompts for each sample in VGGSound, we generate synthetic images and audio at this stage. We use diffusion-based text-to-image and text-to-audio generators, denoted as G_{T2I} and G_{T2A}, respectively. This process is formulated as follows: $\boldsymbol{\mathcal{V}} = \{\mathbf{v}_i = G_{\text{T2I}}(t_i) \mid \forall t_i \in \boldsymbol{\mathcal{T}}_{\text{image}}\}$ and $\boldsymbol{\mathcal{A}} = \{\mathbf{a}_i = G_{\text{T2A}}(t_i) \mid \forall t_i \in \boldsymbol{\mathcal{T}}_{\text{audio}}\}$, where $\boldsymbol{\mathcal{V}}$ and $\boldsymbol{\mathcal{A}}$ represent the sets of generated images and audio samples, respectively. Some examples of the generated images are shown in Figure 2, while the remaining image and audio visualizations are provided in the supplementary material. As can be seen, the generated samples accurately capture the intended semantic information. No filtering is applied to the generated samples. While we use VGGSound categories to create synthetic clones, the framework is adaptable to any dataset or category range.

3.4. Training

Since ACL-SSL [29] is the state-of-the-art method for self-supervised sound source localization, we use this model in our study to better understand the use of synthetic data, as well as explore the potential for further performance improvements, aiming to achieve a new state-of-the-art in sound source localization. ACL-SSL employs the CLIP [31] visual encoder to extract image features, while the audio signal is processed using BEATs [5] as the audio encoder. The audio is first projected into text-like tokens compatible with CLIP's text encoder, then transformed into an audio-driven embedding via the CLIP text encoder. These embeddings are used to highlight sounding regions within visual scenes and align audio and visual features through self-supervised contrastive learning. Notably, this approach operates without explicit text input, relying solely on audio-visual correspondence.

4. Experiments

4.1. Experiment Setup

Downstream Tasks. We conduct all analyses in our study from the perspectives of three different downstream tasks, following [39]: (1) *Single Sound Source Localization*, (2) *Audio-Visual Segmentation*, and (3) *Interactive Localization*. The first task focuses on localizing sound sources from a detection perspective, specifically targeting single sounding objects. This involves localizing one object at a time, rather than detecting multiple objects simultaneously. The second task aims to localize sound sources from a segmentation perspective. The final task evaluates the model's ability to shift the localized region in the image when the same image is paired with a different sound present in the scene.

Datasets. We train the model using the VGGSound [3] dataset, our generated synthetic datasets (which varies in scale), or a combination of both. For evaluation, we employ the following datasets: VGG-SS [4], IS4 [39], VPO-SS [53], VPO-MS [53], and AVSBench-S4 [54] for Single Sound Source Localization; AVSBench-S4 and AVSBench-MS3 [54] for Audio-Visual Segmentation; and IS4 and VPO-MS for Interactive Localization.

Evaluation Metrics. We employ the following evaluation metrics, as in [39]: cIoU, cIoU Adaptive, AUC, and AUC Adaptive for Single Sound Source Localization; mIoU, mIoU Adaptive, F-Score, and F-Score Adaptive for Audio-Visual Segmentation; and IIoU, IIoU Adaptive, IAUC, and IAUC Adaptive for Interactive Localization. All baseline model checkpoints for evaluation are selected based on their peak performance on the VGG-SS dataset. As noted in [29], model performance can vary across different random seeds. Therefore, we train each baseline model six times and report the mean and standard deviation of the results for each dataset to fairly assess the impact of synthetic data. Since each task comprises multiple datasets that may behave differently, we average the performance [14, 15, 44] across all datasets to report the final performance for each task.

Caption & Image & Audio Generation Models. We use Mistral-7B-Instruct-v0.2 [16], Stable Diffusion 3 Medium [8], and Stable Audio Open 1.0 [9] with the default settings released on HuggingFace for caption generation, image generation, and audio generation, respectively. Although stronger and more specialized models could be used (see supp. material), we aim to utilize the most standard and widely available generative models to establish a more cost-effective recipe.

