Scaling Audio-Visual Speech Generation enables Versatile Transfer

Cheng-I Jeff Lai [1 2] **W. Ronny Huang** [1 *] **Tongzhou Chen** [1 *] **Wang Xiao** [1 *] **Xiaohua Zhai** [1 *]
Paul K. Rubenstein [1] **Kyle Kanster** [1] **Scott Wisdom** [1] **James Glass** [2] **Mingqiu Wang** [1 *]

Abstract

Advances in audio tokenization have propelled unit-based Spoken Language Models (SLMs) like VALL-E and AudioLM, while integrating large language model (LLM) backbones has further advanced speech generation. Building on these developments, we investigate *visual-conditioned speech generation*—whether SLMs can seamlessly integrate visual inputs such as images and short videos to generate coherent speech. We introduce **Audio-Visual Gemma (AV-Gemma)**, a family of 3B/10B audio–visual foundation models that generate audio tokens directly from visual inputs. Distinct from prior speech LLMs, AV-Gemma (1) employs a SigLIP encoder for general-purpose visual understanding, and (2) integrates audio tokens generation within the Gemma LLM backbone. Pretrained on 285k hours of web-scale spoken captions, AV-Gemma demonstrates scalability in both data and model size, enabling direct visual-to-speech generation. We validate its versatile transfer on tasks such as Spoken Visual QA, Video-to-Speech Generation, and Audio–Visual Sound Understanding.

1. Introduction

Large Language Models (LLM) (Brown, 2020; Rae et al., 2021; Chowdhery et al., 2022; Anil et al., 2023; Touvron et al., 2023) excels at natural language tasks, and their superior generative capabilities have transformative impact beyond text-only tasks (Achiam et al., 2023; Gemini et al., 2023; Dubey et al., 2024). In vision, LLMs are augmented with additional visual encoders (Radford et al., 2021; Dosovitskiy, 2020; Zhai et al., 2023) to enable general visual understanding and generation, which are commonly referred to as Visual Language Models (VLM), such as LLaVA (Liu et al., 2023; 2024). VLMs combine the strong reasoning capabilities from LLMs and visual representations from the vision encoder (e.g. CLIP), and recently there is a surge of interest in studying and designing the underlying mechanisms of VLMs (Beyer et al., 2024; McKinzie et al., 2024; Lin et al., 2024; Dai et al., 2024; Tong et al., 2024).

Similarly, in speech processing, there has been significant progress in Spoken Language Models (SLMs) (Wang et al., 2023b), which connect audio representations from pretrained encoders such as Whisper (Radford et al., 2023) and USM (Zhang et al., 2023b) to LLMs. In this work, we focus on unit-based SLMs for speech generation (Wu et al., 2024), where an autoregressive decoder model is trained directly on audio tokens. Representative examples include GSLM (Lakhotia et al., 2021), AudioLM (Borsos et al., 2023a), and VALL-E (Wang et al., 2023a). AudioPaLM (Rubenstein et al., 2023) and SpiRit-LM (Nguyen et al., 2024) highlight the advantages of initializing decoder models with pretrained LLMs for joint speech-text modeling tasks, such as speech-to-speech translation and speech-text continuation. Most recently, this framework has enabled real-time spoken dialogue systems like Moshi (Défossez et al., 2024). Collectively, these works demonstrate the success of generative speech modeling in achieving plausible phonetic, linguistic, paralinguistic, and speaker coherence.

Modern AI assistants, such as Project Astra (DeepMind, 2025), are inherently multimodal, designed to enable real-time multi-sensory understanding and spoken responses. This motivated us to explore the synergy between visual perception and unit-based SLMs, leading to our pursuit of *visual-conditioned speech generation*. In the context of SLM, we task our models with generating coherent acoustic and linguistic units of spoken language that are visually grounded. Conversely, from the VLM perspective, our work trains a VLM with audio understanding and speaking capabilities, while preserving VLM's visual-semantic reasoning. Ultimately, this line of inquiry bridges the gap between visual and auditory perception in LLMs, opening new opportunities and practical applications, such as reducing multimodal generation latency, expanding support for non-written spoken languages, and fostering inclusiveness for visually impaired users.

We introduce Audio-Visual Gemma (AV-Gemma), a family of 3B and 10B foundation models designed for visual-conditioned speech generation, capable of speaking what it

[*]Equal advising [1]Google Deepmind [2]MIT. Correspondence to: Cheng-I Jeff Lai <clai24@mit.edu>.

Preprint. Work in progress.

Model	Input	Output	LLM Backbone	Tasks
AudioLM (Borsos et al., 2023a)	speech, sound	speech, sound	x	audio continuation
AudioPaLM (Rubenstein et al., 2023)	speech, text	speech, text	PaLM-8B	speech-to-speech
Qwen-Audio (Chu et al., 2023)	speech, text	speech, text	Qwen-7B	speech-to-text
SLM (Wang et al., 2023b)	speech, text	text	T5-13B	dialogue
TWIST (Hassid et al., 2023)	speech	speech	OPT-1.3B	speech continuation
SpeechGPT (Zhang et al., 2023a)	speech, text	speech, text	LLaMA-13B	spoken dialogue
Pengi (Deshmukh et al., 2023)	sound	text	x	open-ended sound understanding
LTU (Gong et al., 2024)	sound	text	LLaMA-7B	open-ended sound understanding
Audio Flamingo (Kong et al., 2024)	sound	text	OPT-1.3B	open-ended sound understanding
SpiRit-LM (Nguyen et al., 2024)	speech, text	speech, text	LLaMA-7B	expressive speech continuation
Moshi (Défossez et al., 2024)	speech	speech	Helium-7B	real-time spoken dialogue
video-SALMONN (Sun et al., 2024)	video, speech	text	Vicuna-13B	audio-visual QA
AV-Gemma (this work)	image	speech, text	PaliGemma-{3B,10B}	(pretrain) web-scale spoken captioning
	image, speech / text	speech, text	PaliGemma-{3B,10B}	spoken visual QA with spoken / text questions
	video, image	speech, text	PaliGemma-{3B,10B}	video / long-form spoken captioning
	video, sound	speech, text	PaliGemma-{3B,10B}	sound prediction and sound captioning

Table 1. Comparison of Audio-Visual Gemma with prominent speech language models.

perceives. AV-Gemma is built on top of PaliGemma (Beyer et al., 2024) by integrating audio tokens into the VLM generation. At the core, it natively generates visually grounded audio tokens, which is effective across various audio-visual setups, such as video-to-speech generation, spoken visual QA, and sound event prediction and captioning. See Table 1 for a summary of its tasks.

To train AV-Gemma, we created a new spoken captions dataset based on WebLI (Chen et al., 2023), called Spoken WebLI. Spoken WebLI contains 350 million web-scale (image, spoken caption) pairs, totaling approximately 285,000 hours of high-quality synthetic speech. This makes it several orders of magnitude larger than existing audio-visual corpora (Harwath et al., 2016; Hsu et al., 2021). With Spoken WebLI, we systematically study key scaling properties:

1. Data scaling improves both image-to-text and image-to-speech generation.
2. Model capacity scaling further overcomes the bottlenecks observed in smaller models.
3. Visual-to-semantic generation can be learned in a text-free manner.

We demonstrate that on key VLM benchmarks, AV-Gemma closely matches state-of-the-art VLMs, achieving a CIDEr score of 144.2 on COCO Captioning (Chen et al., 2015) and 80.46% on VQAv2 (Goyal et al., 2017), while exhibiting audio understanding and generation capabilities (e.g., generating spoken captions or prompting with spoken questions). We also observe promising transfer learning results on audio-visual generation tasks without relying on specialized tuned model components, such as pretrained audio encoders. Finally, we achieve significant progress in text-free image-to-speech synthesis, scoring significantly higher naturalness MOS (4.30 v.s. 3.84) and visual correspondence MOS (4.21 v.s. 3.79) scores against previous work (Hsu et al., 2021). We summarize our contributions below:

- **Modeling** (Section 3): We introduce AV-Gemma, a family of audio-visual foundation models for direct visual-to-speech generation. It is capable of describing, in both textual and spoken forms, what it perceives visually and auditorily. AV-Gemma is built upon the state-of-the-art VLM, PaliGemma, by integrating audio tokens into the generation process.

