Separating the "Chirp" from the "Chat": Self-supervised Visual Grounding of Sound and Language

Mark Hamilton
MIT, Microsoft
markth@mit.edu

Andrew Zisserman
Oxford, Google

John R. Hershey
Google

William T. Freeman
MIT, Google

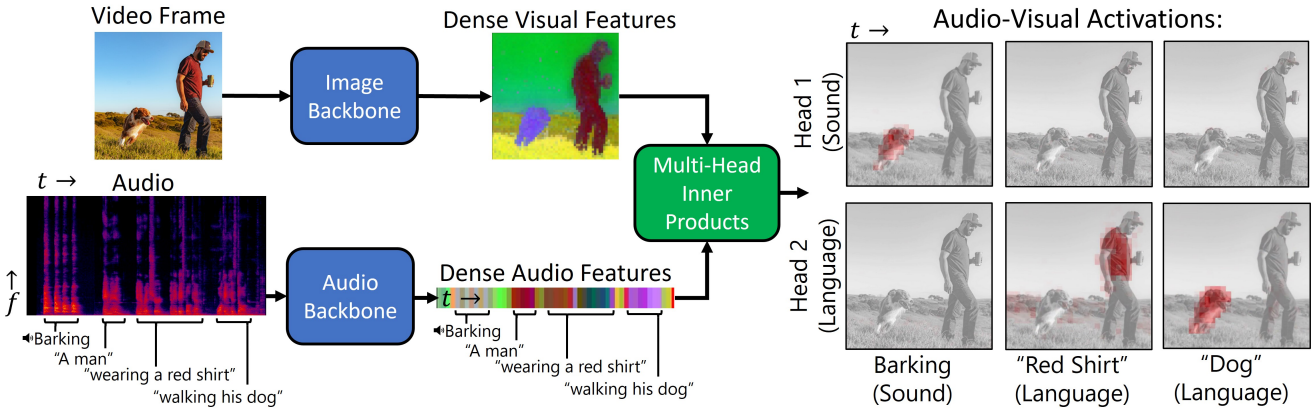

Figure 1. Visual overview of the DenseAV algorithm. Two modality-specific backbones featurize audio and visual signals. We introduce a novel generalization of multi-head attention to extract attention maps that discover and separate the "meaning" of spoken words and the sounds an object makes. DenseAV performs this localization and decomposition solely through observing paired stimuli such as videos.

Abstract

We present DenseAV, a novel dual encoder grounding architecture that learns high-resolution, semantically meaningful, and audio-visual aligned features solely through watching videos. We show that DenseAV can discover the "meaning" of words and the "location" of sounds without explicit localization supervision. Furthermore, it automatically discovers and distinguishes between these two types of associations without supervision. We show that DenseAV's localization abilities arise from a new multi-head feature aggregation operator that directly compares dense image and audio representations for contrastive learning. In contrast, many other systems that learn "global" audio and video representations cannot localize words and sound. Finally, we contribute two new datasets to improve the evaluation of AV representations through speech and sound prompted semantic segmentation. On these and other datasets we show DenseAV dramatically outperforms the prior art on speech and sound prompted semantic segmentation. DenseAV outperforms the current state-of-the-art, ImageBind, on cross-modal retrieval using fewer than half of the parameters. Project Page: https://aka.ms/denseav

1. Introduction

Associating audio and video is a fundamental aspect of human perception. As infants develop, the synchronization and correspondence of sounds enables multi-modal association – a voice with a face, and a 'moo' with a cow [50]. Later, as they acquire language, they associate spoken words with objects they represent [10, 45]. Amazingly, these association abilities, constituting speech recognition, sound event recognition, and visual object recognition, develop without much direct supervision. This work aims to create a model with this capability by learning high-resolution, semantically meaningful, audio-visual (AV) aligned representations. Features with these properties can be used to discover fine-grained correspondences between modalities without localization supervision or knowledge of the textual representation of language.

As an example, consider the spoken caption and accompanying sounds of the image shown in Figure 1. We wish to "ground" both the speech and the sounds with high resolution. For instance, both the spoken word "dog" and the sound of a bark in the audio signal should be associated with the pixels of the dog in the visual signal if present. We seek high quality local representations where simple inner