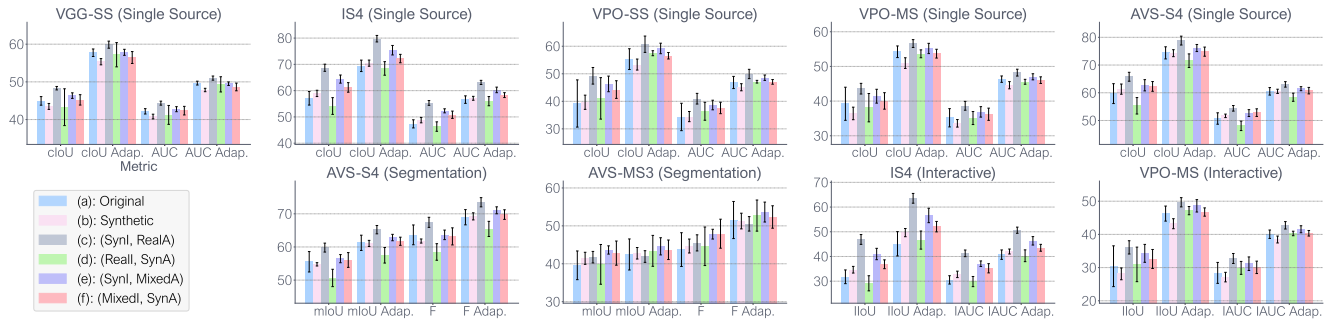

Figure 4. **Sound source localization results on the same data scale.**

Training Details. We strictly follow the same training of ACL-SSL [29]. We use pre-trained ViT-B/16 CLIP [31] model as the image encoder and BEATs [5] as the audio encoder. During training, we use 10-second audio clips sampled at 16 kHz and images resized to 352×352 pixels. The model is optimized for 20 epochs with a batch size of 16.

4.2. Benchmark Analysis on the Same Data Scale

In this section, we verify whether synthetic data can enable sound source localization training and, if so, to what extent it improves performance. Furthermore, we investigate ways to better utilize it, *e.g.*, by replacing VGGSound images with synthetic ones. Additionally, the experiments here offer insights into improving sound source localization models at the same data scale. In this analysis, the data scale matches VGGSound's 144k samples. The results for each task and dataset are shown in Figure 4, while Table 1 summarizes the average performance for each task. Detailed numerical values for Figure 4 are in the supp. material.

Baselines. We design the following variants: (a) *Original* uses entirely real images and audios from VGGSound; (b) *Synthetic* uses fully synthetic audios and images generated by our pipeline; (c) *(SynI, RealA)* uses paired data consisting of synthetic images and real audios from VGGSound; (d) *(RealI, SynA)* similarly uses pair of real images from VGGSound and synthetic audios; (e) *(SynI, MixedA)* uses synthetic images paired with a mixed set of audio samples (#Syn.: 134K + #Real: 10K); and (f) *(MixedI, SynA)* uses a mixed set of images (#Syn.: 134K + #Real: 10K) paired with synthetic audios.

Key findings are as follows:

1. Synthetic data is ready for sound source localization. We study the scenario where no real data is available and instead use synthetic data by comparing the Original and Synthetic baselines. As shown in Table 1, the Synthetic variant achieves comparable or better performance depending on the task. This verifies that synthetic data can effectively substitute for real data in sound source localization.

2. Pairing synthetic images and real audios proves highly beneficial. Our experimental results show that replacing images of VGGSound with synthetic counterparts yields substantial improvement. Specifically, *(SynI, RealA)* achieves performance gains of (**+7.10** cIoU and **+4.94** cIoU Adap.) in Single Sound Source Localization, (**+4.61** mIoU and **+3.50** mIoU Adap.) in Audio-Visual Segmentation, and (**+10.40** IIoU and **+10.91** IIoU Adap.) in the Interactive Localization task. This verifies our hypothesis that refining problematic frames with semantically aligned synthetic images—aligned with the sounding object as shown in Figure 2—can mitigate the misalignment bottleneck.

	Method	cIoU	cIoU Adap.	AUC	AUC Adap.
Single Source	(a) Original	48.03	62.22	41.95	51.99
	(b) Synthetic	47.97	60.88	41.86	51.03
	(c) (SynI,RealA)	55.13	67.16	46.73	55.06
	(d) (RealI,SynA)	46.38	61.66	41.43	51.28
	(e) (SynI,MixedA)	52.24	64.75	44.66	53.37
	(f) (MixedI,SynA)	50.61	62.75	43.96	52.16
	Method	**mIoU**	**mIoU Adap.**	**F-Score**	**F-Score Adap.**
Segmentation	(a) Original	47.56	51.85	53.70	60.26
	(b) Synthetic	48.10	51.92	53.24	60.25
	(c) (SynI,RealA)	50.78	53.57	56.47	62.01
	(d) (RealI,SynA)	45.25	50.45	51.55	59.08
	(e) (SynI,MixedA)	50.06	53.72	55.73	62.35
	(f) (MixedI,SynA)	49.44	52.71	55.56	61.06
	Method	**IIoU**	**IIoU Adap.**	**IAUC**	**IAUC Adap.**
Interactive	(a) Original	31.10	45.71	29.47	40.40
	(b) Synthetic	31.38	46.44	29.95	40.24
	(c) (SynI,RealA)	41.49	56.61	37.01	46.69
	(d) (RealI,SynA)	30.06	46.90	29.88	40.29
	(e) (SynI,MixedA)	37.65	52.61	34.12	43.95
	(f) (MixedI,SynA)	34.72	49.41	32.65	41.93

Table 1. **Sound source localization results on the same data scale.** The best, and the second best results are highlighted. *Mixed* denotes both synthetic and real samples within the same modality are used in a mixed form as (134K Syn. + 10K Real).

