

# Vision-Speech Models: Teaching Speech Models to Converse about Images

Amélie Royer\*

Kyutai

amelie@kyutai.org

Moritz Böhle\*

Kyutai

moritz@kyutai.org

Gabriel de Marmiesse

Kyutai

gabriel@kyutai.org

Laurent Mazaré

Kyutai

laurent@kyutai.org

Neil Zeghidour

Kyutai

neil@kyutai.org

Alexandre Défossez

Kyutai

alex@kyutai.org

Patrick Pérez

Kyutai

patrick@kyutai.org

## Abstract

*The recent successes of Vision-Language models raise the question of how to equivalently imbue a pretrained speech model with visual understanding, an important milestone towards building a multimodal speech model able to freely converse about images. Building such a conversational Vision-Speech model brings its unique challenges: (i) paired image-speech datasets are much scarcer than their image-text counterparts, (ii) ensuring real-time latency at inference is crucial thus bringing compute and memory constraints, and (iii) the model should preserve prosodic features (e.g., speaker tone) which cannot be inferred from text alone. In this work, we introduce Moshivis, augmenting a recent dialogue speech LLM, Moshi, with visual inputs through lightweight adaptation modules. An additional dynamic gating mechanism enables the model to more easily switch between the visual inputs and unrelated conversation topics. To reduce training costs, we design a simple one-stage, parameter-efficient fine-tuning pipeline in which we leverage a mixture of image-text (i.e., “speechless”) and image-speech samples. We evaluate the model on downstream visual understanding tasks with both audio and text prompts, and report qualitative samples of interactions with Moshivis. Our inference code, the image-speech data used for audio evaluation, as well as additional information are available at [github.com/kyutai-labs/moshivis](https://github.com/kyutai-labs/moshivis).*

## 1. Introduction

Vision Language Models (VLMs) have recently gained increasing attention, e.g. [2, 4, 5, 9, 23, 36, 44], showcasing strong capabilities across a variety of visual understanding tasks such as question answering, image captioning or complex reasoning over visual inputs. A core challenge of train-

ing VLMs, or multimodal models in general, is to build well-aligned embeddings of the different input modalities. To this end, the VLM research community has built up vast datasets of paired image and text data over the years, covering many vision understanding tasks [9, 11, 18, 19, 21, 32]. In comparison, such public datasets are very rare in the speech domain, and often limited to captions [6, 15]. This lack of data is particularly apparent when considering the challenge of building open-source multimodal models able to naturally talk about an image as well as other general topics, even though such models are starting to appear in the commercial space [12, 31].

In this work, we aim to effectively integrate vision capabilities into a conversational speech LLM. As our backbone, we use Moshi [8], a recent open-weight speech LLM able to dialogue with the user in real time and in full-duplex, *i.e.* it is able to listen and speak at any time and does not need to be signalled when to talk. Drawing inspiration from VLMs, we aim to adapt Moshi into a **Vision-Speech Model** (VSM) with the same dialoguing abilities. We identify three key challenges specific to building a VSM able to hold natural conversations about visual inputs: (i) overcome the above-mentioned scarcity of image-aligned speech data, and avoid blowing up the complexity of the training pipeline as we are now dealing with three modalities—vision, language and audio, (ii) comply with compute and memory constraints in order to hold real-time conversations at inference, and (iii) maintain the original conversational abilities of the backbone dialogue model, *i.e.*, preserve audio quality as well as prosodic features, and enable seamless switching between image-related and general conversation topics.

We address each of these challenges as follows: **First**, we show that we can adapt the underlying pretrained speech transformer to image inputs using image-text datasets without audio supervision (“speechless” datasets), in combination with a small percentage of speech samples. Specifically, we exploit the fact that Moshi *jointly* predicts text

\* Equal contribution.