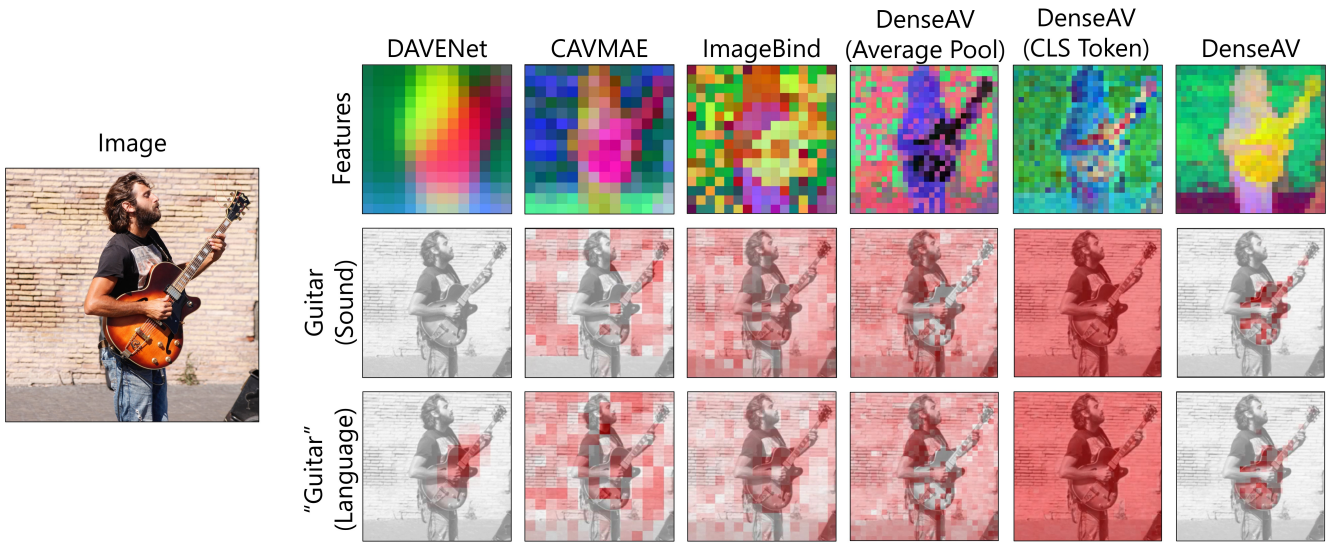

Figure 2. Qualitative comparison of several modern architectures for associating audio and video modalities. Only DenseAV learns a high-resolution and semantically aligned set of local features. This allows us to perform speech and sound prompted semantic segmentation using only the inner products between deep features. Other approaches, such as ImageBind, do not show aligned local feature maps. Approaches that do show some localization capabilities, like DAVENet, do not generalize to sound and language, and do not achieve the high-resolution localization capabilities of DenseAV. Dense features are visualized using PCA as in [20]

products between features exhibit this behavior, which is interestingly absent from popular approaches in the literature.

To achieve this, we make three innovations. First, we introduce DenseAV, a dual-encoder architecture that computes a dense similarity volume over audio and visual features before aggregating this volume into a single similarity score. If we look at a slice of this similarity volume when word is spoken, as in Figure 1, we can visualize the AV activation strength between a word or sound and an image's pixels. The novelty we introduce is to generalize this dense similarity mechanism with multiple heads, much like those of multi-head attention. This generalization allows each head to specialize on particular types of couplings between the visual and audio modalities. Interestingly, we discover that if we give DenseAV two heads and train on a dataset that contains both language and sound, the heads naturally learn to distinguish language from more general sound using only cross-modality supervision. For example, as shown in Figure 1, head 1 focuses on what objects create a sound, like the barking of a dog, whereas head 2 focuses only on the meaning of words.

Second, we show the importance of the "aggregation function" one uses to create a similarity score between an audio clip and a video frame for contrastive learning. The traditional choices; using inner products between global representations such as class tokens [5, 13, 49] or pooled features [18, 58], do not promote AV alignment of dense local features. Because of this, several popular audio-video backbones that excel on cross-modal retrieval *cannot* directly associate objects and sounds using their local features. This limits their ability to be used for downstream tasks such as semantic segmentation, sound localization, or unsupervised language learning and discovery.

Third, we introduce two semantic segmentation datasets to evaluate visual grounding with AV representations for speech and (non-speech) sounds. We build these datasets from the high-quality segmentation masks provided by the ADE20K dataset [59] and measure mean average precision (mAP) and mean intersection over union (mIoU) on a binary mask prediction task. This evaluation is simpler and more thorough than previous efforts to measure visual grounding such as the concept counting metrics of [23] and the "pointing games" of [2, 14, 38] that only check if a heatmap's peak occurs within a target box or segment. Furthermore, our evaluation avoids brittle word-net ontologies [34], clustering, Wu and Palmer distance [55], threshold choices, and a variety of other complicating factors.

To summarize, our main contributions are as follows:

- We introduce DenseAV, a novel self-supervised architecture that learns high-resolution AV correspondences.
- We introduce a local-feature-based image similarity function that significantly improves a network's zero-shot localization ability compared to common strategies such as average pooling or CLS tokens.
- We introduce new datasets for evaluating speech and sound prompted semantic segmentation. We show DenseAV significantly outperforms the current state-of-the-art on these tasks as well as on cross-modal retrieval.
- We discover that our multi-head architecture naturally disentangles audio-visual correspondence into sound and language components using only contrastive supervision.

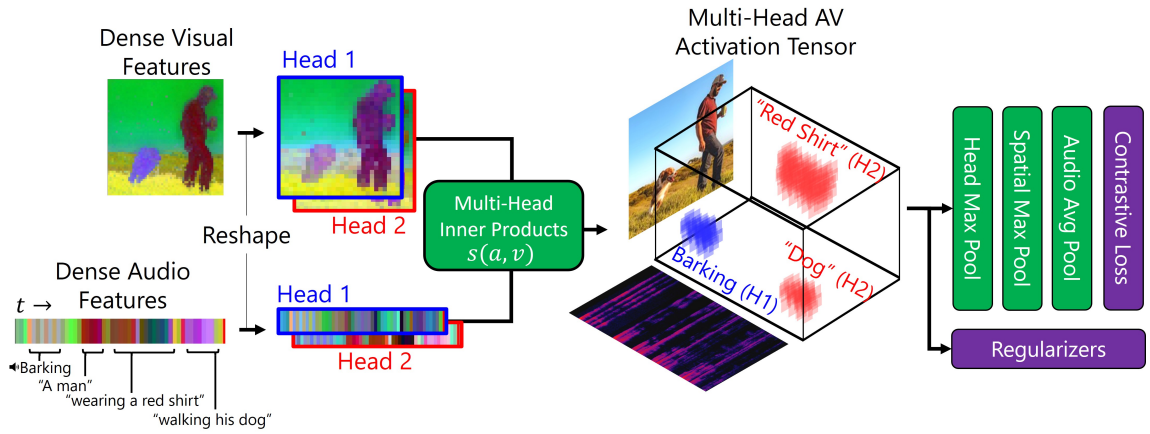

Figure 3. Architectural overview of our multi-head attention aggregator. Dense feature maps are split into K heads ($K = 1, 2$) in our experiments. We form an AV activation tensor by taking the inner-products of each head's features across the spatial and temporal extent of the visual and audio signals respectively as in Equation 1. We then aggregate this similarity volume into a single similarity score by max-pooling head and spatial dimensions and average-pooling audio dimensions. Our approach aims to encourage the network to identify specific shared objects between the audio and visual modalities. In particular, max-pooling of heads disentangles sound and language, and max-pooling spatial dimensions helps localize objects.