3. Pairing with real data makes a difference, even when the amount is very small. Our experimental results show that pairing synthetic data with real data, either fully or partially, improves performance across all tasks, except for the *(RealI, SynA)* variant, which will be discussed in the following paragraphs. We explore mixed-pair training with variants that use a small amount of real data in one modality. Our results show that even with just 10K real images or real audio samples, *(SynI, MixedA)* and *(MixedI, SynA)* achieve performance gains across all tasks compared to the Original. Specifically, *(SynI, MixedA)* outperforms by (**+4.20** cIoU

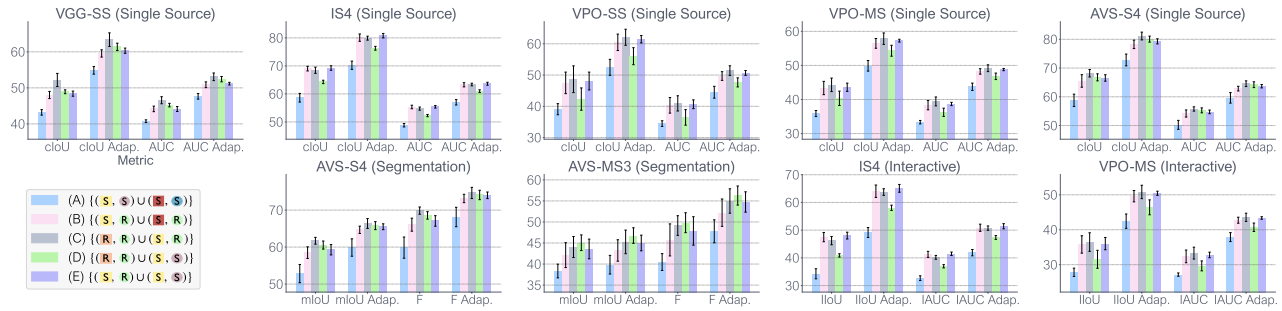

Figure 5. **Sound source localization results on 2× data scale.**

and **+2.53** cIoU Adap.) in Single Sound Source Localization, (**+2.50** mIoU and **+1.87** mIoU Adap.) in Audio-Visual Segmentation, and (**+6.55** IIoU and **+6.90** IIoU Adap.) in Interactive Localization. We hypothesize that mixed-pair training is effective because synthetic data can mitigate the semantic imperfections present in real samples, while real samples can help address the domain gap of synthetic data.

4. The visual modality plays a more significant role in synthetic data. Given the two types of synthetic data, one might wonder which contributes more to the model's performance. Mixed-pair training experiments – comparing Original vs. *(SynI, RealA)* and other variants – reveal that synthetic images offer greater benefits. As discussed earlier, real visual data often contain imperfections due to mid-frame selection, whereas synthetic images provide semantically cleaner and more consistent visual cues. For the audio modality, except for the special case *(RealI, SynA)* discussed in the next paragraph, synthetic audio remains useful, although not as effective as synthetic images: in variant (b), performance is comparable to the Original, and even higher in (e) and (f). The gap between (c) and (e) suggests that synthetic audio is less effective than real audios. This can be attributed to the relative maturity of the two fields; T2I generation is currently more advanced and yields more accurate, semantically aligned results than T2A generation.

5. Combining synthetic audio with real images is not an effective pairing. The *(RealI, SynA)* variant yields the lowest performance among all experiments. As discussed earlier, synthetic audio is less effective, and this issue becomes more pronounced when paired with real data. The likely cause is that real images are already semantically noisy, and combining them with less effective synthetic audio further amplifies these imperfections. The comparison between *(RealI, SynA)* and *(MixedI, SynA)* supports this explanation, as replacing real images with synthetic ones while keeping the same synthetic audio leads to consistent performance improvements across all tasks.

4.3. Scaling Beyond 144K Data

Having analyzed synthetic data at the same scale and identified the key elements of our recipe, we now investigate training models beyond 144K samples, an approach that has not been explored in any previous SSL work. Experiments are first conducted at the 2× scale to identify the most effective variants, which are then used for subsequent experiments at the 3× scale. The results for each task and dataset are presented in Figures 5 and 6, while Tables 2 and 3 summarize the average performance for each task in 2× and 3× scaled data, respectively. Detailed numbers for each dataset are in the supp. material.

