AVFormer: Injecting Vision into Frozen Speech Models for Zero-Shot AV-ASR

Paul Hongsuck Seo Arsha Nagrani Cordelia Schmid
Google Research
{phseo, anagrani, cordelias}@google.com

Abstract

Audiovisual automatic speech recognition (AV-ASR) aims to improve the robustness of a speech recognition system by incorporating visual information. Training fully supervised multimodal models for this task from scratch, however is limited by the need for large labelled audiovisual datasets (in each downstream domain of interest). We present AVFormer, a simple method for augmenting audio-only models with visual information, at the same time performing lightweight domain adaptation. We do this by (i) injecting visual embeddings into a frozen ASR model using lightweight trainable adaptors. We show that these can be trained on a small amount of weakly labelled video data with minimum additional training time and parameters. (ii) We also introduce a simple curriculum scheme during training which we show is crucial to enable the model to jointly process audio and visual information effectively; and finally (iii) we show that our model achieves state of the art zero-shot results on three different AV-ASR benchmarks (How2, VisSpeech and Ego4D), while also crucially preserving decent performance on traditional audio-only speech recognition benchmarks (LibriSpeech). Qualitative results show that our model effectively leverages visual information for robust speech recognition.

1. Introduction

Robustness or adaptation to new, unconstrained domains is a key challenge for automatic speech recognition (ASR) systems. In multimodal video (*e.g.*, TV, online edited videos), the visual stream can provide strong cues for improving the robustness of ASR systems, particularly in cases where the audio is noisy – this is called audiovisual ASR (AV-ASR). Unlike works that simply focus on lip motion [1, 7, 23, 24, 29, 33, 37, 41], we investigate the contribution of entire visual frames. This is particularly useful for videos 'in the wild', where the mouth is not necessarily visible (*e.g.*, egocentric viewpoints, face coverings, and low resolution etc.) [11]. The task is illustrated in Figure 1.

Building audiovisual datasets for training AV-ASR models, however, is challenging. Datasets such as How2 [36] and VisSpeech [11] have been created from instructional videos online, but they are small in size. Not only are datasets for this task small, but models are typically large and consist of both visual and audio encoders. For example the latest AV-ASR model AVATAR [11] shows impressive performance on both datasets, but requires the end-to-end training of visual and audio components in tandem, and consequently a large amount of compute. Like other AV-ASR works [4, 17, 25, 30, 38], it is also only trained and tested on instructional videos, and as we show in the experiments, generalizes poorly to new domains in the zero-shot setting.

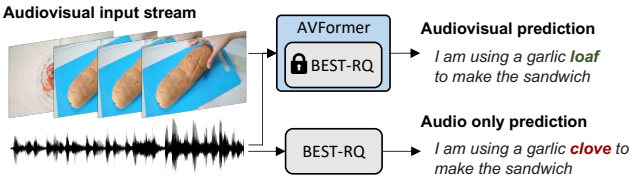

Figure 1. **Unconstrained audiovisual speech recognition.** We inject vision into a frozen speech model (BEST-RQ, in grey) for zero-shot audiovisual ASR via lightweight modules to create a parameter and data efficient model called AVFormer (blue). The visual context can provide helpful clues for robust speech recognition especially when the audio signal is noisy (the visual loaf of bread helps correct the audio-only mistake clove to loaf in the generated transcript).

On the other hand, there have been a number of recently released large-scale audio-only models [6, 8, 19] that are heavily optimised via self-supervised pretraining and large-scale supervised training on *audio-only* data obtained from audio books such as LibriLight [20] and LibriSpeech [31]. These models contain billions of parameters, are readily available, and show strong *generalization across domains*.

Our goal is to reuse the extensive expertise and training time that has been invested in such models, by using their weights. We are inspired by recent works adapting *frozen* foundation models for multi-modal tasks. A popular example is [2] that injects visual information into large language models (LLMs) for vision-text tasks. The benefit of

building on strong frozen LLMs for these tasks is the hope that this will enable the visual-text model to retain powerful *language-only* abilities such as few-shot language adaptation or external knowledge retrieval. Our goal is simple - we wish to do the same for AV-ASR, using strong audio-only ASR models. We add visual inputs to these models in a lightweight manner to enable AV-ASR, but still maintain the benefits of audio-only pretraining for zero-shot generalization.

Our framework is called AVFormer, and injects visual information into a frozen ASR model using lightweight projection layers and trainable adaptors. We show that these can be trained on a small amount of weakly labelled video data (only 5% of the data used by existing state of the art methods [11]) with minimum additional training time and parameters, minimizing the domain shift and catastrophic forgetting that can accompany end-to-end finetuning. In order to further ensure stability during the finetuning of these adapters, we also introduce a simple curriculum scheme during training which we show is crucial to enable the model to jointly process audio and visual information effectively. Finally, we show that our model outperforms existing state of the art zero-shot methods on three different AV-ASR benchmarks (How2, VisSpeech and Ego4D) across different domains, while also crucially preserving decent performance on traditional audio-only speech recognition benchmarks (LibriSpeech).

2. Related Works

State-of-the-Art Speech Recognition Recent state-of-the-art ASR models [6,8,19,45,46] almost all adopt transformer based audio encoders [16, 19, 40] embedding input audio signals into a set of token features thereby extracting local information within a temporal window. Encoders are trained end-to-end using losses such as CTC [15], RNN-T [14] and LAS [5]. In many cases, these encoders are pretrained [6, 8, 19, 45, 46] on large-scale unannotated datasets such as LibriLight [20], and then finetuned for downstream ASR. Consequently, such models incorporate a number of highly-engineered training tricks and techniques suitable for ASR, which we want to reuse for multimodal inference. Rebuilding a multimodal model from scratch incorporating these learnings is expensive and must be redone for each new model. As models get larger and larger [8, 19, 46], this requires a prohibitive amount of compute. Our goal is to reuse this knowledge in a lightweight manner by injecting visual understanding capability into a readily available state-of-the-art ASR model.