2. Related Work

Audio-visual (AV), text-visual, and more general multi-modal models have recently surged in popularity [60]. Broadly speaking DenseAV is an audio-video contrastive learning architecture, this class of methods learns AV representations by aligning paired signals and pushing apart negative signals [11, 27]. Of the models in this class, several stand out for their ability to localize sounds [2, 7, 42] or capture the semantics of language [23, 43]. Many models in this class compare AV signals using inner products between "global" representations formed by pooled deep features [18, 35, 53], or class tokens [17, 32, 42, 43, 49]. Most notably, ImageBind has gained popularity due to its state-of-the-art performance on a variety of tasks and datasets and unified class-token-based contrastive architecture. In this work we show that many of these architectures do not show strong localization properties in their local features, despite excelling at cross-modal retrieval on a "global" level. This limits their applicability to new out-of-domain sounds, sounds that don't have a textual representation, and low-resource languages. We diverge from these works by directly supervising local tokens. In particular, we build on previous works [2, 23] that show max-pooling improves localization capabilities and introduce a new multi-head aggregation operator that generalizes previous losses using an operator similar to self-attention [52].

Another class of methods discover structure in signals through uni- and multi-modal clustering. Early works on audio clustering [41] discovered meaningful utterances without supervision. Similar visual analyses have discovered visual objects [4, 8, 20, 28]. Recent works have applied these ideas to the AV domain [1, 21], but do not focus on extracting *high-resolution* AV representations.

Finally, several works investigate generative audio-video learning. The Sound of Pixels [57] generates the sound of a specific object using a source separation loss. Newer approaches using GANs [30, 31], and diffusion models [9, 17, 33] have generated audio from video and vice versa. Here we focus on improving the local representations of contrastive learners because of their relative scalability, simplicity, and ability to learn high-quality representations.

3. Methods

At a high level, DenseAV tries to determine when a given audio and visual signal belong "together" using dense audio-visual representations. To perform this task robustly, DenseAV must learn how to predict the contents of an audio signal from a visual signal and vice versa. This causes DenseAV to learn dense modality-specific features that capture the mutual information shared between the modalities [51]. Once learned, we can directly query these informative features to perform speech and sound prompted semantic segmentation as illustrated in Figure 1.

More specifically, DenseAV is built from two modality-specific deep featurizers. These backbones produce temporally varying audio features across an audio clip and spatially varying video features for a single randomly selected frame. Our loss computes a similarity between audio and visual signals based on the intuition that two signals are similar if they have a variety of strong couplings or shared objects. More formally, we form a scalar similarity for a pair of audio and video signals by carefully aggregating a volume of pairwise inner products between dense features. We use the InfoNCE [36] contrastive loss to encourage similarity between "positive" pairs of signals and dissimilarity between "negative" pairs formed by in-batch shuffling. Figure 3 graphically depicts this loss function and subsequent

sections detail each component of our architecture.

3.1. Multi-Headed Aggregation of Similarities

DenseAV's key architectural distinction is its loss function that directly supervises the "local" tokens of the visual and audio featurizers. This is a significant departure from other works [5, 17, 19, 39, 46, 49] that pool modality specific information into "global" representations prior to the contrastive loss. Unlike prior works, our loss function aggregates the full pairwise similarities between the local tokens into an aggregate measure of similarity for a given pair of audio and visual signals. We show in Figure 2 that this architectural choice enables DenseAV's local features to align across modalities whereas other approaches such as average pooling, class tokens, and SimPool [44] do not.

We first describe our loss function informally and definite it more precisely in the next paragraph. Our loss function computes the (un-normalized) inner product between every pair of visual and audio features to form a "volume" of inner products. This volume represents how strongly each part of an audio signal "couples" to each part of a visual signal. We aim to find many large couplings between positive pairs of audio and visual signals. Ideally, these couplings should connect visual objects with their references in the audio signal. Conversely, we do not want to find couplings between negative pairs of signals. To compute a single global coupling strength for a pair of signals, we aggregate this volume of pairwise similarities into a single number. There are myriad ways to aggregate this volume ranging in "softness" from average-pooling to max-pooling. Average pooling yields dense gradients and can improve convergence speed and stability. However, max-pooling allows the network to focus on the *best* couplings regardless the object's size or a sound's duration. Our aggregation function combines the benefits of average and max pooling by max-pooling visual dimensions and average pooling audio dimensions as proposed in [23]. Intuitively speaking, this averages the strongest image couplings over an audio signal. It allows small visual objects to have large effects yet provides a strong training gradient to many regions of the signals. Finally, we draw inspiration from multi-head self-attention [52] and generalize this operation to multiple "heads" that we max-pool before pooling the visual and audio dimensions. This allows DenseAV to discover multiple "ways" to associate objects across modalities.