	Method	cIoU	cIoU Adap.	AUC	AUC Adap.
Single Source	(A) {(S, S) ∪ (S, S)}	47.11	60.05	41.53	50.51
	(B) {(S, R) ∪ (S, R)}	54.68	66.96	46.47	54.98
	(C) {(R, R) ∪ (S, R)}	56.36	68.87	47.49	56.32
	(D) {(R, R) ∪ (S, S)}	52.53	65.61	45.09	54.41
	(E) {(S, R) ∪ (S, S)}	55.17	67.82	46.72	55.59
	Method	**mIoU**	**mIoU Adap.**	**F-Score**	**F-Score Adap.**
Segmentation	(A) {(S, S) ∪ (S, S)}	45.60	49.87	50.17	57.96
	(B) {(S, R) ∪ (S, R)}	50.31	53.95	55.90	62.68
	(C) {(R, R) ∪ (S, R)}	52.88	55.83	59.51	64.81
	(D) {(R, R) ∪ (S, S)}	52.81	56.29	59.20	65.19
	(E) {(S, R) ∪ (S, S)}	51.44	55.26	57.47	64.38
	Method	**IIoU**	**IIoU Adap.**	**IAUC**	**IAUC Adap.**
Interactive	(A) {(S, S) ∪ (S, S)}	31.04	45.81	29.91	39.89
	(B) {(S, R) ∪ (S, R)}	41.64	56.83	36.82	46.80
	(C) {(R, R) ∪ (S, R)}	41.35	57.29	36.75	47.15
	(D) {(R, R) ∪ (S, S)}	36.22	52.25	33.31	44.04
	(E) {(S, R) ∪ (S, S)}	42.03	57.77	37.13	47.40

Table 2. **Sound source localization results on 2× scaled data.** R: 144K Real Images, S: 144K Synthetic Images, S: 144K Synthetic Images, R: 144K Real Audios, S: 144K Synthetic Audios, S: 144K Synthetic Audios. The best, and the second best results are highlighted.

Scaling Up Data. We scale the data to twice and three times its original size using two methods. First, we combine the entire existing real dataset **VGGSound** (R: 144K Real Images, R: 144K Real Audios) with our previously generated synthetic dataset, **VGGSyn1** (S: 144K Synthetic Images, S: 144K Synthetic Audios), which is also 144K in size for 2× scaled data experiments. Second, we generate additional synthetic datasets with the same data distribution as in the previous section, namely **VGGSyn2** (S: 144K Synthetic Images, S: 144K Synthetic Audios) and **VGGSyn3** (S: 144K Synthetic Images, S: 144K Synthetic Audios).

Baselines for 2×. We design the following variants, each of which forms a unique training sample by pairing an audio with an image: (A) {(S, S) ∪ (S, S)} uti-

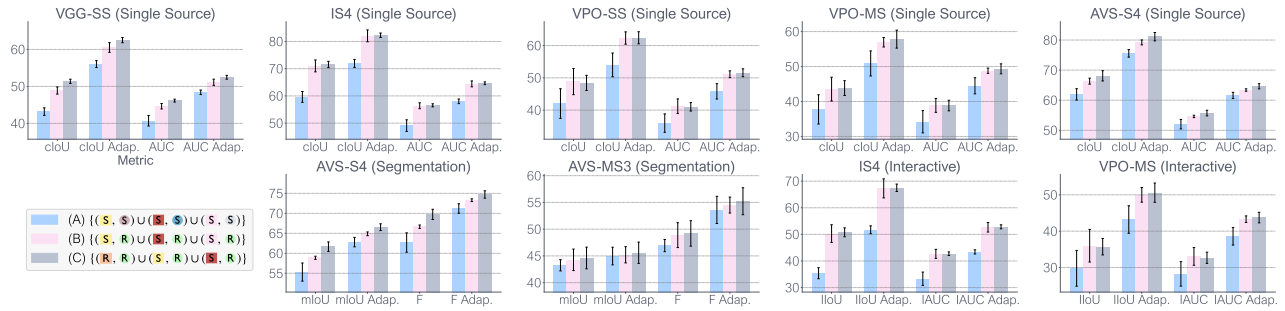

Figure 6. **Sound source localization results on 3× data scale.**

lizes fully synthetic data in both modalities, combination of **VGGSyn1** and **VGGSyn2**; (B) {(S, R) ∪ (S, R)} contains pairs of synthetic images with real audio; (C) {(R, R) ∪ (S, R)} pairs real audio with both real and synthetic images. This can be seen as an extension of the *(SynI, RealA)* variant from the previous section by adding the real dataset on top of it; (D) {(R, R) ∪ (S, S)} combines **VGGSound** and **VGGSyn1** datasets; and (E) {(S, R) ∪ (S, S)} pairs synthetic images with both real and synthetic audios.

Baselines for 3×. We design the following variants based on the experimental results from 2× scaling, adopting the best strategies while narrowing the scope of the experiments. (A) {(S, S) ∪ (S, S) ∪ (S, S)} uses fully synthetic data, combining **VGGSyn1**, **VGGSyn2** and **VGGSyn3**; (B) {(S, R) ∪ (S, R) ∪ (S, R)} pairs synthetic images with real audio; and (C) {(R, R) ∪ (S, R) ∪ (S, R)} pairs real audio with both real and synthetic images at the 3× scale.