Audiovisual Speech Recognition Most AV-ASR works are focused on lip motion, right from early works that use pre-extracted features [29,41] to more recent end-to-end approaches that work on pixels directly [1,7,23,24,33,37]. In contrast, the setting explored in this work is full frame AV-ASR beyond the speaker's mouth movements (also known as 'context-aware' speech recognition). Here the defacto strategy is to use pre-extracted visual context features (due to the high dimensionality of full frame video) – either action features [4, 12, 32, 36], or place and object features [4,17,25,30,38]. Unlike these works which all use visual features from classification models trained on a closed-set of pre-defined objects, places or actions, we use features from CLIP [35], which is trained on image and text paired data, and known to have strong generalization and zero-shot capabilities. This makes our features more suited to unconstrained videos 'in the wild'. An outlier is the recently proposed AVATAR [11], which uses full frame pixels and trains end-to-end on HowTo100. It is the state of the art for this task, achieving good performance on How2 and introducing a new dataset called VisSpeech. Unlike AVATAR, our method reuses strong frozen pretrained models, thereby requiring only 5% of the audiovisual data used in AVATAR, and generalises much better across different domains in the zero-shot setting.

Adapting Large Frozen Pretrained Models There has been a recent flurry of works that adapt frozen foundation models for multi-modal tasks, most notably for injecting visual information to large language models (LLMs) [2]. Architectural details vary: for example MAGMA [10] and Frozen-BiLM [42] add bottleneck adapters [18, 39] to the frozen LLM injecting some visual information; ClipCap [28] learns a vision-to-prefix bridging transformer to map vision features into a prefix for GPT-2, while VC-GPT [22] adds new learnt layers to the frozen LLM. In the AV-ASR domain specifically, multiple works use pre-extracted visual features to improve audio-only ASR [25]. Early work [25] leverages objects and places features from visual classifiers by projecting them to the same space as the audio features in a process known as Visual Adaptive Training (VAT). [17] also uses similar features, but adopts them as the beginning token of each sentence in a language modelling framework. [4] also uses VAT, but for a sequence to sequence model. Unlike these works which use a single visual feature, we show that having multiple visual features improves performance. The closest to our work is LLD [12], which also uses a stream of visual features extracted from the MIL-NCE model [26]. Their fusion method, however, consists of a complicated deliberation decoder, and while they initialize their model with audio-only pretraining, they then finetune the entire audiovisual model end-to-end. In contrast, most of our model remains frozen, and only lightweight adapters are tuned on a small amount of audio-visual data. All previous works are also only focused on the instructional video domain, reporting results either on internally collected datasets or the publicly released How2 [36]. Our focus instead is on zero-shot generalisation across multiple domains, including audio-only

Librispeech [31] (from audiobooks) and Ego4D [13] (egocentric video). We believe this is a more useful setting for actual deployment of such models.

3. Method

Unlike previous AV-ASR works which test only on instructional videos [4, 17, 25, 30, 38], our goal is *zero-shot* generalization across multiple AV domains, while still maintaining good performance on traditional audio-only benchmarks. To do this, we start with an *existing* state-of-the-art ASR model, and adapt it for unconstrained AV-ASR. Visual features are obtained from a strong pretrained visual model, and added to the model via the following two components - (i) we linearly project visual features into the audio token embedding space, and (ii) we inject lightweight adapters into the encoder of the frozen ASR model to allow domain adaptation. During training, we only tune these two sets of additional parameters, while both the ASR model and the visual feature extractor are *frozen* (see Figure 2).

We do this because there are two forms of adaption that are required here - (i) adapting to new video domains and (ii) adapting to multimodal input, both of which we would like to do *without* catastrophic forgetting. Because of the challenges with this setup, we also introduce a curriculum learning strategy to stabilize the learning process, without which the model fails to utilize the visual features effectively. In this section, we first describe the main components of our network architecture (Sec. 3.1) and then introduce our zero-shot curriculum learning strategy and training loss functions (Sec. 3.2).

3.1. Model Architecture

In this section we describe the key components of our architecture - (i) the frozen conformer encoder and decoder, (ii) the visual encoder and projection layers for visual feature extraction and projection, and (iii) additional adaptation layers in the backbone for audio-only domain adaptation. A diagram is show in Figure 2.

3.1.1 Frozen Conformer ASR Model

We start with a frozen ASR model that achieves state-of-the-art performance on traditional ASR benchmarks [31]. Specifically, we use BEST-RQ [6] that adopts a Conformer [16] model with an RNN-Transducer (RNN-T) [14]. The model is pretrained on LibriLight [20] in a self-supervised manner using a random projection quantization technique, after which it is then finetuned for ASR on LibriSpeech [31] using supervised training. The conformer consists of convolution-augmented transformer blocks (conformer blocks), which operate on audio token features that are extracted from a spectrogram via a stack of convolution and linear layers [16]. BEST-RQ uses ConformerXL as a backbone, which has 0.6B parameters [46] – note that training such a large model end-to-end is extremely compute heavy – and requires a large pretraining dataset (made possible by self-supervised learning on Libri-Light). This self-supervised training also enables the model to generalize well across numerous domains. After pretraining, an RNN-T decoder is added to Conformer to generate text output for ASR with 1,024 WordPiece tokens [44]. The RNN-T decoder generates a sequence of tokens consisting of grapheme tokens or a special output token, which represents moving to the next input token (See Figure 2, right for a diagram of the decoder).