More formally, let $\mathcal{S}(a, v) \in \mathbb{R}$ represent the similarity between a tensor of audio features $a \in \mathbb{R}^{CKFT}$ of size (**C**hannel $\times$ **K**-heads $\times$ **F**requency $\times$ **T**ime) and a tensor of visual features $v \in \mathbb{R}^{CKHW}$ of size (**C**hannel $\times$ **K**-heads $\times$ **H**eight $\times$ **W**idth). To define this scalar similarity score, we first create a local similarity volume, $s(a, v) \in \mathbb{R}^{kfthw}$. For simplicity, we consider the aggregated similarity between a single image and audio clip but note one can easily generalize this to max-pool over video-frames. We define the full pairwise volume of similarities as:

$$s(a, v) \in \mathbb{R}^{kfthw} = \sum_{c=1}^{C} a[c, k, f, t] \cdot v[c, k, h, w] \quad (1)$$

Where $a[c, k, f, t]$ represents the value of a at location $[c, k, f, t]$ and $\cdot$ is scalar multiplication. We aggregate this similarity volume into a single score $\mathcal{S}(a, v) \in \mathbb{R}$:

$$\mathcal{S}(a, v) = \frac{1}{FT} \sum_{f=1}^{F} \sum_{t=1}^{T} \max_{k,h,w} \left(s(a, v)[k, f, t, h, w] \right) \quad (2)$$

We note that this operation can be viewed as a multi-head generalization of the MISA loss of [23], and a multi-head multi-time generalization of the MIL loss of [2].

3.2. Loss

We can use the similarity between audio and visual signals defined in Equation 2 to construct a contrastive loss. We follow recent works [15, 17, 54] and use the temperature-weighted InfoNCE [36] to encourage similarity between positive pairs of signals and dissimilarity between negative pairs. In DenseAV, we form B positive pairs by splitting the audio and visual components of a **B**atch of training data. We form $B^2 - B$ negative pairs by comparing a signal to all of the other signals in the training batch. More formally let $(a_b, v_b)_1^B$ be B pairs of audio and visual signals. The visual-retrieval term of our InfoNCE loss is then:

$$\mathcal{L}_{A \to V} = \frac{1}{2B} \sum_{i=1}^{B} \left(\log \frac{\exp\left(\gamma \mathcal{S}(a_b, v_b)\right)}{\sum_{b'=1}^{B} \exp\left(\gamma \mathcal{S}(a_b, v_{b'})\right)} \right) \quad (3)$$

Where $\gamma \in \mathbb{R}^+$ is a trainable inverse temperature parameter. We symmetrize this loss by adding the analogous audio-retrieval term, $\mathcal{L}_{V \to A}$, which iterates over negative *audio* signals in the denominator.

3.3. Audio and Visual Featurizers

The core of DenseAV is two modality-specific backbone networks. We use the DINO vision transformer [5] with ImageNet pretrained weights (without labels) to provide a strong, yet fully unsupervised, vision backbone. Unlike other approaches that use CLIP [46] as a backbone, DINO does not require paired text captions and learns from unlabeled images only. Practically, we find that DINO outperforms CLIP because of its better-behaved local tokens [12], an effect we explore in the Supplement. We append an additional layer norm operation across the channel dimension [3] and a 1×1 Convolution to DINO. The layer-norm and 1×1 convolution ensure the architecture does not start with a saturated loss function. We use the HuBERT audio transformer [25] as DenseAV's audio backbone. HuBERT operates on waveforms and is trained on the LibriSpeech [40] dataset using only self-supervision. Hubert

outputs a single feature per frequency, corresponding to $F = 1$ in Section 3. Though HuBERT was only trained on speech, its audio features can be fine-tuned for more general sounds, much like how vision backbones can be fine-tuned for new datasets [56]. As in the visual branch, we append a channel-wise LayerNorm block and two 3×3 convolutions to the audio branch. These layers help the network avoid saturation and speed convergence. Furthermore, the two convolutions help the model aggregate information, which reduces the cost of the pairwise feature comparison used in our loss function. We refer to these added layers after the pretrained backbones as the "aligners" in later sections.

3.4. Regularizers

Disentanglement Regularizer, $\mathcal{L}_{Dis}$: We add a small regularization term to encourage each head of Equation 1 to specialize and learn independent types of audio-visual associations. Interestingly we find that our 2-head model naturally learns to distinguish the meaning of words with one head and capture the sounds objects produce with another head. To further encourage this unsupervised discovery of concepts, we penalize the network when multiple attention heads are simultaneously active. More precisely, let $(a_b, v_b)_1^B$ be a **B**atch of B paired audio and visual signals. Our disentanglement loss for two heads is then:

$$\mathcal{L}_{Dis} = \text{Mean}(|s(a_b, v_b)[1] \circ s(a_b, v_b)[2]|) \quad (4)$$

Where $\circ$ is elementwise multiplication and $|\cdot|$ is the elementwise absolute value function. $[k]$ mirrors PyTorch slicing notation and refers to selecting the activations for only the kth attention head. Intuitively, this loss encourages one head to be silent if the other head is active and is a "cross-term" generalization of the l^2 regularizer [24] for encouraging activation shrinkage. When $K > 2$ we average contributions from every combination of heads. We ablate this, and our decision to max-pool heads in Table 3.

Stability Regularizers, $\mathcal{L}_{Stability}$: Finally, we add several other small regularization terms to encourage stable convergence. We detail and ablate these terms in the Supplement. Briefly, these terms include standard regularizers like Total Variation [48] smoothness over time and non-negative pressure to encourage the network to focus on similarity instead of dissimilarity. In addition, we add a regularizer to prevent the calibration temperature, γ, from drifting too quickly, and a regularizer to discourage activations during silence and noise. In the supplement we show that each regularizer alone does not have a dramatic effect on final metrics but together they can stop collapses during training.

Combining these losses into a single loss function yields:

$$\mathcal{L} = \mathcal{L}_{A \to V} + \mathcal{L}_{V \to A} + \lambda_{Dis}\mathcal{L}_{Dis} + \mathcal{L}_{Stability} \quad (5)$$

In our experiments we use $\lambda_{Dis} = 0.05$ and refer interested readers to the supplement for the details of our small stability regularizer, $\mathcal{L}_{Stability}$.