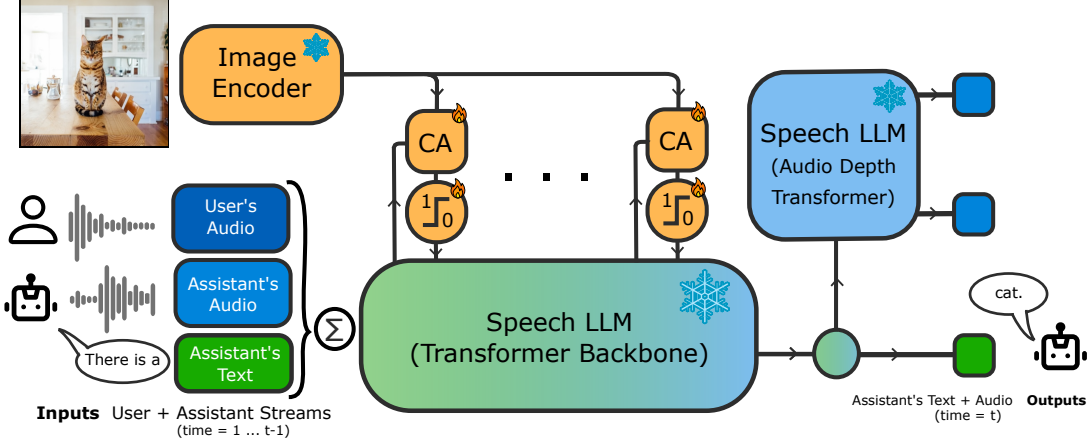


Figure 1. **MoshiVis** is a Vision-Speech model (VSM) able to hold full-duplex real-time conversations about an image, and trained with a light data- and compute- budget. For image representations, we use off-the-shelf transformer-based image encoders from the PaliGemma family [5]. For the speech modelling part, we rely on Moshi [8], a recent speech LLM which *jointly* outputs text and audio tokens in real-time, allowing for full-duplex conversations. At its core, Moshi consists of a standard 7B decoder-only transformer taking as inputs *speech tokens* (which are the sums of temporally aligned *text* tokens and *audio* tokens extracted from the assistant’s and user’s streams), rather than only text like a standard LLM. The output of the transformer is then separately decoded in a text token, as well as passed through a small *depth transformer* which auto-regressively produces a hierarchy of audio codebooks, then decoded into audio frames. First, (Sec. 3.1), we detail how we augment the speech LLM’s transformer with lightweight visual adaptation modules through cross-attention (CA). We then describe our one-stage finetuning pipeline for these modules: We use a mixture of (i) (Sec. 3.2) image+text only data (“*speechless*” data), which, despite incurring a distribution shift due to the lack of audio supervision, allows us to leverage the large body of existing Vision-Language datasets, and (ii) (Sec. 3.3) synthetic spoken visual dialogues which we design to mimic realistic discussions about images.

and audio tokens in a temporally aligned manner. While the produced text tokens differ in distribution from standard language, we find that this form of weak supervision still allows for information transfer from the text to speech modality, despite being out-of-distribution for the backbone model. **Second**, inspired by recent work on perceptual augmentations [34, 38], we inject the visual inputs into the speech LLM backbone through lightweight adaptation modules based on cross-attention. At inference, the keys and values of the attention mechanism can be efficiently cached and thus only need to be computed once for every image. **Third**, to ensure that the base model is still able to discuss general conversation topics other than the input image, we design these cross-attention layers to be able to selectively gate visual inputs based on the conversational context. In addition, expanding on previous work on visual dialogues [7, 40], we design a fully synthetic data pipeline to generate short realistic conversations about images, allowing the model to go beyond the usual setting of “one question - one answer” assumed in most VLM benchmarks.

An overview of our proposed model, **MoshiVis** is given in Figure 1: We augment a pretrained spoken dialogue model, Moshi, with lightweight adaption modules and a simple training pipeline leveraging both image-text and image-speech data. To assess the model’s visual and conversational abilities, we first evaluate MoshiVis’s visual understanding on downstream tasks commonly used in the VLM literature, such as captioning or visual question answering,

in *both* the text and audio realms. To foster further research on Vision-Speech models, we release the audio datasets we use for benchmarking. We then evaluate the model’s ability to switch contexts by measuring how its text reasoning and visual understanding abilities are affected when adding an irrelevant conversation snippet as a prefixing context. Finally, we provide qualitative samples of dialogues with MoshiVis to highlight its conversation abilities and low latency: For instance, on a L4 GPU, we find that MoshiVis only increases latency by 7ms per inference step compared to the base model Moshi, preserving real-time interaction.