Formally speaking, given the log-mel spectrogram $\mathbf{X} \in \mathbb{R}^{\hat{N} \times S}$ with S mel spectrogram bins in a length of $\hat{N}$ converted from the input audio waveform, the tokenizer outputs a set of audio tokens $\{\mathbf{t}_i\}_1^N = h_{\text{tok}}(\mathbf{X})$ where D is the token embedding dimensionality and $N = \hat{N}/4$. The encoder then contextualizes the audio tokens through a series of conformer blocks, each of which is a stack of feed-forward, multi-head self-attention, convolution layers followed by another feed-forward layer. The output of each layer is added with a residual connection. This process produces N contextualized tokens $\hat{\mathbf{t}}_i \in \mathbb{R}^D$, *i.e.*, $\{\hat{\mathbf{t}}_i\}_1^N = h_{\text{enc}}(\{\mathbf{t}_i\}_1^N)$. The decoder finally generates the transcripts by predicting a sequence of K graphemes from the contextualized audio tokens. Given a token $\hat{\mathbf{t}}_i$ and previously generated grapheme w_{j-1}, the decoder generates the next grapheme $w_j = h_{\text{dec}}(\hat{\mathbf{t}}_i, w_{j-1})$ where $w_j \in \mathcal{V} \cup \{\epsilon\}$ with the vocabulary of the predefined graphemes $\mathcal{V}$ and a special blank token ϵ that represents moving to the next token $\hat{\mathbf{t}}_{i+1}$ in the generation process. The decoder h_{dec} is implemented as a two layer LSTM module with a grapheme classification head. Note that at a single audio token index i, multiple graphemes can be emitted (vertical arrows) until an ϵ is emitted (horizontal arrows) as depicted in Figure 2.

3.1.2 Lightweight Adapters

In order to enable domain adaption in the model, we interleave an adapter layer within each conformer block of the encoder. Note that the BEST-RQ model has strong generalization capability, which we want to maintain. Hence we design our adapters to be lightweight, to prevent drastic domain shift and catastrophic forgetting. Given N audio tokens $\mathbf{t}_i$ and M projected visual tokens $\mathbf{t}_j^v$ (which will be described next) at a certain layer l[1], we compute the adapted token features $\tilde{\mathbf{t}}_i$ and $\tilde{\mathbf{t}}_j^v$ using an adapter layer by $\{\tilde{\mathbf{t}}_i\} \cup \{\tilde{\mathbf{t}}_j^v\} = \text{adapt}(\{\mathbf{t}_i\} \cup \{\mathbf{t}_j^v\}; \phi)$ where $\text{adapt}(\cdot)$ is an adapter layer parameterized by ϕ. We introduce and experiment with the following two types of lightweight adapters:

[1]The layer index l is omitted for notational simplicity.

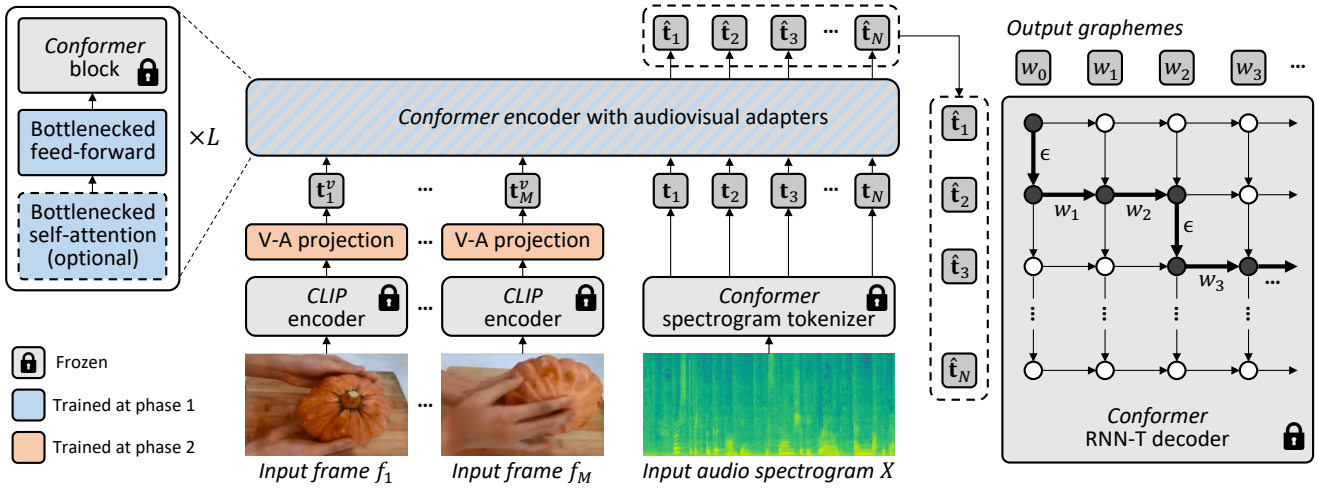

Figure 2. **Overall architecture and training procedure for AVFormer.** Our architecture consists of a frozen Conformer encoder-decoder model [6], and a frozen CLIP [35] encoder (frozen layers shown in grey with a lock symbol), in conjunction with two lightweight trainable modules - (i) visual projection layer (orange) and bottleneck adapters (blue) to enable multi-modal domain adaptation. We propose a two-phase curriculum learning strategy - the adapters (blue) are first trained without any visual tokens, after which the visual projection layer (orange) are tuned while all the other parts are kept frozen.

Feed-forward Adapters (FF). The simplest design is to independently project each token. To achieve this, we use a two-layered MLP with a residual connection as our adapter. To make the layer lightweight, we set the dimensionality of the hidden layer to B, where $B \ll D$. This allows the adaptor to effectively act as a bottleneck, and reduces total additional parameters.

Feed-foward Adapters with Self-Attention (FF+SA). The feed-forward adapters described above operate independently for each token. We can perform an additional contextualization across the input tokens via a self-attention layer [43]. To reduce additional parameters, we apply the same bottleneck projection technique as before, where each input token is transformed into a B dimensional query, key and value for attention, after which the attended feature is projected back into the D dimensional feature space. For multi-head self-attention, each head projects features into B/H dimensional spaces instead where H stands for the number of heads. This module is used with a residual connection and a feed-forward module described above; the combination of these forms a transformer block with bottlenecks. While this (FF+SA) allows additional contextualization across tokens, it introduces four times more parameters than vanilla FF adapters.