Key findings are as follows:

1. Increasing the scale of fully synthetic data leads to performance improvements. Our objective in this experiment is to analyze whether scaling fully synthetic data can match or surpass the performance of models trained on real data and those trained with 1× scale synthetic data. We train the model using 2× and 3× scaled data, respectively. While the 2× scaled variant did not yield performance improvements, the 3× scaled variant {(S, S) ∪ (S, S) ∪ (S, S)} outperforms both the Original and 1× Synthetic variants. Specifically, the 3× fully synthetic variant achieves a performance difference compared to the Original of (**+0.87** cIoU and **-0.54** cIoU Adap.) in Single Sound Source Localization, (**+1.72** mIoU and **+2.04** mIoU Adap.) in Audio-Visual Segmentation, and (**+1.50** IIoU and **+1.77** IIoU Adap.) in Interactive Localization (See (A) in Table 3 and (a) and (b) in Table 1).

2. Increasing the number of synthetic images while keeping real audio further enhances performance. We previously observe a significant improvement in model performance when synthetic images are paired with real audio. Building on this, we apply the same approach at 2× and 3× scales.

Single Source Localization

Method	cIoU	cIoU Adap.	AUC	AUC Adap.
(A) {(S, S) ∪ (S, S) ∪ (S, S)}	48.91	61.68	42.41	51.66
(B) {(S, R) ∪ (S, R) ∪ (S, R)}	55.73	68.21	47.15	55.77
(C) {(R, R) ∪ (S, R) ∪ (S, R)}	56.66	69.23	47.66	56.54

Audio-Visual Segmentation

Method	mIoU	mIoU Adap.	F-Score	F-Score Adap.
(A) {(S, S) ∪ (S, S) ∪ (S, S)}	49.28	53.89	54.82	62.41
(B) {(S, R) ∪ (S, R) ∪ (S, R)}	51.58	55.05	57.77	63.90
(C) {(R, R) ∪ (S, R) ∪ (S, R)}	53.14	56.07	59.47	64.97

Interactive Localization

Method	IIoU	IIoU Adap.	IAUC	IAUC Adap.
(A) {(S, S) ∪ (S, S) ∪ (S, S)}	32.59	47.47	30.79	41.00
(B) {(S, R) ∪ (S, R) ∪ (S, R)}	43.15	58.64	37.87	47.98
(C) {(R, R) ∪ (S, R) ∪ (S, R)}	43.26	59.05	37.71	48.29

Table 3. **Sound source localization results on 3× scaled data.** **VGGSound** (R: 144K Real Images, R: 144K Real Audios), **VGGSyn1** (S: 144K Synthetic Images, S: 144K Synthetic Audios), **VGGSyn2** (S: 144K Synthetic Images, S: 144K Synthetic Audios) and **VGGSyn3** (S: 144K Synthetic Images, S: 144K Synthetic Audios). The best, and the second best results are highlighted.

As shown in Table 3, variant (B) {(S, R) ∪ (S, R) ∪ (S, R)} outperforms its 1× and 2× counterparts ((c) in Table 1 and (B) in Table 2, respectively). Notably, we reaffirm that this variant, at any scale, consistently outperforms the model trained on real data with a large margin.

3. Mixing real and synthetic images helps with domain generalization. We observe that mixing synthetic and real images is a promising approach. Based on the comparisons between (B) and (C), as well as (C) and (E) in Table 2, and (b) and (f) in Table 1, we hypothesize that combining real and synthetic images may enhance domain generalization.

4. Combining real dataset pairs with synthetic image–real audio pairs significantly improves performance. Given our findings that synthetic image–real audio pairs contribute to high performance and that mixing synthetic and real images is a promising approach, we aim to leverage both. In other words, we expand the real VGGSound dataset by integrating synthetic image–real audio pairs. Specifically, we adopt a mixed training strategy that utilizes both synthetic and real image–real audio pairs simultaneously. Our experiments show that variant (C) {(R, R) ∪ (S, R) ∪ (S, R)} in Table 3 outperforms all other variants in this study across all tasks, achieving a new state-of-the-art performance in sound source localization with a significant