3.5. Training

In our experiments we train DenseAV and relevant baselines on the AudioSet [16] dataset for sound prompted segmentation and AudioSet retrieval. We train on the PlacesAudio [22] dataset for speech prompted segmentation, PlacesAudio retrieval, and the ablation studies of Table 4. In our disentanglement experiments of Table 3 and feature visualizations of Figures 1 and 2 we train on both AudioSet and PlacesAudio so that DenseAV can be familiar with both language, the prominent audio signal in PlacesAudio, and more general sounds from AudioSet. In these experiments we sample training data from these two corpora, so each batch has an even split between AudioSet and PlacesAudio.

Warming up Aligners: We find that we can dramatically improve the stability by first training the added aligners (convolutions and layer norms) for 3000 steps while keeping pretrained DINO and HuBERT backbones fixed. This allows the aligners to adapt to these intelligent backbones before modifying each backbone's sensitive weights. We use random resize crops, color jitter, random flips, and random greyscaling as image augmentations. We randomly sample a single video frame to feed to our visual branch. Audio clips are converted to single-channel format and are trimmed or padded with silence to create uniform 10 second clips. We re-sample audio clips according to the requirements of the backbone models used. For HuBERT, we re-sample to 16KhZ. We train on 8 V100 GPUs with an effective batch size of 80, and aggregate negative samples on all GPUs prior to computing the loss to ensure efficient parallelization. We provide additional training information and hyperparameters in the supplement.

Full Training: After warming up the aligners, we train the full model for an additional 800,000 steps using the same loss, batch-size, and training logic. We train all aligner weights and fine-tune all HuBERT audio backbone weights. We use low rank adaptation (LoRA) [26] to fine-tune the "Q", "K", and "V" layers of the DINO visual backbone attention blocks. This allows us to efficiently adapt DINO and stabilize the training as it is quite easy to collapse the carefully trained DINO weights. We use a LoRA rank of 8.

4. Experiments

To evaluate AV representation quality, we perform a variety of analyses including comparative activation visualization, quantitative measurements of speech and sound prompted semantic segmentation, and cross-modal retrieval. Additionally, we quantify our observation that DenseAV can distinguish the meanings of words (language), from the sounds of objects (sound) without supervision.

To adequately measure a representation's AV alignment quality, we found it necessary to introduce two evaluation datasets that measure speech and sound prompted seman-

Method	Speech Semseg. mAP	Speech Semseg. mIoU	Sound Semseg. mAP	Sound Semseg. mIoU
DAVENet [23]	32.2%	26.3%	16.8%	17.0%
CAVMAE [18]	27.2%	19.9%	26.0%	20.5%
ImageBind [17]	20.2%	19.7%	18.3%	18.1%
Ours	**48.7%**	**36.8%**	**32.4%**	**25.5%**

Table 1. **Speech and Sound prompted semantic segmentation**. We analyze the quality of local features using two prompted semantic segmentation tasks. We prompt networks with speech of the form "a picture of a(n) [Object]" to determine whether local feature inner products can segment objects in the ADE20K dataset by name. We create sound prompts for a given ADE20K class using a curated mapping from the ADE20K ontology to the VGSound ontology. DenseAV's local features perform significantly better than all baselines investigated. We bold "first place" results and underline "second place" results.

Method	Places Acc. @10 I → A	Places Acc. @10 A → I	AudioSet Acc. @10 I → A	AudioSet Acc. @10 A → I
[22]*	46.3%	54.8%	-	-
[21]*	54.2%	56.4%	-	-
DAVENet [23]*	52.8%	60.4%	-	-
CAVMAE [18]	81.7%	77.7%	55.7%	50.7%
ImageBind [17]	1.10%	1.10%	64.5%	66.5%
Ours	**94.2%**	**94.3%**	**68.2%**	**68.4%**

Table 2. **Cross-modal retrieval using 1000 evaluation videos from the PlacesAudio and AudioSet validation datasets**. DenseAV dramatically outperforms all approaches tested in all metrics. Most notably, the state-of-the-art image retrieval foundation model, ImageBind, is incapable of recognizing speech. We note that the ImageBind authors do not publish retraining code, so we evaluate their largest pretrained model. Models with a * indicate that they have been previously reported in the literature. Other numbers are calculated by using pretrained models when available or from training with the author's official training scripts.

tic segmentation performance. Our two datasets introduce pairs of speech and sound prompts coupled with matching images and segmentation masks derived from ADE20K. We create these datasets because previous works [23] have not published their datasets or evaluation code. However, we use an experimental setting from the literature for our cross-modal retrieval experiments.

We compare against a variety of prior art including the popular state-of-the art multi-modal retrieval network, ImageBind [17]. We also compare against CAVMAE [18], a leading multimodal backbone trained specifically for AudioSet retrieval, and DAVENet [23], which is trained to localize the meanings of words. We include two other baselines [21, 22] which have reported cross modal retrieval metrics on Places Audio. Finally, we compare our multi-head aggregation strategy to common "global" retrieval methods such as inner products between class-tokens, average-pooled tokens, and SimPooled[44] tokens. We note that SimPool achieves state-of-the-art localization results when compared to 14 other pooling methods. Nevertheless, our multi-head aligner yields better localization results than any of these "global" methods.

4.1. Qualitative Comparison of Feature Maps

Our first experiment in Figure 2 highlights the dramatic differences in quality between DenseAV's features and other approaches in the literature. DenseAV is the only backbone whose local tokens are semantically meaningful and show cross-modal alignment for speech and sound. Though both CAVMAE and ImageBind show high-quality retrieval performance, neither shows high quality aligned local tokens. As a result, DenseAV can associate and localize both sound and language significantly better than other backbones. DAVENet shows coarse correspondences between language and visual objects but cannot associate sound with visual objects and does not match DenseAV's high resolution maps. Furthermore, the right half of Figure 1 demonstrates that DenseAV naturally discovers and separates word semantics from the sound of objects without labels to supervise this separation. In the supplement, we provide additional visualizations of all backbones considered across a wide range of words and sounds.