3.1.3 Visual Feature Extraction and Projection

Given a sequence of M video frames $\mathbf{f}_i$, we extract a $\hat{D}$ dimensional visual feature $\mathbf{v}_i = g(\mathbf{f}_i)$ per frame using a pre-trained visual encoder g. Specifically, we use the CLIP encoder [34] with ViT-L/14 [9] as our visual backbone, which is known to have strong zero-shot generalization capability [34]. Because the CLIP encoder is frozen, we add a linear layer[2] to project the visual features into the audio token embedding space, *i.e.*, $\mathbf{t}_i^v = \text{proj}(\mathbf{v}_i; \theta)$ where $\mathbf{t}_i^v \in \mathbb{R}^D$ and θ is a set of the parameters in the projection layer. The projected visual tokens are fed to the Conformer encoder together with audio tokens $\mathbf{t}_i$. Note that these visual projection layers are essentially performing a type of prompt tuning [21, 28] since the rest of the ASR model is frozen.

3.2. Training Strategy

It is a well-known that AV-ASR is an audio-dominant task, which is why previous works are forced to devise training strategies that prevent the audio stream from dominating training [11]. We observe a similar phenomenon while jointly training both sets of additional parameters (adapters and visual projections). The visual information is not used (similar performance with and without), and training is dominated by the model only adapting to the finetuning *audio* domain. We hence introduce a curriculum training strategy. We first describe our finetuning data, the loss function, and then the curriculum in the next few paragraphs.

Zero-shot Training with Web Videos. Our extended model has two sets of new parameters θ and ϕ introduced for the visual projection layer and the adapters respectively. Since it is labor-intensive and costly to collect new training benchmarks for AV-ASR, we train these new parameters without manually labeled data. We use unlabeled web videos online along with the outputs of an ASR model as

[2]We tested more complex MLP projectors and found that a single linear layer is sufficient for good performance as detailed in the supp. mat.

pseudo ground truth. Our goal is to aid the pretrained ASR model with visual understanding capability using only these automatically collected transcripts; the trained model is then tested in a zero-shot setting on manually annotated public AV-ASR benchmarks.

Loss Function. As the RNN-T decoder in the pretrained ASR model is kept frozen in AVFormer, we adopt the same loss function that is used for ASR pretraining. With an RNN-T decoder, the probability of a transcript $W = \{w_1, w_2, \cdots, w_K\}$ is obtained by marginalizing the probabilities of all valid generation paths y (*e.g.*, the path with bold arrows in Figure 2), *i.e.*,

$$P(W|X) = \sum_{y \in \mathcal{Y}} \prod_{(i,j) \in y} P(w_j | \hat{\mathbf{t}}_i, w_{0:j-1}) \tag{1}$$

where $\mathcal{Y}$ is a set of all valid paths y (paths on the grid from $(0,0)$ to $(N+1, K)$ in Figure 2) which is a sequence of pairs of token and output grapheme indices (i,j), and $P(w_j|\hat{\mathbf{t}}_i, w_{0:j-1})$ is estimated by our decoder $h_{\text{dec}}(\hat{\mathbf{t}}_i, w_{j-1})$. We train our model by minimizing the negative log-likelihood of the pseudo-GT transcripts $\hat{W}$ of input videos:

$$\mathcal{L}(\theta, \phi) = -\sum_i \log P(\hat{W}_i | X_i; \theta, \phi). \tag{2}$$

Curriculum Learning for Visual Processing. We discover empirically that with a naive first round of joint training, our model struggles to learn both the adapters and the visual projectors in one go (as shown in the experiments, the issue becomes more severe as more visual tokens are added). To mitigate this issue, we propose a two-phase curriculum learning strategy that decouples these two factors (domain adaption and visual feature integration) and trains the network in a sequential manner. In the first phase, the adapter parameters ϕ are optimized using $\text{argmax}_\phi \mathcal{L}(\theta, \phi)$ as an objective. Note that at this phase, we do not feed visual tokens at all and thus θ is an empty set. Once ϕ is trained, we add the visual tokens and train the visual projection layers θ using $\text{argmax}_\theta \mathcal{L}(\theta, \phi)$. During this second phase of training, ϕ is kept frozen.

The first stage focuses on audio domain adaptation. By the second phase, the adapters are completely frozen and the visual projector must simply learn to generate visual prompts that project the visual tokens into the audio space. In this way, our curriculum learning strategy allows the model to incorporate visual inputs as well as adapt to new audio domains in AV-ASR benchmarks. We apply each phase just once, as an iterative application of alternating phases leads to performance degradation. This is further discussed in the supp. mat.

Content Word Masking. We adopt the content word masking from [11] to encourage the models to further focus on visual understanding. We observe that the original zero-padded masking introduced in [11] causes instabilities and therefore we add Gaussian noise to the audio input corresponding to masked words, which stabilizes optimization.

4. Experiments

4.1. Experimental Settings

Implementation Details. As mentioned earlier, we use BEST-RQ [6] as the frozen ASR model. Since it has 24 conformer blocks, we add 24 adapters (one in each layer) in all experiments. When added, all adapters and visual projectors are randomly initialized. The decoder predicts WordPiece tokenized graphemes with a vocabulary size of 1,024. In the adapters, we apply layer norm [3] at every residual connection. For both phases of training, we use standard SGD with momentum with a moving average coefficient of 0.9 and a cosine learning rate schedule; the initial learning rate is set to 0.4. We train for 40K and 30K iterations in phase 1 and 2 respectively, with a batch size of 256 on 32 TPU v4 chips. We run 5 independent experiments and report the mean scores for ablation studies. When testing Audiovisual models on audio-only benchmarks, we feed dummy visual inputs (zero tensors).

Metrics. We use word error rate (WER) for all evaluation (lower is better). The alignment between predicted words and ground truth is computed using dynamic programming. The WER is then computed by the number of errors (deletions, substitutions and insertions) across the whole test set divided by the number of ground truth words.

Baselines. We compare AVFormer to two strong baselines proposed this year - (i) the state-of-the art AV-ASR model AVATAR [11] and (ii) the state-of-the-art ASR (audio only) model BEST-RQ [6]. We apply both models to the same settings as AVFormer for a fair comparison.