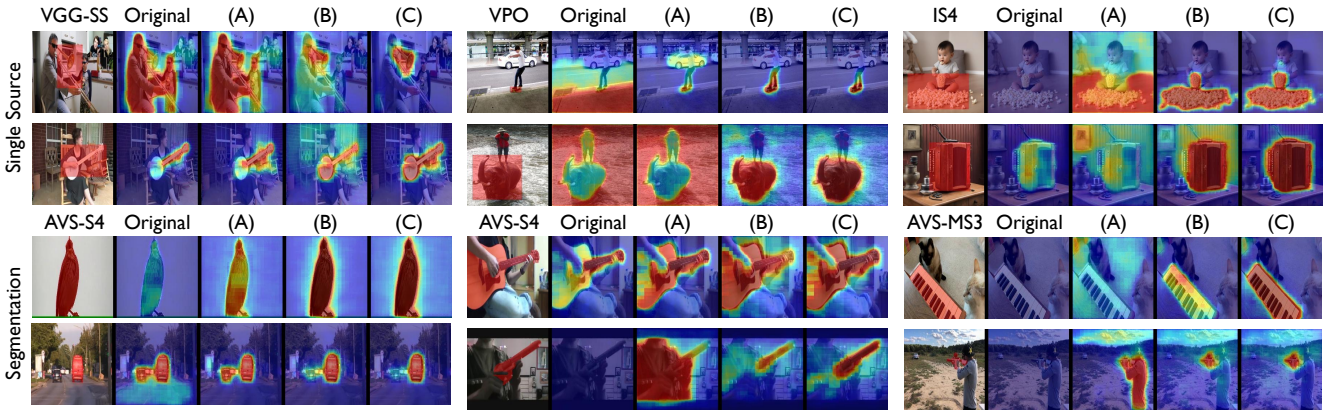

Figure 7. **Qualitative results across various datasets, comparing 3× scale variants with the Original.**

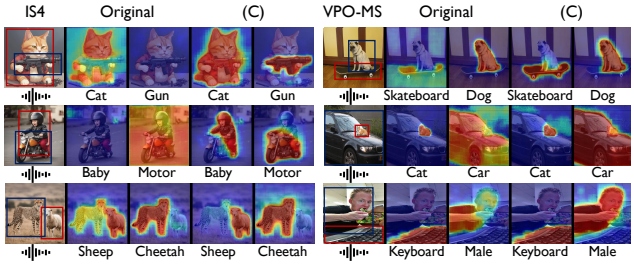

Figure 8. **Interactive Localization results comparing the 3× scale variant (C) with the Original.**

margin. In particular, it achieves a substantial performance difference compared to the Original of (**+8.62** cIoU and **+7.00** cIoU Adap.) in Single Sound Source Localization, (**+5.58** mIoU and **+4.22** mIoU Adap.) in Audio-Visual Segmentation, and (**+12.16** IIoU and **+13.35** IIoU Adap.) in Interactive Localization. Note that the same variant at the 2× scale ((C) in Table 2) also achieves high performance, falling only marginally behind the 3× scale results. We hypothesize that, in this variant, the model benefits from well-aligned real image–audio pairs, which help address the domain gap in real test sets, while also correcting noisy correspondence samples using synthetic image–real audio pairs. This finding, namely extending the real VGGSound dataset with large-scale synthetic image–real audio pairs, represents our final recipe and the most viable path toward developing a more accurate and generalizable SSL model within our experimental scale.

4.4. Qualitative Results

Figure 7 compares the Original model with the top-performing 3× scale variants (A, B, and C) from Table 3 in Single Source Localization and Audio-Visual Segmentation tasks. As shown in the numerical results, (B) and (C) localize sound sources more accurately than both the Original and variant (A), even detecting small objects like the '*skateboard*' in VPO (first row) and the '*gun*' in AVS-MS3 (last row). Additionally, where the Original model fails to generate activations, (B) and (C) provide accurate predictions, such as the '*popcorn*' in IS4. Compared to the Original, (A) (fully synthetic at 3× scale) performs similarly overall but shows better performance in the segmentation task, particularly in cases like the 'guitar' and 'bird' in AVS-S4 (first row), aligning with numerical results in Table 1 and Table 3.

Figure 8 visualizes the Interactive Localization task results, comparing the Original model with variant (C) from Table 3. This 3× scale variant, which combines real data with synthetic image–real audio pairs, clearly outperforms the real-data-only model, demonstrating the benefits of scalability and hybrid training.

5. Conclusion and Discussion

In this work, we introduce the first scalable pipeline for training sound source localization models (SSL) using synthetic data from text-to-X models. Our approach shifts the focus from model-centric to data-centric learning, showing that synthetic data can replace, refine, and scale real datasets. By generating synthetic clones of VGGSound, we enable training on fully synthetic and hybrid real-synthetic datasets. Extensive experiments validate that synthetic data can match or surpass real data in SSL. Replacing real images with synthetic counterparts mitigates semantic misalignment, while scaling up training data with real and synthetic datasets achieves state-of-the-art performance, significantly outperforming real-data-only baseline. To the best of our knowledge, this is the first systematic study on synthetic training data for SSL across multiple tasks and datasets. Our findings highlight the potential of synthetic data in advancing scalable and generalizable SSL models.

What can be further done? Future work may explore more diverse synthetic data generation strategies, such as automating concept dictionaries with LLMs. Another direction is improving generalization beyond VGGSound's class labels by incorporating open-world categories. Lastly, scaling to larger datasets would be beneficial, as limited computational resources restricted us from exceeding the 3× dataset size. These remain as future work.