**In short**, our contributions are: (i) a simple one-stage training recipe that leverages “speechless” data in combination with speech samples, tapping into the large amounts of pre-existing vision-language datasets; (ii) a lightweight gating mechanism to facilitate context switches in conversations, in particular between image-related and non-relevant content; (iii) a synthetic data pipeline for generating realistic visual dialogues. To facilitate reproducibility, we release our inference code as well as the image-speech benchmarks used to evaluate the model in audio form.

## 2. Related Work

**Vision Language Models.** VLMs transfer to visual inputs the strong reasoning abilities of LLMs, to achieve complex visual understanding [2, 4, 5, 9, 20, 36, 44]. By combining an LLM and an image encoder, they show re-

markable results on various visual understanding tasks, such as captioning, question answering or optical character recognition. While early-stage joint pre-training is a popular technique to train VLMs [2, 4, 5, 9, 44], it generally requires large amounts of image-text data and well-tuned multi-stage training pipelines. Expanding this approach to the additional audio modality, as done in [42] for instance, comes with costly training data and compute requirements. Instead, we draw inspiration from VLM “perceptual augmentations”, which have proven data- and parameter-efficient while still achieving strong visual capabilities [3, 25, 34, 38]. Such methods typically first project the image tokens to a more amenable embedding space, then inject these tokens in the text token flow via prefixing or cross-attention. Similarly, in this work, we introduce adaption modules based on gated cross-attention to adapt a speech model into a VSM. This choice is primarily motivated for practical reasons: Direct insertion of the image tokens effectively takes space in the context window of the model, thus limiting the length of the conversation which is often bottlenecked by the size of the KV cache at inference.

**Speech Modelling and Visual Inputs.** A straightforward way to augment a VLM with speech is to use ad-hoc text-speech conversion: an input module transcribing the input speech, and an output text-to-speech (TTS) module producing speech from the LLM outputs [17]. However, it is well known in the speech modelling literature [8, 10, 29] that this cascaded setup has severe flaws: it causes noticeable latency, loses prosodic information such as the user’s tone or emotion because of the input speech transcription, and it imposes separated speaker turns. Another alternative would be to include the audio modality directly at the pretraining stage of a VLM, building for instance on image-speech encoders such as SpeechCLIP [33]. While this joint training approach is being successfully explored in projects such as Mini-Omni2 [42] or AnyGPT [43], aiming to reproduce the abilities of closed commercial multimodal assistants such as GPT-4o, it requires a carefully crafted multi-stage training pipeline and datasets selection to balance all modalities across the stages. Instead, in this work, we leverage a pre-trained speech LLM, *i.e.*, a voice model with strong built-in conversational abilities, and we expand on VLM perceptual augmentation techniques to propose a simple training pipeline for turning the speech model into a VSM. Specifically, we rely on Moshi [8], a recent open-weight Speech LLM which *jointly* produces text and audio. As we will show, the presence of this text stream, despite having a distribution different from standard text, provides a strong basis for leveraging VLM techniques and thus to adapt Moshi into a VSM. Moreover, as Moshi was designed as a real-time conversational model, it proves itself to be a good starting point for designing a real-time conversational VSM.

**Towards Multimodal Dialogue Models.** Extending multimodal models from the standard “one question - one answer” paradigm to more natural multi-turn conversations is a challenging task, both from the training and evaluation perspective. In the language domain, early work on visual dialogues [7, 40] introduced the task of answering a sequence of around 10 questions about an image. Expanding on this, our work also aims to further explore how one can go from a VSM to a model able to dialogue about an image at will. In particular, we investigate the model’s ability to switch context between image-relevant and more general conversation topics, inspired from task switching analysis in LLMs [14]. We also design a synthetic data pipeline modelling realistic conversations about images (*e.g.* questions with different levels of details, misleading questions, *etc.*).