- **Data** (Section 4): To pretrain AV-Gemma, we collected the largest-scale audio-visual dataset to date, Spoken WebLI. Compared to existing corpora, its data source is web-scale, and the captions are generated via a high-quality production speech synthesizer.

- **Scaling** (Section 5): We carefully study audio-visual scaling properties, exploring *how* larger data or model capacity boosts performance and examining the role of textual supervision in visual-to-semantic generation.

- **Versatile Transfer** (Section 6): AV-Gemma achieves competitive performance on spoken visual QA, demonstrates robust transfer capabilities across diverse spoken captioning scenarios, and performs well in sound event understanding out-of-the-box.

We believe the findings in this work provide valuable insights for building future audio-visual foundation models.

2. Related Work

2.1. Visual Language Models

PaLI series (Chen et al., 2023; Beyer et al., 2024)

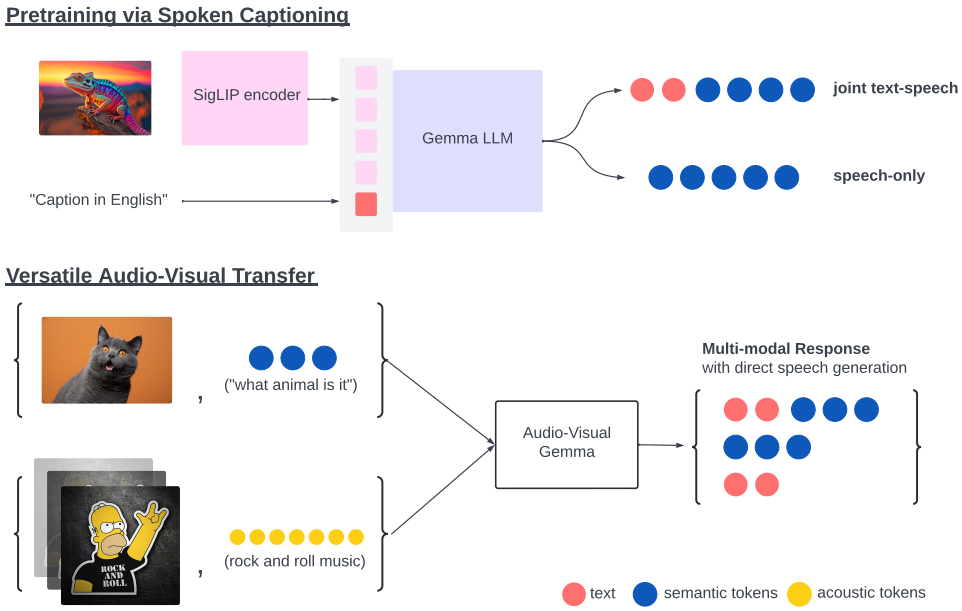

Figure 1. **Overview of Audio-Visual Gemma (AV-Gemma).** AV-Gemma consists of a SigLIP visual encoder, a Gemma LLM, and an expanded vocabulary to query and generate audio tokens. **(top)** AV-Gemma is pretrained via a direct spoken captioning objective on web-scale spoken captions. We provide two pretraining variants: joint text-speech and speech-only. In either case, audio tokens are predicted directly from the Gemma LLM natively without having a speech decoder component. **(bottom)** AV-Gemma works with a suite of audio-visual transfer setups, such as video-to-speech, spoken visual QA, and sound events understanding. It adopts an encoder-free approach to contextual audio understanding – speech and general sounds are presented directly as hard tokens in-context to the model.

2.2. From Audio Codec to Spoken Language Model

2.3. Visually-Grounded Speech Models

2.4. Direct Visual-to-Audio Generation

3. Model

AV-Gemma is designed to describe what it perceives visually and auditorily, in both textual and spoken forms. Figure 1 provides an overview of the model setup. Building upon PaliGemma (Beyer et al., 2024; Steiner et al., 2024), AV-Gemma is available in two sizes (3B and 10B), two resolutions (224px^2 and 448px^2), and supports two generation modes (joint text-speech and speech-only).

In this section, we begin by outlining the pretraining objective that enables AV-Gemma's native ability for visual-to-speech generation. We then describe the model components and training procedures that implement this objective effectively. Finally, we detail how contextual audio understanding is integrated into this framework during finetuning.

3.1. Pretraining Objective

AV-Gemma is pretrained via spoken captioning of images. Given a dataset $\mathcal{D} = (\mathbf{v}_i, \mathbf{x}_i, \mathbf{y}_i)$, where $\mathbf{v}$ represents an image, $\mathbf{x}$ is the corresponding text caption, and $\mathbf{y}$ is the corresponding spoken caption, the pretraining objective is defined for two distinct modes:

1. Joint text-speech generation: AV-Gemma predicts both the text caption $\mathbf{x}$ and the spoken caption $\mathbf{y}$ jointly, modeling the probability $p(\mathbf{x}, \mathbf{y}|\mathbf{v})$. The model assumes paired $(\mathbf{x}_i, \mathbf{y}_i)$ samples during training and optimizes both text generation and audio-text alignment conditioned on the visual input $\mathbf{v}$.

2. Speech-only generation: A specialized variant of AV-Gemma is trained to predict the spoken caption $\mathbf{y}$ directly from the visual input $\mathbf{v}$, modeling $p(\mathbf{y}|\mathbf{v})$. This mode bypasses intermediate textual representations, focusing solely on maximizing the likelihood of the spoken output.

As our primary interest in both $p(\mathbf{x}, \mathbf{y}|\mathbf{v})$ and $p(\mathbf{y}|\mathbf{v})$ lies in generating the underlying semantics of spoken content that aligns with the visual inputs, AV-Gemma simplifies the modeling task by focusing on the underlying semantic (also known as phonetic) representations of $\mathbf{y}$. This approach ignores acoustic representations like speaker variation and fine-grained acoustic details that are irrelevant to visual semantics. Following AudioPaLM (Rubenstein et al., 2023), we represent $\mathbf{y}$ as sequences of discrete semantic tokens, denoted as $\mathbf{s} = \{s_k\}_{k=1}^{C_s}$, where C_s is the semantic codebook size. AV-Gemma then learns to autoregressively predict semantic tokens conditioned on visual inputs, modeling the speech-only generation task as $p(\mathbf{s}_t|\mathbf{s}_{<t}, \mathbf{v})$.

Under this formulation, the joint text-speech generation objective $p(\mathbf{x}, \mathbf{y}|\mathbf{v})$ becomes $p(\mathbf{x}, \mathbf{s}|\mathbf{v})$, and

$$p(\mathbf{x}, \mathbf{y}|\mathbf{v}) = p(\mathbf{x}|\mathbf{v}) \cdot p(\mathbf{s}|\mathbf{v}, \mathbf{x}) = \prod_{n=1}^{N} p(\mathbf{x}_n|\mathbf{x}_{<n}, \mathbf{v}) \cdot \prod_{t=1}^{T} p(\mathbf{s}_t|\mathbf{s}_{<t}, \mathbf{v}, \mathbf{x}), \quad (1)$$

Maximizing this objective optimizes both image-to-text generation and the speech-text alignment.

By predicting semantic tokens from images, AV-Gemma directly learns visual-to-speech generation[1]. The difference between the two modes lies in the complexity of the learning task: joint text-speech generation decomposes the problem into two subtasks—visual-to-semantic learning followed by semantic-to-speech learning—making it inherently easier. In contrast, speech-only generation combines these steps, requiring the model to learn both, making it a more challenging setup.