4.2. Datasets

The additional parameters in our model are finetuned on the HowTo100M dataset, which contains instructional videos from YouTube. In order to assess generalization, we evaluate across different domains – LibriSpeech (audiobooks), How2 and VisSpeech (YouTube instructional videos) and Ego4D (egocentric video of daily-life activities). Note that VisSpeech consists of more unconstrained video (background noise, challenging accents etc) than How2. More details for each dataset are provided below.

LibriLight [20] and LibriSpeech [31]. LibriLight is an unlabelled speech dataset that is used to pretrain BEST-RQ. The model is then finetuned for ASR on LibriSpeech containing 960 hours audio with manually annotated GT transcripts. For a fair comparison, we also use LibriSpeech for pretraining some of our baselines in the ablations.

HowTo100M [27]. This dataset contains 1.2M instructional

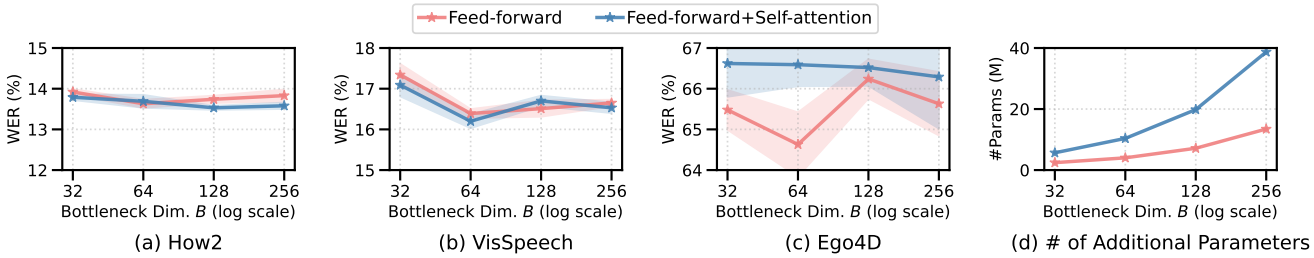

Figure 3. **Effects of different architectures (feed-forward (FF) vs feed-forward + self-attention (FF+SA)) and the bottleneck dimensionality B of adaptor layers on performance.** Results are for audiovisual models trained with our curriculum learning, and are shown on 3 datasets in the zero-shot setting (lower WER% is better). We show that a bottleneck dimension of 64 with FF layers achieves the best or almost the best performance (a,b,c) with the least number of additional parameters (d). Best viewed in color.

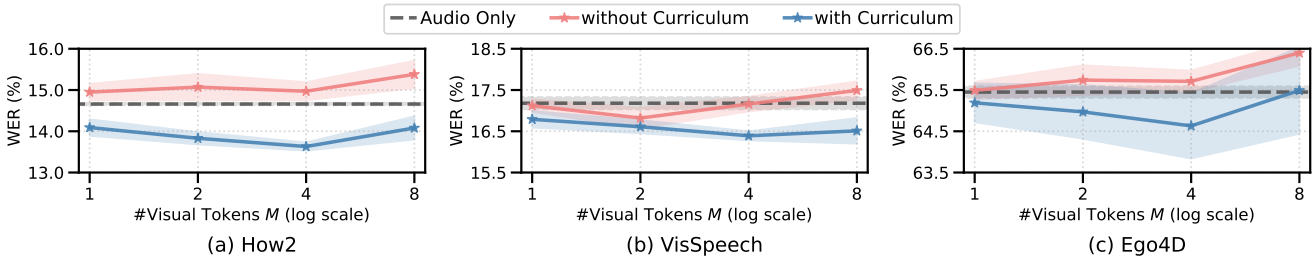

Figure 4. **Effects of curriculum learning and the number of visual tokens M on performance.** Red and blue lines are for audiovisual models and are shown on 3 datasets in the zero-shot setting (lower WER% is better). Using the curriculum helps on all 3 datasets (for How2 (a) and Ego4D (c) it is crucial for outperforming audio-only performance). Performance improves up until 4 visual tokens, at which point it saturates. Best viewed in color.

videos without manual annotations. ASR is used to obtain pseudo-GT transcripts for training our adapters and visual projector. We remove videos present in VisSpeech and How2 (described next).

How2 [36]. We use the 300hr version of How2, which consists of instructional videos with automatically collected user uploaded captions. The videos are segmented into 5.8s short clips with 20-word long transcripts in average. We use the validation (2,022 clips) and test (2,305 clips) splits to evaluate our model in a zero-shot setting.

VisSpeech [11]. VisSpeech is an AV-ASR test benchmark that consists of 503 video clips with manually annotated transcripts, which are sampled from HowTo100M. The dataset curation process focuses on samples where an audio-only ASR model fails and where strong visual correlations are observed.

Ego4D [13]. Ego4D consists of egocentric video from 74 worldwide locations and 9 countries, with over 3,670 hours of daily-life activity video. We use the audiovisual diarization benchmark in the Ego4D challenge[3]. It consists of 585 5-minutes long egocentric video clips splited into train (397 clips), validation (51 clips) and test (137 clips) sets. We report zero-shot results on the validation set as the test annotations are not released. We evaluate transcripts on segmented clips based on GT boundaries.

[3] https://ego4d-data.org/docs/challenge/

4.3. Results

In this section, we show ablations of the various design choices in our model (adapter architecture and bottleneck dimension), and then discuss the impact of curriculum learning and the benefit of adding visual tokens (including the impact of the number of visual tokens). We then show an ablation discussing the impact of adding both adapters and visual tokens, and the impact of finetuning dataset size. Finally, we show zero-shot performance of our model compared to state of the art baselines. Note that all ablations and results are provided on all 3 downstream datasets in a zero-shot setting – How2, VisSpeech and Ego4D.