### 3. Design and Training of MoshiVis

We now describe how we augment a speech LLM such as Moshi [8] to handle visual inputs, while maintaining its conversational capabilities and real-time latency. In [Section 3.1](#), we describe how we inject visual information into the stream of speech tokens, as shown in [Figure 1](#). In [Section 3.2](#), we discuss how we leverage standard image-text data, allowing us to directly tap into the large body of vision-language understanding datasets, instead of having to collect large amounts of dedicated image-speech data. Nevertheless, for training a visual *conversational* model we still lack adequate, freely accessible dialogue datasets. To remedy this, we introduce in [Section 3.3](#) a synthetic data generation pipeline for producing spoken visual dialogues.

#### 3.1. Image-Speech Adaptation

As the core backbone of the proposed architecture, we use Moshi [8], a recent end-to-end speech dialogue model which jointly predicts text and audio tokens in real time. Our aim is to augment this backbone to interpret visual inputs given by a pretrained image encoder, such that it preserves the low inference latency required for real-time conversations, and while keeping a reasonable training budget. Our proposed pipeline is agnostic to the specific choice of the image encoder, as long as it outputs a tokenized representation of the image. In practice, we use off-the-shelf state-of-the-art image encoders from PaliGemma [5].

**Preliminaries: Speech LLMs.** As shown in [Figure 1](#), during its forward pass, Moshi first encodes the input dialogue into multiple temporally-aligned streams of tokens: A text stream capturing the assistant’s speech content, and two audio streams, one for the assistant and the user respectively. These tokens are then summed to form a single stream of tokens, the *speech tokens*, which are then fed to a transformer. The output sequence of speech embeddings are finally decoded back into separate text and audio tokens with a lightweight *depth transformer*. For further details

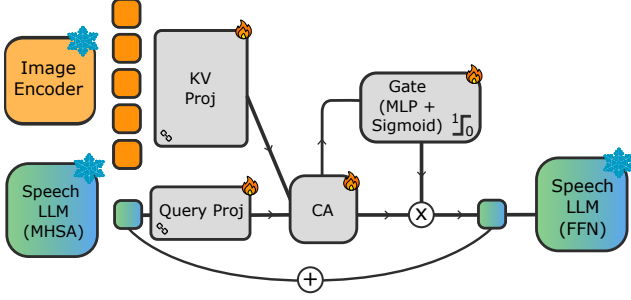


Figure 2. **Adaptation modules.** The image tokens are injected into the current speech token via residual cross-attention (CA) layers, placed between the multi-head self attention (MHSA) and the feedforward network (FFN) in every transformer block. As the cross-attention’s QKV projections are shared across layers ( $\infty$ ), at inference, we only need to compute the keys and values once per image, thus reducing the memory cost needed to store the image embeddings. To enable more context switch, we modulate the output of the cross-attention with a binary gate. The resulting output is fed back into the speech token stream as a residual.

about Moshi, in particular on audio processing, please refer to the original paper [8]. Importantly, while speech tokens contain information from the text stream, their distribution differ from standard text used in language modelling: (i) they are summed with audio tokens hence contain additional non-semantic acoustic information, and (ii) the underlying text stream processed by Moshi contains many additional padding tokens to preserve the temporal alignment between text and speech. Nevertheless, the core backbone of Moshi can be seen as a standard transformer acting on *speech* tokens, which we aim to further augment to be able to process visual inputs.

**Gated Cross-Attention.** To fuse image information into the stream of speech tokens, we introduce a cross-attention layer in each transformer block, as illustrated in Figure 2. The cross-attention takes as queries the tokens output by the self-attention layer, and uses the image embeddings as keys and values. The output is then used to compute a residual update of the speech tokens. However, introducing this additional source of information may be detrimental to the model’s initial conversational abilities, in particular its ability to switch context (see ablation experiments in Section 4.2). To promote context switching abilities, we further modulate the output of the cross-attention module with a self-gating mechanism. Intuitively, a gate output of zero would turn off the image information and exactly recover the base model behaviour, while higher values facilitate the flow of image information. Specifically, the gate is a small 2-layer MLP with a hidden size reduction factor of 1/8, followed by a sigmoid activation. During training, we do not supervise the gate’s outputs and instead let it implicitly learn an image relevance score for each token.