References

[1] Relja Arandjelović and Andrew Zisserman. Objects that sound. In *ECCV*, 2018. 1, 2

[2] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from diffusion models improves imagenet classification. *TMLR*, 2023. 3

[3] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, 2020. 2, 4

[4] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *CVPR*, 2021. 1, 2, 4

[5] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, and Furu Wei. BEATs: Audio pre-training with acoustic tokenizers. In *ICML*, 2023. 4, 5

[6] Yuhua Chen, Wen Li, Xiaoran Chen, and Luc Van Gool. Learning semantic segmentation from synthetic data: A geometrically guided input-output adaptation approach. In *CVPR*, 2019. 2, 3

[7] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *ICCV*, 2015. 2, 3

[8] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 4

[9] Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open. *arXiv preprint arXiv:2407.14358*, 2024. 4

[10] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *NeurIPS*, 2023. 3

[11] Lijie Fan, Kaifeng Chen, Dilip Krishnan, Dina Katabi, Phillip Isola, and Yonglong Tian. Scaling laws of synthetic images for model training... for now. In *CVPR*, 2024. 3

[12] Dennis Fedorishin, Deen Dayal Mohan, Bhavin Jawade, Srirangaraj Setlur, and Venu Govindaraju. Hear the flow: Optical flow-based self-supervised visual sound source localization. In *WACV*, 2023. 2

[13] Mark Hamilton, Andrew Zisserman, John R Hershey, and William T Freeman. Separating the" chirp" from the" chat": Self-supervised visual grounding of sound and language. In *CVPR*, pages 13117–13127, 2024. 1

[14] Hasan Abed Al Kader Hammoud, Hani Itani, Fabio Pizzati, Philip Torr, Adel Bibi, and Bernard Ghanem. Synthclip: Are we ready for a fully synthetic clip training? *arXiv preprint arXiv:2402.01832*, 2024. 3, 4

[15] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? In *ICLR*, 2023. 3, 4

[16] Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023. 4

[17] Inho Kim, Youngkil Song, Jicheol Park, Won Hwa Kim, and Suha Kwak. Improving sound source localization with joint slot attention on image and audio. In *CVPR*, 2025. 1

[18] Yo-whan Kim, Samarth Mishra, SouYoung Jin, Rameswar Panda, Hilde Kuehne, Leonid Karlinsky, Venkatesh Saligrama, Kate Saenko, Aude Oliva, and Rogerio Feris. How transferable are video representations based on synthetic data? *NeurIPS*, 2022. 2, 3

[19] Byung-Ki Kwon, Sung-Bin Kim, and Tae-Hyun Oh. The devil in the details: Simple and effective optical flow synthetic data generation. *The Visual Computer*, 40(12):8665–8674, 2024. 2, 3

[20] Jia Li, Wenjie Zhao, Ziru Huang, Yunhui Guo, and Yapeng Tian. Do audio-visual segmentation models truly segment sounding objects? *arXiv preprint arXiv:2502.00358*, 2025. 1

[21] Che Liu, Zhongwei Wan, Haozhe Wang, Yinda Chen, Talha Qaiser, Chen Jin, Fariba Yousefi, Nikolay Burlutskiy, and Rossella Arcucci. Can medical vision-language pre-training succeed with purely synthetic data? *arXiv preprint arXiv:2410.13523*, 2024. 3

[22] Jinxiang Liu, Chen Ju, Weidi Xie, and Ya Zhang. Exploiting transformation invariance and equivariance for self-supervised sound localisation. In *ACM MM*, 2022. 2

[23] Juncheng Ma, Peiwen Sun, Yaoting Wang, and Di Hu. Stepping stones: a progressive training strategy for audio-visual semantic segmentation. In *ECCV*, 2024. 2

[24] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *CVPR*, 2016. 2, 3

[25] Shentong Mo and Pedro Morgado. Localizing visual sounds the easy way. In *ECCV*, 2022. 1, 2

[26] Shentong Mo and Pedro Morgado. A closer look at weakly-supervised audio-visual source localization. *NeurIPS*, 2022. 1, 2

[27] Pedro Morgado, Ishan Misra, and Nuno Vasconcelos. Robust audio-visual instance discrimination. In *CVPR*, 2021. 1

[28] Sooyoung Park, Arda Senocak, and Joon Son Chung. MarginNCE: Robust sound localization with a negative margin. In *ICASSP*, 2023. 2

[29] Sooyoung Park, Arda Senocak, and Joon Son Chung. Can clip help sound source localization? In *WACV*, 2024. 2, 4, 5

[30] Albert Pumarola, Jordi Sanchez-Riera, Gary Choi, Alberto Sanfeliu, and Francesc Moreno-Noguer. 3DPeople: Modeling the geometry of dressed humans. In *ICCV*, 2019. 2, 3