3.2. Architecture

The AV-Gemma architecture builds upon PaliGemma (Beyer et al., 2024; Steiner et al., 2024). The vision backbone is a contrastively pretrained SigLIP-So400m vision encoder (Alabdulmohsin et al., 2023; Zhai et al., 2023), followed by a linear projection layer. This projection maps the visual features into sequences of "image tokens." A 224px^2 image generates 256 tokens, while a 448px^2 image produces 1024 tokens. Video inputs are processed as sequences of images, where each frame is tokenized similarly; for example, 10 frames of 224px^2 resolution yield 2560 image tokens.

These image tokens are concatenated with text and audio prefixes to form a unified input sequence, which is then fed into the Gemma LLM (Gemma, 2024a). Predictions are generated autoregressively from the LLM. By default, AV-Gemma is initialized with pretrained checkpoints from PaliGemma, leveraging its strong visual-semantic foundational capabilities. A summary of the model configurations is provided in Table 2.

Semantic Tokens: Following AudioPaLM, we utilize the pretrained USM-v2 semantic tokenizer to convert speech waveforms into discrete tokens (Rubenstein et al., 2023)[2]. The tokenizer features a vocabulary of 32k tokens, sampled at a fixed rate of 25Hz (40ms per token). Compared to other popular semantic tokenization methods, such as w2v2+kmeans, USM-v2 demonstrates significantly better speaker invariance and token efficiency measured in BPE, attributed to the auxiliary ASR loss incorporated during training (Vashishth et al., 2024).

Direct Semantic Token Generation: We first modify the Gemma LLM decoder to handle both text and semantic tokens by expanding the token embedding matrix E from size $V \times d$ (where V is Gemma's SentencePiece (Kudo, 2018) vocab size) to $(V + C_s) \times d$. New rows are added to E for the semantic tokens, which are initialized from scratch while retaining the pretrained text embeddings, as in (Zhang et al., 2023a; Rubenstein et al., 2023). The resulting vocab size is 256k + 32k, corresponding to text and semantic tokens, respectively. Next, we serialize the joint text-speech token sequence to the decoder as follows:

```
tokens = [image tokens, BOS, prefix, SEP,
text suffix, audio suffix, EOS, PAD]
```

During pretraining, the prefix is fixed to "Caption in English." For speech-only generation, the text suffix is omitted. Notably, as text-free generation is also our focus, AV-Gemma avoids incorporating additional components, such as a separate speech decoder in LLAMA-OMNI (Fang et al., 2024), for speech modeling. This architectural simplicity underpins our emphasis on direct visual-to-audio modeling.

Waveform Synthesis: Finally, the semantic tokens are converted into SoundStream tokens (Borsos et al., 2023a) using the SoundStorm decoder (Borsos et al., 2023b), which efficiently reconstructs high-quality audio and controls speaker characteristics. By default, the speaker is randomly sampled unless a 3-second voice prompt specifies the target speaker (Wang et al., 2023a).

3.3. Training and Decoding

Following the PaLI series, AV-Gemma's training consists of three stages, with the role of textual annotations in each stage carefully analyzed in Section 5:

- **Stage 0:** Text-image pretraining, initialized with existing PaliGemma checkpoints. In this stage, the model is trained on a diverse mixture of vision-language tasks to develop a broad range of "skills" (Beyer et al., 2024). These tasks include image captioning, OCR, visual question answering (VQA), object detection, segmentation, and grounded captioning. The task mixture is empirically weighted, and the model is trained for 1 billion examples.
- **Stage 1:** Audio-visual pretraining for joint text-speech or speech-only generation. Following Stage 0, the model undergoes continued pretraining for direct spoken captioning, as detailed in Section 3.1. Based on findings from LiT (Zhai et al., 2022b), the image encoder is frozen while the linear projection layer and Gemma decoder are fine-tuned. Departing from the "infinite" learning-rate scheduler (Zhai et al., 2022a) used in PaliGemma, we

[1]Most audio-visual learning models, such as DAVEnet (Harwath et al., 2016), VG-HuBERT (Peng & Harwath, 2022), SpeechCLIP (Shih et al., 2022), and CAV-MAE (Gong et al., 2023), are trained with utterance-level contrastive objectives. These models also learn visual-to-speech alignment but are limited to retrieval and classification tasks.

[2]Although USM-v2 is designed to encode primarily semantic information, it retains some acoustic details, such as pronunciation, within the token sequences. Consequently, AV-Gemma still inherently models certain acoustic characteristics alongside semantics.

Model	Vision Encoder	LLM	Params.	Pretraining Costs (TPU hours)	
				$224px^2$	$448px^2$
AV-Gemma 3B	SigLIP-So400m	Gemma1 2B	3.0B	110k	160k
AV-Gemma 10B		Gemma2 9B	9.7B	490k	N/A

Table 2. **AV-Gemma model configurations.** All models were pretrained for 2B audio-visual examples on 512 Cloud TPUv3 instances. Compared to PaliGemma (Steiner et al., 2024), the higher costs stem from the longer ($\sim$10x) audio token sequences.

found that a simple cosine learning-rate scheduler with 10% linear warm-up is sufficient. We borrow the data-parallel and model-parallel sharding implementations from PaliGemma, enabling AV-Gemma to be trained on 512 TPUv3 chips with a batch size of 25600, corresponding to about 20.8 hours of audio per batch. Logits soft-capping is applied to the attention and logits for the 10B model following (Gemma, 2024b). For each model configuration, extensive hyperparameter tuning is performed over learning rate and weight decay. The model is trained for 2 billion examples during this stage.

- **Tranfser:** Task-specific fine-tuning to adapt pretrained AV-Gemma for transfer tasks. The token prefixes and suffixes described in Section 3.2 are modified depending on the task, with representative examples in Table 9. We observe that jointly fine-tuning the image encoder yields improved overall performance, with the same set of hyperparameter tuning applied for each task.

As studied extensively in (Beyer et al., 2024), we apply prefix-LM masking to the token sequence across all stages. This involves block attention over image and prefix tokens, while suffix tokens follow an autoregressive attention pattern. The next-token prediction loss is computed exclusively over the suffix tokens. For generation, AV-Gemma employs greedy decoding by default for both joint text-speech and speech-only modes. Our experience with the commonly-used decoding methods like Top-K or Top-P sampling perform less effectively in our setups.

3.4. Contextual Audio Understanding

Inspired by the in-context learning capabilities of VALL-E (Wang et al., 2023a) for target speaker modeling, AV-Gemma also employs *prompting* to leverage contextual audio understanding for certain transfer tasks, such as spoken visual QA and general sound event recognition. When the model is presented with a reference audio $\mathbf{a}$, we modify the generation objectives, $p(\mathbf{x}, \mathbf{s}|\mathbf{v})$ and $p(\mathbf{s}|\mathbf{v})$, to additionally condition on $\mathbf{a}$, yielding $p(\mathbf{x}, \mathbf{s}|\mathbf{v}, \mathbf{a})$ and $p(\mathbf{s}|\mathbf{v}, \mathbf{a})$ respectively, where $\mathbf{a}$ serves as an audio prompt providing relevant linguistic or acoustic context.