4.2. Speech Prompted Image Segmentation

Dataset: We introduce a speech prompted segmentation dataset using the ADE20K dataset, which is known for its comprehensive ontology and pixel-precise annotations [59]. From this dataset, we curate an evaluation subset of image-class pairs by sampling up to 10 images for each object class in ADE20K, excluding images where the selected class was tiny ($< 5\%$ of pixels). We only consider classes with at least 2 images that pass the tiny object criterion. For each class and image, we formed a binary target mask by selecting the semantic segmentation mask for that class. This resulted in 3030 image-object pairs spanning 478 ADE20K classes.

We created paired speech signals by speaking the prompt "A picture of a(n) [object]" where [object] is the name of the ADE20K class. We create clear, controlled, and consistent audio prompts using Microsoft's neural text to speech service [47]. This service also provides exact timing of the "[object]" utterance within the broader prompt and ensures each class is measured equally. Grammar was manually verified for the utterances to ensure proper singular/plural and a/an agreement with the class name. We release images, masks, and audio prompts for reproducibility.

Evaluation Measure: We evaluate methods based on how well their speech-prompted activations align with ground truth masks for the visual object's class. We quantify this with the binary Average Precision (AP) and Intersection over Union (IoU) metrics. These quantify how close activations match with the binary label mask from the ADE20K dataset. To compute an aggregate score over all of the object

Method	Pred. Dis.	Act. Dis.
No $\mathcal{L}_{Dis}$, No Head Max Pool	64.1%	70.3%
No $\mathcal{L}_{Dis}$	**99.9%**	86.5%
Ours	**99.9%**	**91.2%**

Table 3. Quantitative ablation study of the impact of max-pooling attention heads and adding our disentanglement loss, $\mathcal{L}_{Dis}$. Intuitively, max-pooling attention heads allows each head to specialize on its own specific set of triggers. Our disentanglement loss further encourages the heads to operate independently and orthogonally.

classes considered, we compute the mean average precision (mAP) and mean intersection over union (mIoU) by averaging AP scores across all object categories considered.

The mAP is particularly well suited for evaluating feature similarities because it is unaffected by monotonic transformations of the similarity scores. This eliminates the need for arbitrary thresholding and calibration. This is particularly important because many networks' inner products are not centered at zero, and the best thresholding strategy can be nontrivial, and dependent on the network and object class. Average Precision avoids these confounding factors and ensures a fair comparison across methods. Unfortunately, unlike the mAP, the mIoU metric requires selecting a threshold. To ensure our mIoU measurement is similarly invariant to monotonic transformations we evaluate 20 uniformly spaced thresholds between the smallest and largest activations of each model. For each baseline, we report results for the best threshold to ensure a fair comparison between all networks considered.

Implementation: We compute image heatmaps by evaluating each modality-specific network on the image-audio pairs from our dataset. We extract dense features from the final layer of each network and form their similarity volume according to Equation 1. For DenseAV we max-pool the head dimension to properly compare with single-headed models. We average activations over the temporal extent of the "[object]" utterance using the word timing information from the ground truth audio clip. This creates a heatmap over the image features that can be bi-linearly resized to the original image's size. We then compare these per-pixel activation scores to ground truth object masks from our dataset.

Results: In Speech mAP and mIoU columns of Table 1 we show that DenseAV achieves a **51% (+16.5 mAP)** relative increase in speech-prompted semantic segmentation over previous methods. Approaches that use global token based contrastive strategies such as CAVMAE and ImageBind perform particularly poorly in this task, and this observation aligns with the qualitative results of Figure 2.

4.3. Sound Prompted Image Segmentation

Dataset: To evaluate how well deep features localize sound, we build on Section 4.2 and create a dataset of sound prompts that align with ADE20K classes. We first select the same (large) image-object pairs from ADE20K. We then create a mapping between the ADE20K and VGGSound [6] ontologies. To compute a robust mapping, we first embed ADE20K class names and VGGSound class names with the GPT Ada 2 text embedding model [37]. For each ADE20K class, we create a list of at most three candidates from the VGGSound ontology that have a cosine similarity ($> .85$). We then manually review these candidates to select the best VGGSound class for each ADE20K class and remove any spurious or mistaken matches. This produces a set of 95 ADE20K classes with strong matches in the VGGSound ontology. For each of our original 3030 image-object pairs we select a random VGGSound validation clip with a matching class according to our mapped ontology. This yields 106 image-object pairs across 20 ADE20K classes.

Evaluation Measure: We use the same mAP and mIoU evaluation metrics as Section 4.2, but instead average over the 20 ADE20K classes considered.

Implementation: We compute sound prompted image activations as in section 4.2 but with one key change: we average activations over the entire clip because we do not have ground-truth sound timing information.

Results: The "Sound mAP and mIoU" columns of Table 1 show that DenseAV achieves a **25% (+6.4mAP)** relative improvement in sound prompted segmentation compared to the prior art. Most notably, ImageBind's features cannot localize sound despite their high cross-modal retrieval performance learned from millions of hours of sound.

Method	Speech mAP	Places Acc. @10	
		V $\rightarrow$ A	A $\rightarrow$ V
Average Pool	20.1%	92.0%	91.2%
CLS Token	20.6%	86.4%	89.8%
SimPool [44]	35.3%	92.6%	92.8%
Multi-Head (Ours)	**48.2%**	**93.5%**	**93.8%**

Table 4. Quantitative ablation of different feature aggregation strategies. Though the common practice of average pooling and using a learned CLS token to aggregate features have little effect on retrieval performance, they dramatically degrade performance on speech prompted semantic segmentation.