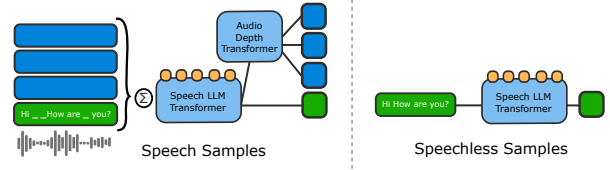


Figure 3. **MoshiVis forward pass during mixed data training.** *Speech samples* are composed of the user’s and assistant’s audio streams (■) and a text stream (■) (only for the assistant) containing extra padding tokens (.) to maintain the temporal alignment with speech. The input streams are summed and passed to the transformer. The output audio streams are auto-regressively decoded by a small transformer (Audio Depth Transformer). In practice, *we only train the first two audio streams for speech samples*. This allows for faster training as we need fewer parallel calls to the depth transformer. In contrast, *speechless samples* only contain standard text; in this case, MoshiVis acts as a standard transformer augmented with additional adaptation modules (■).

**Compute Efficiency.** At inference, as the image tokens are independent of the speech tokens, we can precompute and cache their KV projections once at the beginning of the temporal stream. In addition, we use the same cross-attention QKV projection weights in every layer of the transformer, which lessens the memory cost of the aforementioned cache, and we find that it does not significantly impact performance (see Section 4.1). As for training cost, note that in all of our experiments we keep the weights of the image embedder and the speech transformer frozen. We find that this has two positive effects: (i) It enables a lightweight training pipeline that is accessible to a wider audience for downstream task tuning; in total, we only train the adaptation modules which amounts to a total of 206M trainable parameters, and (ii) it avoids a potential degradation of the backbone speech transformer’s ability to converse about general topics other than the input image. In particular, our model exactly recovers the backbone speech transformer when the gates’ outputs are all zeros.

### 3.2. Leveraging “Speechless” Datasets for Training

While image-text datasets have flourished in recent years, equivalent datasets in speech form are scarcely available and mainly consist of transcripts of COCO-Captions [6, 15, 16]. In Section 3.3, we further expand this line of work by introducing a pipeline to synthetically generate realistic spoken visual dialogues. However, the cost of generating such data (and associated training time) to cover all aspects of visual understanding tasks, as well as speech properties (*e.g.*, variety of prosody, emotions, speaker interruptions, *etc.*) would quickly blow up. As an alternative, we aim to tap into existing image-text data which already covers a wide variety of visual understanding tasks. To that end, a key observation is that the backbone speech model explicitly predicts (and takes as input) a stream of text to-



kens. However, as illustrated in Figure 3, this stream of text tokens contains many occurrences of padding tokens—required to temporally align text and audio streams—thus does not follow the same distribution as standard text (as would, *e.g.*, a direct transcript of the audio). Nevertheless, we hypothesize it is possible to train our adaptation modules on “speechless” data, even though this incurs a distribution shift as (i) the model expects summed audio and text tokens, and (ii) speechless samples do not provide any supervisory signal to the audio codebooks output by the model.

To alleviate this, we train MoshiVis with mixed supervision: Each batch of data is composed of  $p_{\text{audio}}\%$  speech samples with audio streams, and  $(100 - p_{\text{audio}})\%$  speechless samples. As shown in Figure 3, speech samples are in-distribution with respect to the base Moshi model: the corresponding text stream only contains the text of the assistant, while the user speech is only present as audio (as the corresponding text would not be available at inference). In contrast, speechless samples contain the whole transcript in text (including the user’s questions)—as such, they differ significantly from the speech inputs: their stream of text tokens does not include any alignment padding tokens, the user’s input is given in text instead of audio, and finally, they do not contain any audio information. Interestingly, we nonetheless find that even a few audio samples in the batch are sufficient for the model to learn from the text-only signal while preserving coherent speech in the output, as we show later in experiments (Section 4.1). Importantly, this means we can now finetune the model on specialized downstream vision tasks using readily available image-text data, with little audio supervision.

Next, we discuss the case of visual dialogue datasets, which are scarcely available in text and inexistent in speech.

### 3.3. Generating Synthetic Visual Dialogues

Visual dialogue datasets only exist in text form [7, 40] and often consist of fixed-length sequences of short question-answer pairs. To promote more natural conversational flow, we design a synthetic data pipeline for spoken visual dialogues, which we use to train the final dialogue model.