[31] 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 *ICML*, 2021. 2, 4, 5

[32] Swami Sankaranarayanan, Yogesh Balaji, Arpit Jain, Ser Nam Lim, and Rama Chellappa. Learning from synthetic data: Addressing domain shift for semantic segmentation. In *CVPR*, 2018. 2, 3

[33] Mert Bülent Sarıyıldız, Karteek Alahari, Diane Larlus, and Yannis Kalantidis. Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In *CVPR*, 2023. 3

[34] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *CVPR*, 2018. 1, 2

[35] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound sources in visual scenes: Analysis and applications. *IEEE TPAMI*, 2021. 1, 2

[36] Arda Senocak, Hyeonggon Ryu, Junsik Kim, and In So Kweon. Less can be more: Sound source localization with a classification model. In *WACV*, 2022. 1

[37] Arda Senocak, Hyeonggon Ryu, Junsik Kim, and In So Kweon. Learning sound localization better from semantically similar samples. In *ICASSP*, 2022. 2

[38] Arda Senocak, Hyeonggon Ryu, Junsik Kim, Tae-Hyun Oh, Hanspeter Pfister, and Joon Son Chung. Sound source localization is all about cross-modal alignment. In *ICCV*, 2023. 2

[39] Arda Senocak, Hyeonggon Ryu, Junsik Kim, Tae-Hyun Oh, Hanspeter Pfister, and Joon Son Chung. Toward interactive sound source localization: Better align sight and sound! *IEEE TPAMI*, 2025. 1, 2, 4

[40] Zengjie Song, Yuxi Wang, Junsong Fan, Tieniu Tan, and Zhaoxiang Zhang. Self-supervised predictive learning: A negative-free method for sound source localization in visual scenes. In *CVPR*, 2022. 2

[41] Weixuan Sun, Jiayi Zhang, Jianyuan Wang, Zheyuan Liu, Yiran Zhong, Tianpeng Feng, Yandong Guo, Yanhao Zhang, and Nick Barnes. Learning audio-visual source localization via false negative aware contrastive learning. In *CVPR*, 2023. 2

[42] 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 *CVPR*, 2023. 1

[43] Yonglong Tian, Lijie Fan, Phillip Isola, Huiwen Chang, and Dilip Krishnan. Stablerep: Synthetic images from text-to-image models make strong visual representation learners. *NeurIPS*, 2023. 3

[44] Yonglong Tian, Lijie Fan, Kaifeng Chen, Dina Katabi, Dilip Krishnan, and Phillip Isola. Learning vision from models rivals learning vision from data. In *CVPR*, 2024. 3, 4

[45] Sung Jin Um, Dongjin Kim, Sangmin Lee, and Jung Uk Kim. Object-aware sound source localization via audio-visual scene understanding. In *CVPR*, 2025. 1

[46] Gul Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017. 2, 3

[47] Yaoting Wang, Peiwen Sun, Dongzhan Zhou, Guangyao Li, Honggang Zhang, and Di Hu. Ref-avs: Refer and segment objects in audio-visual scenes. In *ECCV*, 2024. 2

[48] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. *arXiv preprint arXiv:2309.16671*, 2023. 3

[49] Hanyu Xuan, Zhiliang Wu, Jian Yang, Yan Yan, and Xavier Alameda-Pineda. A proposal-based paradigm for self-supervised sound source localization in videos. In *CVPR*, 2022. 2

[50] Moon Ye-Bin, Nam Hyeon-Woo, Wonseok Choi, Nayeong Kim, Suha Kwak, and Tae-Hyun Oh. Synaug: Exploiting synthetic data for data imbalance problems. *Pattern Recognition Letters*, 2025. 2, 3

[51] Qiying Yu, Quan Sun, Xiaosong Zhang, Yufeng Cui, Fan Zhang, Yue Cao, Xinlong Wang, and Jingjing Liu. Capsfusion: Rethinking image-text data at scale. In *CVPR*, 2024. 3

[52] Zhuoran Yu, Chenchen Zhu, Sean Culatana, Raghuraman Krishnamoorthi, Fanyi Xiao, and Yong Jae Lee. Diversify, don't fine-tune: Scaling up visual recognition training with synthetic images. *arXiv preprint arXiv:2312.02253*, 2023. 3

[53] Chen Yuanhong, Liu Yuyuan, Wang Hu, Liu Fengbei, Wang Chong, and Carneiro Gustavo. Unraveling instance associations: A closer look for audio-visual segmentation. In *CVPR*, 2024. 1, 2, 4

[54] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio-visual segmentation. In *ECCV*, 2022. 2, 4

[55] Jinxing Zhou, Xuyang Shen, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, et al. Audio-visual segmentation with semantics. *IJCV*, 133(4):1644–1664, 2025. 1