Audio Prompt Representation: We represent audio input $\mathbf{a}$ as a sequence of discrete tokens, analogous to how VALL-E encodes target speaker voices via Encodec codes (Défossez et al., 2022). When $\mathbf{a}$ contains spoken content (e.g., a spoken question), we convert it into semantic token sequences (analogous to $\mathbf{s}$ in Section 3.1). For general acoustic events or non-speech sounds, we instead encode $\mathbf{a}$ into acoustic token sequences via the SoundStream tokenizer (Zeghidour et al., 2021). Following AudioLM's approach, we linearize SoundStream tokens from the Residual Vector Quantization (RVQ) codebooks into a single sequence by interleaving codebook samples in time, aligning with the "Flattening Pattern" in MusicGen (Copet et al., 2024). Each RVQ codebook has a vocabulary of 1024 tokens; thus, if we use the top Q codebooks, we expand the Gemma LLM vocabulary to $256\text{k} + 32\text{k} + 1024 \times Q$. In either case, the tokenized audio prompt $\mathbf{a}$ is included as part of the prefix tokens, enabling AV-Gemma to incorporate contextual information from both linguistic and non-linguistic audio during generation.

4. Large-Scale Data Curation

As described in Section 3, AV-Gemma's visual-to-speech generation hinges on predicting spoken captions from images, necessitating pretraining on large-scale (image, spoken caption) pairs. In this section, we first examine existing audio-visual corpora, which remain insufficient in size for our purposes. We then introduce Spoken WebLI, the largest-scale spoken captioning corpus to date.

4.1. Existing Audio-Visual Corpora

Audio-visual research has often been driven by curated spoken captioning corpora. For instance, the Places Audio corpus (Harwath et al., 2016) spurred initial progress in speech-image retrieval and demonstrated preliminary signs of "word-like" unit acquisition from speech (Harwath & Glass, 2017; Harwath et al., 2019). Meanwhile, the introduction of SpokenCOCO (Hsu et al., 2021) enabled text-free image-to-speech synthesis and subsequently advanced the state-of-the-art in speech-image retrieval (Shih et al., 2022) and word discovery (Peng & Harwath, 2022). Table 3 summarizes three widely used visually-grounded speech corpora. Flickr8k Audio and SpokenCOCO each extend an existing image captioning dataset by collecting spoken captions from crowd workers, while Places Audio comprises free-form spoken descriptions of large-scale scene images from the Places205 dataset.

Dataset	# Hours	# Spoken Captions	Language	Collection Method	Data Source
Flickr8k Audio (Harwath & Glass, 2015)	46	40k	English	crowd-sourced (scripted)	Flickr 8k
Places Audio (Harwath et al., 2016)	936	500k	English and Hindi	crowd-sourced (spontaneous)	Places 205
SpokenCOCO (Hsu et al., 2021)	742	605k	English	crowd-sourced (scripted)	MSCOCO
Spoken WebLI (this work)	285,000	350M	English	production TTS	WebLI

Table 3. Comparison of Spoken WebLI with notable audio-visual spoken captioning datasets.

Although each dataset provides paired spoken captions and images, together they add up to fewer than 2,000 hours of speech, which is significantly below the 100k+ hours used in large-scale speech-to-text training (Radford et al., 2023; Zhang et al., 2023b; Peng et al., 2023). For comparison, they are also much smaller than the billion-scale image-caption corpora utilized by vision-language models, such as PaLI (Chen et al., 2023) and SigLIP (Zhai et al., 2023). Hence, despite their utility, existing audio-visual corpora fall short in scale for training AV-Gemma.

4.2. Spoken WebLI: A Web-Scale Audio-Visual Corpus

To overcome the stated limitations and enable large-scale training, we introduce *Spoken WebLI*, a new audio-visual corpus derived from the WebLI dataset (Chen et al., 2023). Spoken WebLI comprises 350M speech-image pairs, totaling roughly 285k hours of speech, which is orders of magnitude larger than existing datasets (see Table 3). As it is derived from the public web, the dataset spans a broad range of domains, offering greater diversity and robustness.

Cost-Efficient Synthetic Annotation. At this scale, crowd-sourcing spoken captions would be prohibitively expensive and time-consuming. For reference, collecting about 600k spoken captions in SpokenCOCO already cost $10k. To mitigate these costs, we instead synthesize WebLI's text caption[3] using text-to-speech (TTS) engines. This approach takes advantage of high-quality TTS while eliminating factors irrelevant to audio-visual modeling (e.g. microphone artifacts, speaker variation, speaking pauses). Furthermore, the growing popularity of synthetic data in speech research (Gong et al., 2024; Yeo et al., 2024; Cornell et al., 2024; Lu et al., 2024) affirms its viability for large-scale training.

Large-Scale TTS Inference. We start with the English subset of WebLI and apply additional quality filtering to obtain 350M high-quality text-image pairs. We then leverage production English TTS that is based on Virtuoso (Saeki et al., 2023) to synthesize the text into spoken captions. To speed up the process, we parallelize the TTS inference to 1000 TPUs. As a result, we produce 285k hours of synthetic speech at a fraction of the cost of manual recording, yielding the largest audio-visual dataset to date. Compared to SpokenCOCO or Places Audio, Spoken WebLI not only scales to hundreds of thousands of hours but also encompasses real-world linguistic and visual diversity from web data. To inspect the dataset samples, please refer to the demo page.

[3]The WebLI text captions are extracted from the "alt-text" field.

5. Audio-Visual Scaling Properties

In this section, we discuss three key scaling properties observed during audio-visual pretraining of **AV-Gemma** (Section 3) on **Spoken WebLI** (Section 4). Through these observations, we aim to empirically dissect *how* scaling a spoken captioning objective on large-scale speech-image pairs is effective. Before diving in, we highlight the following claims:

Claim 1 (Data Scaling): Increasing the scale of pretraining data boosts AV-Gemma's ability to generate text (image-to-text) and speech (image-to-speech).

Claim 2 (Model Capacity): Larger model capacity further improves both the semantic quality of visual-to-text outputs and the fidelity of speech generation.

Claim 3 (Textual Annotation): Effective visual-to-semantic generation can be learned end-to-end, without relying on explicit text–image alignment.

5.1. Scaling Experimental Setup

We conducted a series of AV-Gemma pretraining experiments to support these claims by systematically varying dataset size, model capacity, and model generation modes. Our base configuration is an AV-Gemma 3B model with 224^2 input resolution and a joint text–speech generation objective, trained on the full 285k hours of Spoken WebLI.

Model Generation Evaluation. For our scaling studies, we measure spoken captioning quality on the Spoken WebLI test split and on SpokenCOCO (via further finetuning):

- *Joint text–speech generation*: We use CIDEr (Vedantam et al., 2015) to compare the model's text output against ground-truth captions, and measure speech quality via WER, using the text output as the reference (since speech generation essentially acts as TTS of the text output)
- *Speech-only generation*: We use ASR-CIDEr by transcribing the model-generated speech, then computing the CIDEr score against ground-truth text.

We convert the decoded audio tokens into speech waveforms using a vocoder with an enrolled speaker. We then

employ Whisper (Radford et al., 2023) to transcribe the resulting audio. Table 4 reports the captioning results of our base AV-Gemma configuration (fine-tuned on Spoken-COCO), while Table 5 shows the speech-generation results from the same fine-tuned checkpoint. Notably, our base model's speech generation quality closely matches the system's upper bound.Moreover, despite its simultaneous speech-generation capability, AV-Gemma's text generation performance approaches that of state-of-the-art VLMs (Table 4).

Method	CIDEr ↑
AnyGPT (8B, 224^2) (Zhan et al., 2024)	107.5
Flamingo (9B, 288^2) (Alayrac et al., 2022)	138.1
Sim VLM (632M, 224^2) (Wang et al., 2022)	143.3
CoCa (2.1B, 288^2) (Yu et al., 2022)	143.6
BLIP-2 (2.7B, 364^2) (Li et al., 2023)	145.8
PaliGemma (3B, 224^2) (Beyer et al., 2024)	141.9
PaliGemma (10B, 224^2) (Steiner et al., 2024)	143.7
PaliGemma (28B, 224^2) (Steiner et al., 2024)	144.0
AV-Gemma (3B, 224^2, joint text-speech generation)	144.2

Table 4. Captioning results on COCO Caption (Karpathy test split).