**Spoken Visual Dialogue Generation.** Our first step is to generate realistic conversation about images in text form. For this, we prompt two separate Mistral-Nemo [28] models in turns, one with the goal of asking questions (the *user*) the other to answer them (the *assistant*). Both LLMs are also fed with the same text caption of an image to use as support for their respective roles. To start off the dialogue, we prompt the user to ask a general question about the image (*e.g.* “*what’s in the image?*”) and for the assistant to give a global description in a few sentences. The initial question prompt is also designed to broadly cover different question lengths, conversation tones and vocabulary. The models then continue the dialogue for 8 to 16 turns (a turn

being a question-answer pair), while being prompted with a randomly selected instruction at each turn. We design several instructions, each capturing a different type of conversation about an image, such as general questions about the image content, about fine-grained details (object locations and their properties), as well as misleading questions (*e.g.* about objects not present in the image).

All prompts used for data generation are given in Appendix E.1. Once the text dialogues are generated, we convert them to speech using the same text-to-speech model as in [8], ensuring a consistent assistant voice across samples.

**Data Augmentation.** To further enhance the model’s ability to switch topics during a conversation, we also generate a set of generic spoken dialogues, not related to any image, following the synthetic data procedure described in [8]. At training time, each visual dialogue has a  $p_{\text{concat}}$  chance of being concatenated on-the-fly with a prefix and suffix conversation, randomly sampled from this set of unrelated dialogues. In addition, we randomly sample and trim the length of each of the three dialogues being concatenated (*i.e.*, the prefix, suffix and visual dialogue).

## 4. Experiments

In this section, we discuss the performance of MoshiVis in practice. First, in Section 4.1 we evaluate its downstream accuracy on classical vision tasks including generic image understanding (captioning, question answering) and more specialized tasks (text reading). In particular, we evaluate each task in both the text and audio domains, and carefully analyse how the proportion of audio data available at training affects visual understanding as well as audio quality. Secondly, we address our initial target task and discuss the model’s ability to hold a spoken conversation about visual inputs. In Section 4.2, we measure the model’s ability to switch between different contexts in a single conversation, *i.e.*, going from talking about the image to an entirely different topic, and vice-versa, and how this behaviour is affected by the gating mechanism. Finally, in Section 4.3, we discuss practical usage of MoshiVis “in-the-wild”, such as real-time inference latencies and qualitative samples.

### 4.1. Vision-Speech Benchmark

As discussed in Section 3.2, we train MoshiVis with mixed data, each batch containing a proportion  $p_{\text{audio}}$  of speech samples. In this section, we assess whether training the model with speechless data still translates to actual vision understanding when queried via audio/speech, and, in particular, how  $p_{\text{audio}}$  affects this performance. Note that in this section we do not use any of the synthetic visual dialogues introduced in Section 3.3. Instead, we focus on downstream performance on specific vision-language benchmarks.