Adapter Architecture and Bottleneck Dimensionality. Figure 3 compares results with feed-forward adapters (FF) only vs adapters with both feed-forward and self-attention (FF+SA). We also vary the bottleneck dimension from 32 to 256. We observe that on How2 (Figure 3a) and VisSpeech (Figure 3b), both adapter types perform similarly although FF+SA uses significantly more parameters than FF (Figure 3d), indicating that a simple projection is enough for strong adaptation. On Ego4D (Figure 3c), simple FF outperforms FF+SA by a large margin, potentially because of the larger domain gap (instructional edited videos online to egocentric daily activity videos). The greater number of parameters in FF+SA may result in a larger shift to the instructional video domain and away from Ego4D. Figure 3

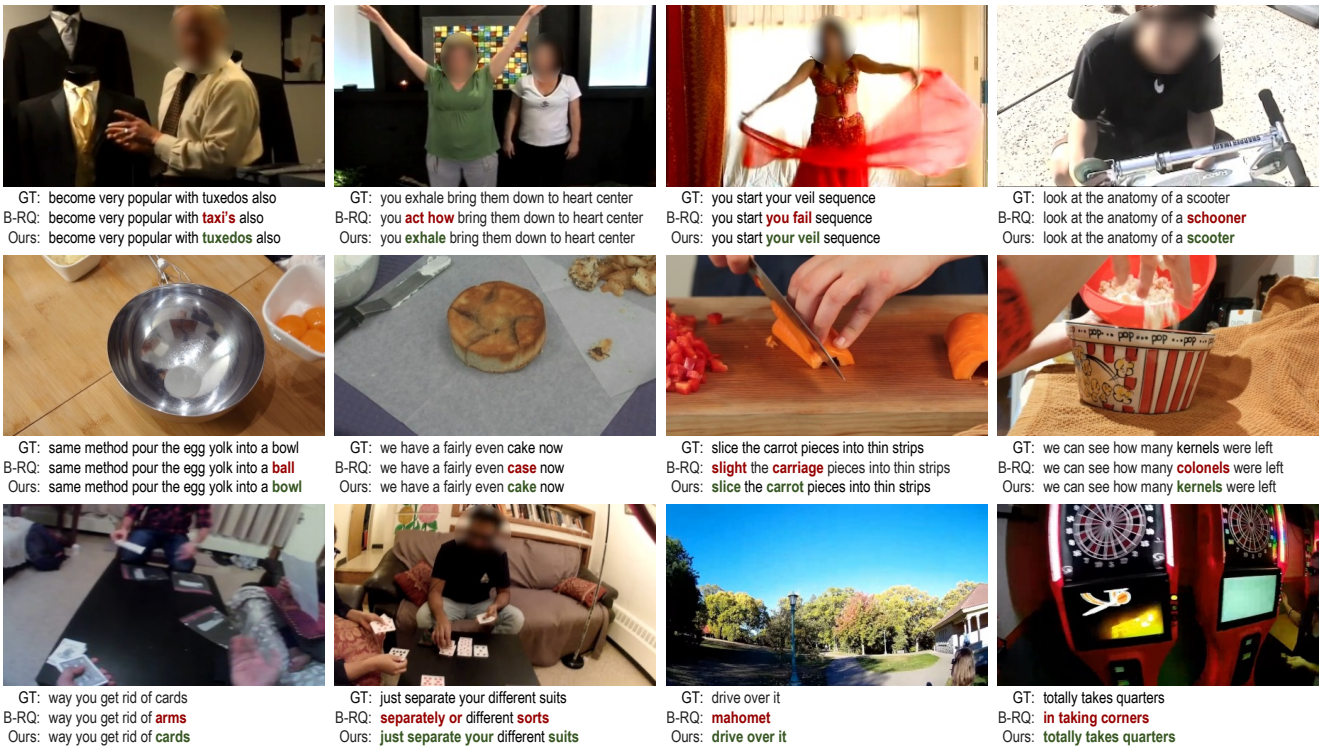

Figure 5. **Qualitative Results on How2 (top), VisSpeech (middle) and Ego4D (bottom).** We show the ground truth (GT), and predictions from the audio only BEST-RQ model (B-RQ) and our audiovisual AVFormer (Ours) in the zero-shot setting. For each clip we show a single visual frame. Note how the visual context helps with visual objects (tuxedos, veil, scooter, bowl, cake, carrot *etc*), as well as actions (exhale, drive over) and works well even in the ego-centric domain (learns driving from input of road in row 3, column 3). Errors in the predicted words compared to the GT are highlighted in red. Faces are blurred for privacy.

Table 1. **Effect of visual tokens (VT) and adapter layers.** Results on 3 datasets are obtained in the zero-shot setting (lower WER% is better). The first row corresponds to the vanilla pre-trained BEST-RQ. Visual projector is added only when feeding VT. The gains from both VT and adapters are complementary.

VT	Adapters	How2	VisSpeech	Ego4D
		21.90	31.61	77.98
✓		19.74 ± 0.04	31.13 ± 0.06	76.50 ± 0.11
	✓	14.66 ± 0.03	17.18 ± 0.15	65.45 ± 0.14
✓	✓	13.63 ± 0.10	16.39 ± 0.11	64.63 ± 0.79

also shows the effect of different bottleneck dimensions. In general the WER comes down from 32 to 64, but saturates at $B = 64$ across all datasets with FF, while introducing only few additional parameters (0.6% of the number of parameters in BEST-RQ). Hence in the rest of experiments, we adopt FF adapters with $B = 64$.

Curriculum Learning and Visual Tokens. We show the results of AVFormer with and without the proposed two-stage curriculum in Figure 4, and also compare to an audio-only baseline which had only FF adapters with $B = 64$ and no visual information. Without curriculum learning, our AV-ASR model is worse than the audio-only baseline across all datasets, with the gap increasing as more visual tokens are added. In contrast, when the proposed two-phase curriculum is applied, our AV-ASR model performs significantly better than the baseline audio-only model. We also test our model with different number of visual input tokens (where one token corresponds to one frame). More visual tokens improves the model up until $M = 4$ with up to 7.0% relative improvement, after which performance begins to degrade. Hence we set $M = 4$ in all experiments.

Table 2. **Effect of training dataset size.** Results are for audiovisual models trained with our curriculum learning, and are shown on 3 datasets in the zero-shot setting (lower WER% is better). Only 5% of HowTo100M is required.