Model	WER / CER ↓
AV-Gemma 3B speech generation + vocoder	5.39 / 3.28
Ground Truth semantic tokens + vocoder	6.63 / 3.03
Ground Truth text caption + Virtuoso TTS	1.01 / 0.11

Table 5. AV-Gemma's speech generation quality against our system upper bound (middle row) on COCO Caption.

5.2. Claim 1: Data Scaling Improves Generation

Setup. Building on the base AV-Gemma configuration from Section 5.1, we investigate how varying the pretraining dataset size affects model generation quality. We construct multiple subsets of Spoken WebLI, ranging from as little as 3k hours to the full 285k hours. All models are pretrained for the same number of steps with early stopping, using the same hyperparameter sweeps. Stage 1 pretraining hours (e.g., 3k, 30k, 150k) serve as our primary "data scale" measure. To quantify data-scaling gains, we employ a *CIDEr regret* metric, measuring the relative performance gap compared to using the entire 285k-hour Spoken WebLI (Section 4). A lower regret suggests that adding more training data yields diminishing returns, whereas a higher regret indicates missed gains when training with fewer hours. Figure 2 illustrates how AV-Gemma's text generation quality evolves with increasing data scale. Meanwhile, Figure 3 shows how speech generation quality (via ASR-CIDEr) changes over the number of training examples seen for both joint text–speech and speech-only generation modes.

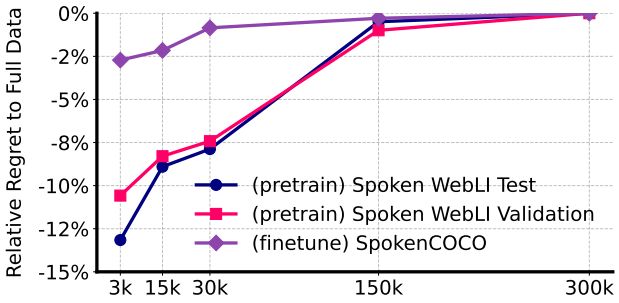

Figure 2. **AV-Gemma 3B data scaling curves.** The x-axis shows the hours of Spoken WebLI data used during Stage 1 pretraining, and the y-axis plots the relative pretraining and finetuning performance regret (in CIDEr) compared to using the full 285k hours.

Key Findings. Our results exhibit a consistent positive trend: scaling up the training data yields higher-quality text captions (image-to-text) and more accurate speech generation (image-to-speech). From Figure 2, we observe substantial improvement when increasing the pretraining data from 3k to 150k hours, after which the 3B AV-Gemma model experiences diminishing returns. Nonetheless, the overall gains remain meaningful at larger scales. In parallel, the ASR-CIDEr curves from Figure 3 confirm that more training data provides broader coverage, improving speech generation in both joint text–speech and speech-only modes. Overall, these findings suggest that *more data systematically boosts* AV-Gemma's audio-visual generation capabilities and justify our large-scale data curation efforts.

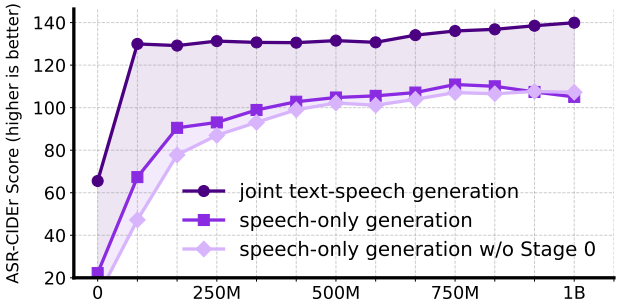

Figure 3. **AV-Gemma 3B speech-only generation scaling curves.** The x-axis depicts number of training examples seen, and the y-axis plots ASR-CIDEr (higher is better). Shaded regions highlight relative performance differences between pretraining via (*joint text-speech*), (*speech-only*), and (*speech-only without VLM init*).

5.3. Claim 2: Model Capacity Improves Generation

Setup. Recall from Section 5.2 (Figure 2) that we observed diminishing text-generation returns beyond 150k hours for AV-Gemma 3B. Additionally, Figure 4 (blue curve) shows that the model's text generation CIDEr score improves steadily over training, whereas the speech generation WER (pink curve) plateaus at ∼20% relatively early in training.

These observations suggest that *vertical scaling* of model capacity might unlock further gains in both CIDEr and WER. Thus, we train a 10B AV-Gemma model (Table 2) on the full Spoken WebLI and compare its audio–visual generation performance with the 3B base model.

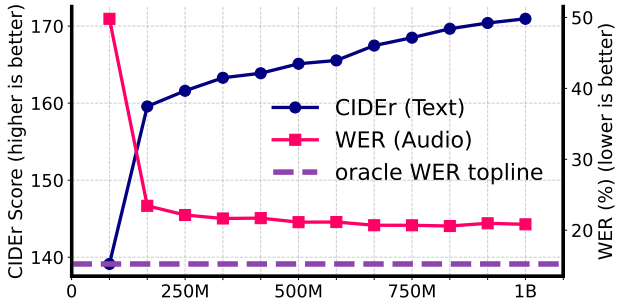

Figure 4. **AV-Gemma 3B joint text-speech generation scaling curves.** The x-axis depicts pretraining compute (number of training examples seen), while the left y-axis plots CIDEr for text generation and the right y-axis shows WER for audio generation. The bottom dashed line indicates our system's WER upper bound.

Key Findings. Table 6 shows that moving from AV-Gemma 3B to 10B boosts CIDEr from 176 to 217 on the Spoken WebLI test set, effectively eliminating the plateau observed in Figure 2. Meanwhile, Table 7 reveals that WER drops from 21.9% to 17.25%, approaching the top-line performance indicated by the dashed line in Figure 4.[4] Thus, larger model capacity notably improves both text (CIDEr) and speech (WER) generation, mitigating the scaling bottlenecks seen with the 3B base configuration.

Model	Image Res.	CIDEr ↑
AV-Gemma 3B	224x	176.27
AV-Gemma 3B	448x	182.33
AV-Gemma 10B	224x	217.02

Table 6. Comparison of captioning results on the Spoken WebLI test split for different AV-Gemma capacities and input resolutions.

Model	WER / CER ↓
AV-Gemma 3B speech generation + vocoder	21.9 / 11.33
AV-Gemma 10B speech generation + vocoder	17.25 / 6.86
Ground Truth semantic tokens + vocoder	15.24 / 4.87

Table 7. Scaling AV-Gemma from 3B to 10B improves speech generation quality on Spoken WebLI test split.

[4]The high WER on Spoken WebLI partly stems from noisy web captions.

5.4. Claim 3: Effective Visual-to-Semantic Generation without Text–Image Supervision

Setup. VLMs typically learn visual–semantic representations from large text–image corpora, treating text as the main output. Our base AV-Gemma adds a joint text–speech generation objective (Stage 1), initialized from a pretrained PaliGemma checkpoint (Stage 0). One might argue that Stage 1 pretraining merely aligns speech with text (TTS), given the strong visual-to-text skills inherited from PaliGemma. To examine the role of textual annotation, we train two *speech-only* AV-Gemma variants: (i) one still initialized from Stage 0 (VLM init), and (ii) one without any VLM initialization.

Prior work on visually grounded speech (Harwath & Glass, 2019; Harwath et al., 2019) has shown that training on audio–visual pairs can induce linguistic unit acquisition. VG-HuBERT (Peng & Harwath, 2022) and SpeechCLIP (Shih et al., 2022) demonstrate how word boundaries align with visual words. Meanwhile, (Hsu et al., 2021) built a direct image-to-speech system via learned audio–visual units, demonstrating the feasibility of text-free visual-to-speech synthesis. Following this line of work, we compare our speech-only AV-Gemma variant with (Hsu et al., 2021) using subjective Mean Opinion Scores (MOS) that account for both synthesis quality (naturalness) and visual correctness.