4.4. Cross-Modal Retrieval

We show that DenseAV's representations are not only better for localization, but significantly outperform other approaches on cross-modal retrieval. We adopt the evaluation setting of [23] and measure cross modal retrieval accuracy at 1, 5, and 10 in a thousand-way retrieval task. In particular, we use the same thousand images from the validation set of [23] and also replicate this analysis on one-thousand random clips from the AudioSet validation data. Table 2 shows results for 1000-way retrieval tasks on both the Places Audio and AudioSet datasets. We show cross-modal accuracy at 10, but also show larger tables in the supplement that echo these results using accuracy at 1 and 5. DenseAV significantly outperforms all baselines across all metrics. Interest-

ingly, DenseAV outperforms ImageBind with *less than half* of the trainable parameters and no reliance on text.

4.5. Measuring Disentanglement

We observe that DenseAV's heads naturally learn to differentiate audio-visual couplings that capture the meaning of words (language) and those that capture the sounds of objects (sound). Furthermore this effect generalizes to novel clips, including those with both sound and language as shown in Figure 1. We quantify this observation in two ways, the first measures if a head's average activation strength predicts whether a clip contains mainly "language" or "sound". The second method quantifies how often the "sound" head is incorrectly active when the "language" head should be active and vice versa. We leverage the fact that AudioSet dataset contains mostly clips with ambient sound and rarely contains language. In contrast, Places Audio is entirely language-based without external ambient sound. We note that these analyses are specifically for our architecture with two heads $K = 2$ and trained on both AudioSet and PlacesAudio data.

For both measures of disentanglement, we first compute a clip's aggregated similarity for each head. In particular, we remove the max-pooling over heads in Equation 2 to create a single-head similarity, $\mathcal{S}(a, v)_k$. We then min-max scale the scores of each head across both datasets to lie in the $[0, 1]$ interval, which we refer to as $\hat{\mathcal{S}}(a, v)_k$. Using these normalized scores, we can create metrics that capture how well a given head responds only to a specific dataset.

Our first metric measures how well a head's scores predict whether a clip is from the "sound" or "language" dataset. Let $(a_b, v_b)_1^B$ be tuples of paired audio and visual signals. let $l[k']_b$ be an indicator variable of whether the signal (a_b, v_b) arises from the sound dataset, AudioSet, $(k' = 1)$, or the language dataset Places Audio $(k' = 2)$.

$$\delta_{pred}(k, k') = \text{AP}\left((\hat{\mathcal{S}}(a_b, v_b)_k)_1^B, (l[k']_b)_1^B\right) \quad (6)$$

Where $AP(\cdot, \cdot)$ is the binary average precision with prediction and label arguments respectively. Intuitively, this measures whether the scores of head k are direct predictors of whether the data is from dataset k'. We can find the best assignment between heads and datasets such that each head is maximally predictive of the given dataset:

$$\text{PredDis} = \frac{1}{2}\max\left(\delta_{pred}(0,0) + \delta_{pred}(1,1), \delta_{pred}(1,0) + \delta_{pred}(0,1)\right) \quad (7)$$

The prediction disentanglement score, PredDis, is a percentage that ranges from 50% for completely entangled signals to 100% if one can perfectly classify the signals using the scores of either head. The maximum over the two possible assignments makes this metric invariant to permutations of the heads. We note that this metric is a Hungarian matching assignment [29] over two entries, a common technique to asses unsupervised classification performance [20, 28].

Our second measure quantifies "spurious activations" in the non-dominant head. A truly disentangled system should have a head that only fires on sound, and another head that only fires on language. We create another disentanglement measure, ActDis, by replacing δ_{pred} in Equation 7 with:

$$\delta_{act}(k, k') = 1 - \frac{1}{\sum_{b'} l[k']_{b'}} \sum_{b=1}^{B} \hat{\mathcal{S}}(a_b, v_b)_k \cdot l[k']_b \quad (8)$$

Intuitively, this measures the "inactivity" of head k on dataset k'. If head k is totally silent on dataset k' then $\delta_{act}(k, k') = 1$. Like PredDis, ActDis is a percentage ranging from 50% to 100% with 100% representing perfect disentanglement where the sound head is completely silent during the language clips, and vice versa.

Table 3 shows that DenseAV achieves near perfect predictive (99%) and activation (91%) disentanglement. It also shows that our disentanglement regularizer and max-pooling over heads improves DenseAV's natural ability to distinguish sound from language without supervision.

5. Conclusion

We presented DenseAV, a novel contrastive learning architecture that can discover the meaning of words and localize the sounds of objects using only video supervision. We are the first to observe both qualitatively and quantitatively that it's possible to disentangle the meaning of words from the sound of objects with only a contrastive learning signal. DenseAV's success stems from its novel multi-head attention aggregation mechanism that encourages its modality-specific backbones to create high-resolution, semantically meaningful, and AV aligned representations. These properties of DenseAV's representation are not seen in other state-of-the-art models in the literature. Consequently, DenseAV significantly surpasses other leading models in dense prediction tasks such as speech and sound-prompted semantic segmentation as well as in cross-modal retrieval.

Acknowledgements

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 2021323067. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors(s) and do not necessarily reflect the views of the National Science Foundation. This work is supported by the National Science Foundation under Cooperative Agreement PHY-2019786 (The NSF AI Institute for Artificial Intelligence and Fundamental Interactions, http://iaifi.org/). This work is funded by a Royal Society Research Professorship RSRP\R\241003, and EPSRC Programme Grant VisualAI EP/T028572/1.