Training-set size	How2	VisSpeech	Ego4D
5%	13.69 ± 0.17	16.60 ± 0.17	64.75 ± 1.05
100%	13.63 ± 0.10	16.39 ± 0.11	64.63 ± 0.79

Complementary Gain of Additional Components. Table 1 shows the effect of our additional lightweight components (projection layer for visual tokens and adapter layers) for zero-shot AV-ASR. The first row is simply the vanilla baseline (frozen BEST-RQ). We observe that adding projected visual tokens and adapters bring individual gains to the baseline (the former adding visual information and the latter aiding with audio-domain adaptation), and when com-

Table 3. **Comparison to state-of-the-art methods for zero-shot performance across different AV-ASR datasets.** We also show performance on LibriSpeech which is audio-only. Results are reported as WER % (lower is better). Note that AVATAR and BEST-RQ are finetuned end-to-end (all parameters) on HowTo100M, whereas for our model, only the visual projectors (VP) and adapters are finetuned on 5% of the dataset. PT means pretraining. When a model is marked with both LibriSpeech and HowTo100M pretraining, we first train the model on LibriSpeech and then on HowTo100M next. For LibriSpeech evaluation, we report numbers on test-clean set. *LibriSpeech trained model is evaluated directly on LibriSpeech test set.

Method	Modality	LibriSpeech PT	HowTo100M PT		LibriSpeech	How2	VisSpeech	Ego4D
			Pretrained params	Data %				
AVATAR [11]	A	✓	–	–	8.85	39.43	65.33	110.86
AVATAR [11]	A+V	–	All	100	24.65	17.23	35.66	92.03
AVATAR [11]	A+V	✓	All	100	24.08	18.37	35.59	71.97
BEST-RQ [6]	A	✓	–	–	1.60*	21.90	28.62	77.98
BEST-RQ [6]	A	✓	All	100	5.60	15.32	16.69	68.34
AVFormer (Ours)	**A+V**	✓	VP + Adapters	5	4.36	**13.69**	**16.60**	**64.75**

bined with our curriculum learning, are complementary to performance, achieving the lowest WER.

Training Dataset Size. Given our additional components are so lightweight, we test whether adaptation can be done with a small amount of weakly labelled data. The results in Table 2 show only 5% of HowTo100M training data performs on par with the full dataset – *i.e*. the pretrained knowledge in BEST-RQ and CLIP yields considerable data efficiency to the model. Ablation results with more data fractions are provided in the supp. mat.

Comparisons to Zero-shot Baselines on AV-ASR. We compare our model to baselines in Table 3, for zero-shot performance on all 3 AV-ASR benchmarks[4] AVFormer outperforms AVATAR and BEST-RQ on all, even outperforming both AVATAR and BEST-RQ when they are fully finetuned on LibriSpeech and then 100% of HowTo100M (3rd and 5th row). Note for BEST-RQ, this involves finetuning 0.6B params. Our model, in contrast only finetunes 4M params on 5% of HowTo100M.

Comparisons to Zero-shot Baselines on LibriSpeech. Even though this is not the main goal of this work, we also investigate performance on LibriSpeech, which is audio-only (Table 3). Note other AV-ASR works do not do this, but we believe it is important for deployment of AV-ASR models. We first note that AVATAR pretrained on LibriSpeech and then finetuned on HowTo100M performs poorly when re-evaluated on LibriSpeech (showing severe catastrophic forgetting between rows 1 and 3). We believe this is because all parameters are trained end-to-end. On the other hand, AVFormer performs much better on LibriSpeech (4.36 vs 24.08), and is much closer to BEST-RQ's 1.60 which is a model tuned only for LibriSpeech and incapable of AV-ASR, while AVFormer achieves SOTA on AV-ASR as well.

Qualitative Results. Qualitative examples are provided in Fig. 5 comparing our method to audio-only BEST-RQ for zero-shot ASR. We show that for all 3 downstream AV-ASR datasets, visual context improves mistakes that are made on objects (*eg.* tuxedos, veil and scooter from the top row), actions (exhale - top row, second column), and even corrects a homophone [5] (colonels to kernals, row 2 column 4).

[4]Note that the original AVATAR and BEST-RQ papers do not report this. We apply these models in the same setting as ours for a fair comparison.

Table 4. **Finetuning performance on How2 and Ego4D.** We outperform all previous works on How2 that use frozen visual features. AVATAR is trained end-to-end, with all visual parameters finetuned. Scores are in WER %.

Method	Frozen visual feats	How2	Ego4D
VAT [4]	✓	18.0	–
MultiRes [32]	✓	20.5	–
LLD [12]	✓	16.7	–
AVATAR [11]		9.11	55.27
AVFormer (Ours)	✓	10.22	**55.23**

Comparisons to SOTA after Finetuning. For completeness, we also show finetuning results on two domains - instructional (How2) and egocentric (Ego4D) videos in Table 4. We outperform all previous works on How2 that use frozen visual features. Our model is also not too much worse (How2) or on par (Ego4D) with AVATAR, even though AVATAR is trained end-to-end, and all parameters (including a large visual encoder) are finetuned.

5. Conclusion

We present AVFormer, a lightweight method for adapting existing, frozen state-of-the-art ASR models for AV-ASR. Our approach is practical and achieves impressive zero-shot performance. As ASR models get larger and larger, tuning the entire parameter set of pre-trained models becomes impractical for different domains. Our method seamlessly allows both domain transfer and visual input mixing in the same, parameter efficient model.