Key Findings. Figure 3 compares the speech-only AV-Gemma variants scaling curves against the joint text–speech model, measured by ASR-CIDEr. Despite lacking direct text–image supervision, both speech-only variants learn adequate visual-to-semantic mappings (e.g., reaching ∼110 vs. 140 in ASR-CIDEr). Interestingly, both speech-only models perform similarly, with the Stage 0 VLM initialization mainly speeding up training convergence rather than yielding a higher final score. These indicate that textual annotation is *not strictly necessary* for image-to-speech generation, provided there is sufficient spoken captioning data and model capacity. Table 8 further shows that AV-Gemma surpasses (Hsu et al., 2021) in both naturalness and visual-content MOS. Overall, these results underscore that visual–semantic alignment can emerge without direct text–image pairs, and that AV-Gemma effectively leverages large-scale audio–visual data to learn robust speech generation grounded in visual context.

Text-Free Model	Naturalness	Visual Content
Hsu et al. (2021)	3.84	3.79
AV-Gemma (speech-only)	4.30	4.21

Table 8. AV-Gemma attains higher naturalness and more accurate visual correspondence MOS scores than previous text-free models.

Task / Dataset	AV-Gemma Inference	
	Task Prompt	**Model Generation**
	PRETRAINING VIA SPOKEN CAPTIONING	
Spoken WebLI	[IMAGE] Caption in English	[USM8190][USM200][USM719][USM909]...[USM238] 🔊 *watercolor floral bouquet with blue and white flowers*
	VERSATILE AUDIO-VISUAL TRANSFER	
Spoken VQA	[IMAGE][USM23][USM89]...[USM238] 🔊 *Is this a beautiful zebra?*	[USM54][USM78][USM90] 🔊 *Yes*
Moments in Time	[VIDEO] Caption in English	[USM111][USM289][USM134]..[USM90] 🔊 *Two deer are grazing in the woods*
Image Paragraph	[IMAGE] Caption in English	[USM3198][USM4123][USM23901]...[USM17982] 🔊 *(Long-form caption) A red barn with a silver roof is on a farm. There is a red gate on the side of the barn. There is a wooden fence on the side of the barn.*
Epic-Kitchens-100	[EGOCENTRIC VIDEO] Instruct	[USM14][USM44][USM913]...[USM143] 🔊 *(Instruction) Pour the water into the pot*
AudioSet-400k	[VIDEO][SS19][SS21][SS299]...[SS14] 🔊 *(Rock and roll background music)*	Rock and roll
AudioCaps	[VIDEO][SS89][SS285][SS97]...[SS101] Caption in English 🔊 *(A man demonstrates spray painting)*	[USM910][USM798][USM1879]...[USM3891] 🔊 *A man speaks and sprays paint.*

Table 9. **Representative prompts and generations for AV-Gemma.** 🔊 indicates a spoken caption, question, answer, or background audio. Tokens like [USM90] represent semantic tokens, and [SS19] represents SoundStream acoustic tokens.

6. Versatile Transfer

Thus far, we have focused on analyzing AV-Gemma's scaling properties for spoken captioning on images. In this section, we showcase its *versatile transfer* capabilities by evaluating three distinct tasks: (1) spoken visual question answering, (2) spoken captioning in diverse domains, and (3) audio-visual sound understanding. Table 9 outlines representative prompts and model generations across various tasks. Each transfer task requires a different combination of input and output modalities, and Table 1 places these tasks in the context of other prominent speech language models.

6.1. Spoken Visual Question Answering (VQA)

Task Introduction. We evaluate AV-Gemma on the VQAv2 benchmark (Goyal et al., 2017), which tests a model's ability to answer open-ended visual questions. Unlike conventional VLMs that only accept textual queries, AV-Gemma is tasked to handle *spoken questions* and produce *spoken answers*, a task we call *Spoken VQA*. To adapt AV-Gemma for Spoken VQA, we fine-tune on image–question–answer tuples that include both textual and spoken inputs/outputs (see Section 3.4 for how audio prompts are constructed).

Approach. A naive fine-tuning of AV-Gemma for speech-in, speech-out VQA reduces accuracy by about 10% compared to text-based VQA. This reduction can be attributed to the fact that, during pretraining, AV-Gemma was not exposed to spoken questions presented as semantic tokens within its prompts (Table 9). We discovered that additionally prompting the model to "Repeat the question before answering" significantly narrows this gap. This compels the model to first understand the spoken question by transcribing it into text, before proceeding to generate the spoken responses.

Results. Table 10 compares AV-Gemma with strong VLMs, showing that AV-Gemma (with joint text–speech generation) closely matches or slightly underperforms the text-only baseline when all inputs/outputs remain textual. However, AV-Gemma uniquely enables spoken queries or speech-based responses. With both question and answer in speech form, AV-Gemma achieves an accuracy of 80.46% on VQAv2. These results also underscore the effectiveness of our spoken captioning pretraining objective.

Model	Question	Answer	Accuracy
Flamingo (80B, 288^2)	text	text	51.8%
Sim VLM (632M, 224^2)	text	text	80.03%
CoCa (2.1B, 288^2)	text	text	82.3%
BLIP-2 (2.7B, 364^2)	text	text	65.0%
PaliGemma (3B, 224^2)	text	text	82.36%
AV-Gemma (3B, 224^2)	text	text	81.88%
AV-Gemma (3B, 224^2)	speech	text, speech	71.9%
+ (prompt) repeat the question before answering			80.46%

Table 10. **VQAv2 test-dev results.** AV-Gemma supports both spoken queries and spoken responses with competitive accuracy.

6.2. Spoken Captioning in Diverse Domains

Task Introduction. While Spoken WebLI covers a broad range of linguistic and visual diversity for image captions, modern AI assistants increasingly require speech-based descriptions in scenarios beyond static images (e.g., short video clips, first-person views). To explore this broader scope, we extend AV-Gemma's spoken captioning to three additional domains, each with TTS-synthesized speech annotations (as in Spoken WebLI):

- **Moments in Time** (Monfort et al., 2019): A dataset of one million 3-second video clips depicting everyday activities with dynamic visual scenes (people, animals, objects, natural phenomena). Originally annotated in text, which we convert to spoken form.
- **Epic-Kitchen-100** (Damen et al., 2022): A collection of 100 hours of first-person, unscripted cooking videos. We convert the 90k action labels into "spoken instructions."
- **Image Paragraph** (Krause et al., 2017): 20k fine-grained image captions, each consisting of one paragraph. We convert these paragraphs into *long-form* spoken captions of up to 20 seconds.

Approach. We subsample each video to 16 frames at 224^2 resolution, encoding these as separate images. This results in 4096 image tokens. We then concatenate the resulting image tokens, without special separators, to form a unified visual context in the prefix token sequence. We follow the general transfer setup in Section 3.3 and fine-tune AV-Gemma using a joint text-speech generation objective.

Results. Qualitative inspection shows that AV-Gemma adapts successfully to these varied settings, producing accurate video-conditioned spoken captions or instructions, and coherent long-form paragraphs (up to 20 seconds). Quantitatively, AV-Gemma attains CIDEr scores of 38.96 on Moments in Time, 230.21 on Epic-Kitchen-100, and 28.68 on Image Paragraph. Notably, for Epic-Kitchen-100, initializing the model with a Moments-in-Time pretrained checkpoint further improves the score to 237.82.