References

[1] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. *Advances in Neural Information Processing Systems*, 33:9758–9770, 2020. 3

[2] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *Proceedings of the European conference on computer vision (ECCV)*, pages 435–451, 2018. 2, 3, 4

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 4

[4] Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze. Deep clustering for unsupervised learning of visual features. In *Proceedings of the European conference on computer vision (ECCV)*, pages 132–149, 2018. 3

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 2, 4

[6] 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. 7

[7] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16867–16876, 2021. 3

[8] Jang Hyun Cho, Utkarsh Mall, Kavita Bala, and Bharath Hariharan. Picie: Unsupervised semantic segmentation using invariance and equivariance in clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16794–16804, 2021. 3

[9] Jeongsoo Choi, Joanna Hong, and Yong Man Ro. Diffv2s: Diffusion-based video-to-speech synthesis with vision-guided speaker embedding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7812–7821, 2023. 3

[10] Noam Chomsky. *Language and problems of knowledge: The Managua lectures*. MIT press, 1987. 1

[11] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, pages 539–546. IEEE, 2005. 3

[12] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023. 4

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2

[14] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE international conference on computer vision*, pages 3429–3437, 2017. 2

[15] Nicholas Frosst, Nicolas Papernot, and Geoffrey Hinton. Analyzing and improving representations with the soft nearest neighbor loss. In *International conference on machine learning*, pages 2012–2020. PMLR, 2019. 4

[16] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017. 5

[17] 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. 3, 4, 6

[18] Yuan Gong, Andrew Rouditchenko, Alexander H Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James R Glass. Contrastive audio-visual masked autoencoder. In *The Eleventh International Conference on Learning Representations*, 2022. 2, 3, 6

[19] 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. 4

[20] Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T Freeman. Unsupervised semantic segmentation by distilling feature correspondences. *arXiv preprint arXiv:2203.08414*, 2022. 2, 3, 8

[21] David Harwath and James R Glass. Learning word-like units from joint audio-visual analysis. *arXiv preprint arXiv:1701.07481*, 2017. 3, 6

[22] David Harwath, Antonio Torralba, and James Glass. Unsupervised learning of spoken language with visual context. *Advances in Neural Information Processing Systems*, 29, 2016. 5, 6

[23] David Harwath, Adria Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. Jointly discovering visual objects and spoken words from raw sensory input. In *Proceedings of the European conference on computer vision (ECCV)*, pages 649–665, 2018. 2, 3, 4, 6, 7

[24] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. 5

[25] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. 4

[26] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen.

Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 5

[27] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. *Technologies*, 9(1): 2, 2020. 3

[28] Xu Ji, João F Henriques, and Andrea Vedaldi. Invariant information clustering for unsupervised image classification and segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9865–9874, 2019. 3, 8

[29] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 8

[30] Neeraj Kumar, Srishti Goel, Ankur Narang, and Mujtaba Hasan. Robust one shot audio to video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 770–771, 2020. 3

[31] Shiguang Liu, Sijia Li, and Haonan Cheng. Towards an end-to-end visual-to-raw-audio generation with gan. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(3): 1299–1312, 2021. 3

[32] Shuang Ma, Zhaoyang Zeng, Daniel McDuff, and Yale Song. Active contrastive learning of audio-visual video representations. *arXiv preprint arXiv:2009.09805*, 2020. 3

[33] Yuxin Mao, Jing Zhang, Mochu Xiang, Yunqiu Lv, Yiran Zhong, and Yuchao Dai. Contrastive conditional latent diffusion for audio-visual segmentation. *arXiv preprint arXiv:2307.16579*, 2023. 3

[34] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 2

[35] Mathew Monfort, SouYoung Jin, Alexander Liu, David Harwath, Rogerio Feris, James Glass, and Aude Oliva. Spoken moments: Learning joint audio-visual representations from video descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14871–14881, 2021. 3

[36] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3, 4

[37] OpenAI. Gpt-4 technical report, 2023. 7

[38] Maxime Oquab, Léon Bottou, Ivan Laptev, and Josef Sivic. Is object localization for free?-weakly-supervised learning with convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 685–694, 2015. 2

[39] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2023. 4

[40] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE, 2015. 4

[41] Alex S Park and James R Glass. Unsupervised pattern discovery in speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(1):186–197, 2007. 3

[42] Puyuan Peng and David Harwath. Self-supervised representation learning for speech using visual grounding and masked language modeling. *arXiv preprint arXiv:2202.03543*, 2022. 3

[43] Puyuan Peng and David Harwath. Word discovery in visually grounded, self-supervised speech models. *arXiv preprint arXiv:2203.15081*, 2022. 3

[44] Bill Psomas, Ioannis Kakogeorgiou, Konstantinos Karantzalos, and Yannis Avrithis. Keep it simpool: Who said supervised transformers suffer from attention deficit? In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5350–5360, 2023. 4, 6, 7

[45] Geoffrey K Pullum and Barbara C Scholz. Empirical assessment of stimulus poverty arguments. *The linguistic review*, 19(1-2):9–50, 2002. 1

[46] 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. 4

[47] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. *arXiv preprint arXiv:2006.04558*, 2020. 6

[48] Leonid I Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992. 5

[49] 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. 2, 3, 4

[50] Linda Smith and Chen Yu. Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3):1558–1568, 2008. 1

[51] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 776–794. Springer, 2020. 3

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3, 4

[53] Luyu Wang and Aaron van den Oord. Multi-format contrastive learning of audio representations. *arXiv preprint arXiv:2103.06508*, 2021. 3

[54] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on

the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR, 2020. 4

[55] Zhibiao Wu and Martha Palmer. Verb semantics and lexical selection. *arXiv preprint cmp-lg/9406033*, 1994. 2

[56] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014. 5

[57] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pages 570–586, 2018. 3

[58] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2921–2929, 2016. 2

[59] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 2, 6

[60] Hao Zhu, Man-Di Luo, Rui Wang, Ai-Hua Zheng, and Ran He. Deep audio-visual learning: A survey. *International Journal of Automation and Computing*, 18:351–376, 2021. 3