[5]same pronunciation, different spelling

References

[1] Triantafyllos Afouras, Joon Son Chung, Andrew W. Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE transactions on pattern analysis and machine intelligence*, 2018. 1, 2

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 1, 2

[3] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 5

[4] Ozan Caglayan, Ramon Sanabria, Shruti Palaskar, Loïc Barrault, and Florian Metze. Multimodal grounding for sequence-to-sequence speech recognition. In *ICASSP*, 2019. 1, 2, 3, 8

[5] William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *ICASSP*, 2016. 2

[6] Chung-Cheng Chiu, James Qin, Yu Zhang, Jiahui Yu, and Yonghui Wu. Self-supervised learning with random-projection quantizer for speech recognition. In *ICML*, 2022. 1, 2, 3, 4, 5, 8

[7] Joon Son Chung, Andrew W. Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. *CVPR*, 2017. 1, 2

[8] Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 244–250. IEEE, 2021. 1, 2

[9] 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. In *ICLR*, 2020. 4

[10] Constantin Eichenberg, Sidney Black, Samuel Weinbach, Letitia Parcalabescu, and Anette Frank. Magma–multimodal augmentation of generative models through adapter-based finetuning. *arXiv preprint arXiv:2112.05253*, 2021. 2

[11] Valentin Gabeur, Paul Hongsuck Seo, Arsha Nagrani, Chen Sun, Karteek Alahari, and Cordelia Schmid. Avatar: Unconstrained audiovisual speech recognition. In *Interspeech*, 2022. 1, 2, 4, 5, 6, 8

[12] Shahram Ghorbani, Yashesh Gaur, Yu Shi, and Jinyu Li. Listen, look and deliberate: Visual context-aware speech recognition using pre-trained text-video representations. In *IEEE Spoken Language Technology Workshop (SLT)*, 2021. 2, 8

[13] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, 2022. 3, 6

[14] Alex Graves. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*, 2012. 2, 3

[15] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In *ICML*, 2006. 2

[16] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech*, 2020. 2, 3

[17] Abhinav Gupta, Yajie Miao, Leonardo Neves, and Florian Metze. Visual features for context-aware speech recognition. In *ICASSP*, 2017. 1, 2, 3

[18] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019. 2

[19] 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. 1, 2

[20] Jacob Kahn, Morgane Rivière, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, et al. Libri-light: A benchmark for asr with limited or no supervision. In *ICASSP*, 2020. 1, 2, 3, 5

[21] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *EMNLP*, 2021. 4

[22] Ziyang Luo, Yadong Xi, Rongsheng Zhang, and Jing Ma. Vc-gpt: Visual conditioned gpt for end-to-end generative vision-and-language pre-training. *arXiv preprint arXiv:2201.12723*, 2022. 2

[23] Pingchuan Ma, Stavros Petridis, and Maja Pantic. End-to-end audio-visual speech recognition with conformers. *ICASSP*, 2021. 1, 2

[24] Takaki Makino, Hank Liao, Yannis M. Assael, Brendan Shillingford, Basi García, Otavio Braga, and Olivier Siohan. Recurrent neural network transducer for audio-visual speech recognition. *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019. 1, 2

[25] Yajie Miao and Florian Metze. Open-domain audio-visual speech recognition: A deep learning approach. In *INTERSPEECH*, 2016. 1, 2, 3

[26] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020. 2

[27] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a Text-Video Embedding by

Watching Hundred Million Narrated Video Clips. In *ICCV*, 2019. 5

[28] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 2, 4

[29] Kuniaki Noda, Yuki Yamaguchi, Kazuhiro Nakadai, H. Okuno, and Tetsuya Ogata. Audio-visual speech recognition using deep learning. *Applied Intelligence*, 42:722–737, 2014. 1, 2

[30] Shruti Palaskar, Ramon Sanabria, and Florian Metze. End-to-end multimodal speech recognition. *ICASSP*, 2018. 1, 2, 3

[31] 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. 1, 3, 5

[32] Georgios Paraskevopoulos, Srinivas Parthasarathy, Aparna Khare, and Shiva Sundaram. Multimodal and multiresolution speech recognition with transformers. In *ACL*, 2020. 2, 8

[33] Stavros Petridis, Themos Stafylakis, Pingchuan Ma, Feipeng Cai, Georgios Tzimiropoulos, and Maja Pantic. End-to-end audiovisual speech recognition. *ICASSP*, 2018. 1, 2

[34] 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. 4

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 4

[36] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. How2: a large-scale dataset for multimodal language understanding. In *Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL)*, 2018. 1, 2, 6

[37] Dmitriy Serdyuk, Otavio Braga, and Olivier Siohan. Audio-visual speech recognition is worth 32x32x8 voxels. *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021. 1, 2

[38] Tejas Srinivasan, Ramon Sanabria, and Florian Metze. Looking enhances listening: Recovering missing speech using images. *ICASSP*, 2020. 1, 2, 3

[39] Yi-Lin Sung, Jaemin Cho, and Mohit Bansal. Vl-adapter: Parameter-efficient transfer learning for vision-and-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5227–5237, 2022. 2

[40] Gabriel Synnaeve, Qiantong Xu, Jacob Kahn, Tatiana Likhomanenko, Edouard Grave, Vineel Pratap, Anuroop Sriram, Vitaliy Liptchinsky, and Ronan Collobert. End-to-end asr: from supervised to semi-supervised learning with modern architectures. *arXiv preprint arXiv:1911.08460*, 2019. 2

[41] Satoshi Tamura, Hiroshi Ninomiya, Norihide Kitaoka, Shin Osuga, Yurie Iribe, K. Takeda, and Satoru Hayamizu. Audio-visual speech recognition using deep bottleneck features and high-performance lipreading. *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*, 2015. 1, 2

[42] Yang *et. al.* Zero-shot video question answering via frozen bidirectional language models. In *NeurIPS*, 2022. 2

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 4

[44] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016. 3

[45] Qiantong Xu, Alexei Baevski, Tatiana Likhomanenko, Paden Tomasello, Alexis Conneau, Ronan Collobert, Gabriel Synnaeve, and Michael Auli. Self-training and pre-training are complementary for speech recognition. In *ICASSP*, 2021. 2

[46] Yu Zhang, James Qin, Daniel S Park, Wei Han, Chung-Cheng Chiu, Ruoming Pang, Quoc V Le, and Yonghui Wu. Pushing the limits of semi-supervised learning for automatic speech recognition. *arXiv preprint arXiv:2010.10504*, 2020. 2, 3