6.3. Audio-Visual Sound Understanding

Task Introduction. Although AV-Gemma is primarily designed for visual-conditioned speech generation, many real-world applications benefit from additional *audio* comprehension. To demonstrate AV-Gemma's ability to parse generic audio inputs via prompting (Section 3.4), we evaluate two audio–visual sound understanding tasks:

- **AudioSet-400k** (Gemmeke et al., 2017): 400k 10-second sound clips from YouTube,[5] with 527 audio event classes. We follow LTU (Gong et al., 2024) to have the model output text labels directly (open-vocab) at inference.
- **AudioCaps** (Kim et al., 2019): A subset of AudioSet containing 40k sound clips, each with a human-annotated caption. We convert these text captions into spoken form.

Approach. We subsample each video to 10 frames at 224^2 resolution, with a 1-fps frame rate, yielding 2560 image tokens. For acoustic prompts, we linearize SoundStream tokens from the first RVQ codebook (Q=1), as prompting with a larger Q or with semantic tokens results in worse mAP scores. On AudioCaps, AV-Gemma is fine-tuned for joint text-speech generation.

Results. Table 11 reports classification mAP on AudioSet. Without a specialized audio encoder, AV-Gemma lags behind LTU (Gong et al., 2024) and GAMA (Ghosh et al., 2024) in the audio-only setup, but integrating video frames raises the mAP to 21.1%. On AudioCaps, AV-Gemma achieves a 75.41 CIDEr score for spoken captioning, again without any specialized audio components. Qualitative analyses suggest that AV-Gemma often leverages visual context for certain predictions, while relying purely on acoustic prompts for others. Overall, these results illustrate AV-Gemma's capacity to flexibly incorporate audio cues alongside visual context, pointing toward future extensions that may further enhance sound understanding performance.

Model	Input Representations	AudioSet
Pengi	CLAP (Elizalde et al., 2023)	11.5%
Qwen-Audio	Whisper	13.4%
LTU	AST (Gong et al., 2021)	18.7%
GAMA	AST	19.2%
AV-Gemma 3B	audio-only (SoundStream)	12.07%
AV-Gemma 3B	video-only (224^2, 1fps)	17.60%
AV-Gemma 3B	video + audio	21.1%

Table 11. **Open-vocab sound event classification on AudioSet (mAP)**. Although not specialized for audio labeling, AV-Gemma effectively leverages additional visual context for sound prediction.

7. Conclusion

We introduced **AV-Gemma**, a family of audio–visual foundation models for direct visual-to-speech generation, trained on large-scale spoken image captions (Spoken WebLI). Our pretrain scaling studies demonstrated that (1) increasing data systematically boosts both image-to-text and image-to-speech generation, (2) larger model capacity alleviates performance bottlenecks observed in smaller model configurations, and (3) visual-to-semantic generation can still emerge without explicit text–image alignment.

Beyond strong baseline performance on image-based spoken captioning, AV-Gemma's flexible architecture and audio prompting enable seamless transfer to a wide array of

[5]The original dataset contains 2M clips; we retain 400k following compliance policy.

audio–visual tasks—including spoken visual QA, video-conditioned spoken captioning, and open-vocabulary audio-visual sound event recognition—without relying on specialized components. We believe that AV-Gemma's ability to natively integrate visual understanding with speech generation opens pathways for advanced multimodal applications such as real-time assistants, broader support for non-written languages, and expressive audiovisual content creation. Future work will explore improved acoustic tokenization schemes for audio generation, expressivity modeling for visual-conditioned speech, multilingual support, and more efficient joint text-speech decoding mechanism.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv*, 2023.

Alabdulmohsin, I. M., Zhai, X., Kolesnikov, A., and Beyer, L. Getting vit in shape: Scaling laws for compute-optimal model design. *NeurIPS*, 2023.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al. Flamingo: a visual language model for few-shot learning. *NeurIPS*, 2022.

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al. Palm 2 technical report. *arXiv*, 2023.

Beyer, L., Steiner, A., Pinto, A. S., Kolesnikov, A., Wang, X., Salz, D., Neumann, M., Alabdulmohsin, I., Tschannen, M., Bugliarello, E., et al. Paligemma: A versatile 3b vlm for transfer. *arXiv*, 2024.

Borsos, Z., Marinier, R., Vincent, D., Kharitonov, E., Pietquin, O., Sharifi, M., Roblek, D., Teboul, O., Grangier, D., Tagliasacchi, M., et al. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 2023a.

Borsos, Z., Sharifi, M., Vincent, D., Kharitonov, E., Zeghidour, N., and Tagliasacchi, M. Soundstorm: Efficient parallel audio generation. *arXiv*, 2023b.

Brown, T. B. Language models are few-shot learners. *NeurIPS*, 2020.

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., and Zitnick, C. L. Microsoft coco captions: Data collection and evaluation server. *arXiv*, 2015.

Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., et al. Pali: A jointly-scaled multilingual language-image model. *ICLR*, 2023.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., et al. Palm: Scaling language modeling with pathways. *arXiv*, 2022.

Chu, Y., Xu, J., Zhou, X., Yang, Q., Zhang, S., Yan, Z., Zhou, C., and Zhou, J. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv*, 2023.

Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y., and Défossez, A. Simple and controllable music generation. *NeurIPS*, 2024.

Cornell, S., Darefsky, J., Duan, Z., and Watanabe, S. Generating data with text-to-speech and large-language models for conversational speech recognition. *arXiv*, 2024.

Dai, W., Lee, N., Wang, B., Yang, Z., Liu, Z., Barker, J., Rintamaki, T., Shoeybi, M., Catanzaro, B., and Ping, W. Nvlm: Open frontier-class multimodal llms. *arXiv*, 2024.

Damen, D., Doughty, H., Farinella, G. M., Furnari, A., Kazakos, E., Ma, J., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision*, pp. 1–23, 2022.

DeepMind. A research prototype exploring future capabilities of a universal ai assistant. https://deepmind.google/technologies/project-astra/, 2025.

Défossez, A., Copet, J., Synnaeve, G., and Adi, Y. High fidelity neural audio compression. *arXiv*, 2022.

Défossez, A., Mazaré, L., Orsini, M., Royer, A., Pérez, P., Jégou, H., Grave, E., and Zeghidour, N. Moshi: a speech-text foundation model for real-time dialogue. *arXiv*, 2024.

Deshmukh, S., Elizalde, B., Singh, R., and Wang, H. Pengi: An audio language model for audio tasks. *NeurIPS*, 2023.

Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv*, 2020.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Yang, A., Fan, A., et al. The llama 3 herd of models. *arXiv*, 2024.

Elizalde, B., Deshmukh, S., Al Ismail, M., and Wang, H. Clap learning audio concepts from natural language supervision. 2023.

Fang, Q., Guo, S., Zhou, Y., Ma, Z., Zhang, S., and Feng, Y. Llama-omni: Seamless speech interaction with large language models. *arXiv*, 2024.

Gemini, Anil, R., Borgeaud, S., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., Millican, K., et al. Gemini: a family of highly capable multimodal models. *arXiv*, 2023.

Gemma, T. Gemma: Open models based on gemini research and technology. *arXiv*, 2024a.

Gemma, T. Gemma 2: Improving open language models at a practical size. *arXiv*, 2024b.

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. Audio set: An ontology and human-labeled dataset for audio events. 2017.

Ghosh, S., Kumar, S., Seth, A., Evuru, C. K. R., Tyagi, U., Sakshi, S., Nieto, O., Duraiswami, R., and Manocha, D. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities. *arXiv*, 2024.

Gong, Y., Chung, Y.-A., and Glass, J. Ast: Audio spectrogram transformer. *Interspeech*, 2021.

Gong, Y., Rouditchenko, A., Liu, A. H., Harwath, D., Karlinsky, L., Kuehne, H., and Glass, J. Contrastive audio-visual masked autoencoder. *ICLR*, 2023.

Gong, Y., Luo, H., Liu, A. H., Karlinsky, L., and Glass, J. Listen, think, and understand. *ICLR*, 2024.

Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., and Parikh, D. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *CVPR*, 2017.

Harwath, D. and Glass, J. Deep multimodal semantic embeddings for speech and images. 2015.

Harwath, D. and Glass, J. Towards visually grounded subword speech unit discovery. 2019.

Harwath, D. and Glass, J. R. Learning word-like units from joint audio-visual analysis. *ACL*, 2017.

Harwath, D., Torralba, A., and Glass, J. Unsupervised learning of spoken language with visual context. *NeurIPS*, 2016.

Harwath, D., Hsu, W.-N., and Glass, J. Learning hierarchical discrete linguistic units from visually-grounded speech. *ICLR*, 2019.

Hassid, M., Remez, T., Nguyen, T. A., Gat, I., Conneau, A., Kreuk, F., Copet, J., Defossez, A., Synnaeve, G., Dupoux, E., et al. Textually pretrained speech language models. *NeurIPS*, 2023.

Hsu, W.-N., Harwath, D., Song, C., and Glass, J. Text-free image-to-speech synthesis using learned segmental units. *ACL*, 2021.

Kim, C. D., Kim, B., Lee, H., and Kim, G. Audiocaps: Generating captions for audios in the wild. 2019.

Kong, Z., Goel, A., Badlani, R., Ping, W., Valle, R., and Catanzaro, B. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. *arXiv*, 2024.

Krause, J., Johnson, J., Krishna, R., and Fei-Fei, L. A hierarchical approach for generating descriptive image paragraphs. 2017.

Kudo, T. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *EMNLP*, 2018.

Lakhotia, K., Kharitonov, E., Hsu, W.-N., Adi, Y., Polyak, A., Bolte, B., Nguyen, T.-A., Copet, J., Baevski, A., Mohamed, A., et al. On generative spoken language modeling from raw audio. *TACL*, 2021.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. 2023.

Lin, J., Yin, H., Ping, W., Molchanov, P., Shoeybi, M., and Han, S. Vila: On pre-training for visual language models. In *CVPR*, 2024.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *NeurIPS*, 2023.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning. In *CVPR*, 2024.

Lu, Y., Song, J., Chang, X., Bian, H., Maiti, S., and Watanabe, S. Syneslm: A unified approach for audio-visual speech recognition and translation via language model and synthetic data. *arXiv*, 2024.

McKinzie, B., Gan, Z., Fauconnier, J.-P., Dodge, S., Zhang, B., Dufter, P., Shah, D., Du, X., Peng, F., Weers, F., et al. Mm1: Methods, analysis & insights from multimodal llm pre-training. *arXiv*, 2024.

Monfort, M., Andonian, A., Zhou, B., Ramakrishnan, K., Bargal, S. A., Yan, T., Brown, L., Fan, Q., Gutfreund, D., Vondrick, C., et al. Moments in time dataset: one million videos for event understanding. *IEEE transactions on pattern analysis and machine intelligence*, 2019.

Nguyen, T. A., Muller, B., Yu, B., Costa-Jussa, M. R., Elbayad, M., Popuri, S., Duquenne, P.-A., Algayres, R., Mavlyutov, R., Gat, I., et al. Spirit-lm: Interleaved spoken and written language model. *arXiv*, 2024.

Peng, P. and Harwath, D. Word discovery in visually grounded, self-supervised speech models. *Interspeech*, 2022.

Peng, Y., Tian, J., Yan, B., Berrebbi, D., Chang, X., Li, X., Shi, J., Arora, S., Chen, W., Sharma, R., et al. Reproducing whisper-style training using an open-source toolkit and publicly available data. 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. Robust speech recognition via large-scale weak supervision. *ICML*, 2023.

Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv*, 2021.

Rubenstein, P. K., Asawaroengchai, C., Nguyen, D. D., Bapna, A., Borsos, Z., Quitry, F. d. C., Chen, P., Badawy, D. E., Han, W., Kharitonov, E., et al. Audiopalm: A large language model that can speak and listen. *arXiv*, 2023.

Saeki, T., Zen, H., Chen, Z., Morioka, N., Wang, G., Zhang, Y., Bapna, A., Rosenberg, A., and Ramabhadran, B. Virtuoso: Massive multilingual speech-text joint semi-supervised learning for text-to-speech. 2023.

Shih, Y.-J., Wang, H.-F., Chang, H.-J., Berry, L., Lee, H.-y., and Harwath, D. Speechclip: Integrating speech with pre-trained vision and language model. *SLT*, 2022.

Steiner, A., Pinto, A. S., Tschannen, M., Keysers, D., Wang, X., Bitton, Y., Gritsenko, A., Minderer, M., Sherbondy, A., Long, S., et al. Paligemma 2: A family of versatile vlms for transfer. *arXiv*, 2024.

Sun, G., Yu, W., Tang, C., Chen, X., Tan, T., Li, W., Lu, L., Ma, Z., Wang, Y., and Zhang, C. video-salmonn: Speech-enhanced audio-visual large language models. *ICML*, 2024.

Tong, S., Brown, E., Wu, P., Woo, S., Middepogu, M., Akula, S. C., Yang, J., Yang, S., Iyer, A., Pan, X., et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv*, 2024.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv*, 2023.

Vashishth, S., Singh, H., Bharadwaj, S., Ganapathy, S., Asawaroengchai, C., Audhkhasi, K., Rosenberg, A., Bapna, A., and Ramabhadran, B. Stab: Speech tokenizer assessment benchmark. *arXiv*, 2024.

Vedantam, R., Lawrence Zitnick, C., and Parikh, D. Cider: Consensus-based image description evaluation. 2015.

Wang, C., Chen, S., Wu, Y., Zhang, Z., Zhou, L., Liu, S., Chen, Z., Liu, Y., Wang, H., Li, J., et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv*, 2023a.

Wang, M., Han, W., Shafran, I., Wu, Z., Chiu, C.-C., Cao, Y., Chen, N., Zhang, Y., Soltau, H., Rubenstein, P. K., et al. Slm: Bridge the thin gap between speech and text foundation models. In *ASRU*, 2023b.

Wang, Z., Yu, J., Yu, A. W., Dai, Z., Tsvetkov, Y., and Cao, Y. Simvlm: Simple visual language model pretraining with weak supervision. *ICLR*, 2022.

Wu, H., Chen, X., Lin, Y.-C., Chang, K.-w., Chung, H.-L., Liu, A. H., and Lee, H.-y. Towards audio language modeling-an overview. *arXiv*, 2024.

Yeo, J. H., Kim, M., Watanabe, S., and Ro, Y. M. Visual speech recognition for languages with limited labeled data using automatic labels from whisper. 2024.

Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., and Wu, Y. Coca: Contrastive captioners are image-text foundation models. arxiv 2022. *TMLR*, 2022.

Zeghidour, N., Luebs, A., Omran, A., Skoglund, J., and Tagliasacchi, M. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2021.

Zhai, X., Kolesnikov, A., Houlsby, N., and Beyer, L. Scaling vision transformers. 2022a.

Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., and Beyer, L. Lit: Zero-shot transfer with locked-image text tuning. 2022b.

Zhai, X., Mustafa, B., Kolesnikov, A., and Beyer, L. Sigmoid loss for language image pre-training. In *CVPR*, 2023.

Zhan, J., Dai, J., Ye, J., Zhou, Y., Zhang, D., Liu, Z., Zhang, X., Yuan, R., Zhang, G., Li, L., et al. Anygpt: Unified multimodal llm with discrete sequence modeling. *ACL*, 2024.

Zhang, D., Li, S., Zhang, X., Zhan, J., Wang, P., Zhou, Y., and Qiu, X. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *EMNLP*, 2023a.

Zhang, Y., Han, W., Qin, J., Wang, Y., Bapna, A., Chen, Z., Chen, N., Li, B., Axelrod, V., Wang, G., et al. Google usm: Scaling automatic speech recognition beyond 100 languages. *arXiv*, 2023b.