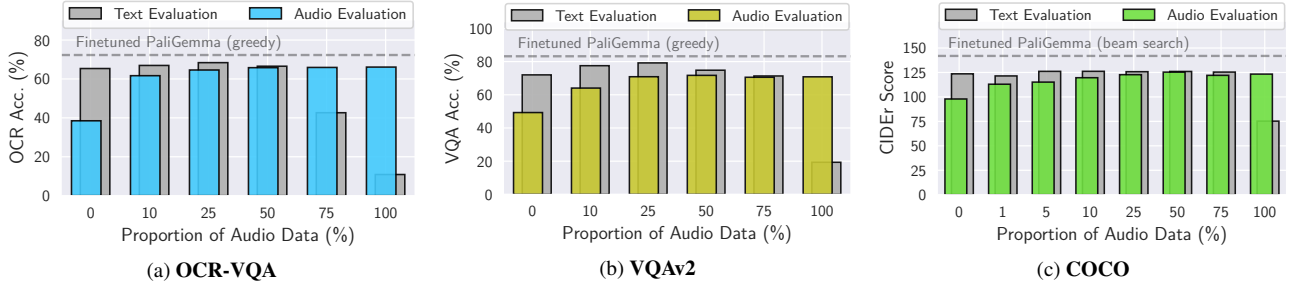


Figure 4. **Training Moshivis with different amounts of audio data** on a) OCR-VQA, b) VQAv2, and c) COCO. In particular, we show the scores obtained by the model when prompting it either with text (■) or audio (■) and using greedy decoding. Note that even when training *with no audio data at all*, the cross-attention mechanism enables the speech model to obtain results substantially above chance on all benchmarks. While this can come at the cost of audio quality, we find that adding as little as 1% of audio data effectively recovers the base model’s audio quality (Table 1). For reference, we also report the results of the fine-tuned PaliGemma (stage 3 of [5]), as we use the same image encoder as a starting point for fine-tuning; *note that in contrast to PaliGemma, we keep the image encoder and LLM frozen*.

**Experimental Setup.** As backbones, we use Moshi [8] for the speech modality ( $\sim 7B$  parameters), and the “stage 2” vision encoder of PaliGemma [5] ( $\sim 400M$  parameters) for images. Both backbones are kept frozen during training, and we only train the adaptation modules ( $\sim 206M$  parameters). We employ benchmarks covering a range of visual understanding tasks: COCO [22] (image captioning), OCR-VQA [27] (text recognition), and VQAv2 [13] (question answering). As we primarily want to evaluate our vision-speech model in the audio domain, we also convert these datasets to speech using the same text-to-speech model as in our synthetic data generation pipeline; the resulting image-speech datasets for evaluation are available on our project page, see [github.com/kyutai-labs/moshivis](https://github.com/kyutai-labs/moshivis). Note that prompting the model with speech instead of text introduces interesting challenges, for instance due to certain benchmarks being sensitive to formatting such as text punctuations, which are not necessarily transcribed in speech data. We discuss these in more detail in Appendix B.

**Main Results.** In Figure 4, we report downstream performance after training the adaptation modules in Moshivis on three separate tasks (OCR-VQA, VQAv2, and COCO), while varying the proportion of samples with audio in the batch  $p_{\text{audio}} \in \{0\%, 10\%, 25\%, 50\%, 75\%, 100\%\}$ . We evaluate the model by prompting it in both text and audio form. In both scenarios, we directly use the text tokens generated by Moshi alongside the speech as predictions to compare to the ground-truth. As reference, we also report the results of “stage 3” PaliGemma, as reported in [5], which starts from the same vision encoder (stage 2) used as frozen backbone in Moshivis; Note, however, that in contrast to our setting, both the vision encoder and LLM are finetuned for the downstream task in stage 3 of PaliGemma.

First, we observe promising transfer from the image to audio modality even when learning only from speechless samples: when trained with  $p_{\text{audio}}=0\%$  and prompted in the audio domain, the model yields 38.5% on OCR-

VQA, 49.3% on VQAv2, and a CIDEr [39] score of 113 on COCO. Interestingly, the reverse is not true: at  $p_{\text{audio}}=100\%$ , the text evaluation performance is negatively impacted, highlighting the benefits of a mixed data supervision strategy. However, the speech produced by a model trained with  $p_{\text{audio}}=0\%$  is not coherent and also of lower audio quality when comparing to the frozen Moshi backbone, as measured by their MOSNet [24] score in Table 1; for qualitative examples, see Appendix C. Nevertheless, the same table also shows that adding even small amounts of audio quickly recovers audio quality. Moreover, as shown in Figure 4, increasing  $p_{\text{audio}}$  also benefits downstream accuracy on all tasks, reaching scores comparable to stage 3 PaliGemma [5], even when prompting the model in audio form. Across these tasks, we observe that  $p_{\text{audio}}=25\%$  generally yields the best trade-off between downstream performance and amount of speech training data required.

$p_{\text{audio}}$	0%	1%	5%	10%	Moshi
MOSNet	2.78	3.59	3.47	3.56	3.34

Table 1. **Audio quality as a function of the proportion of speech samples used in training.** For each model, we evaluate its MOSNet [24] scores on 1000 randomly generated audio samples of roughly 40 seconds. While training with *no audio* severely impacts speech quality, it quickly recovers to the same quality level as the backbone model even with just a few speech samples; for qualitative audio samples, see Appendix C.

**Text-to-Audio Transfer across Tasks.** Secondly, we investigate whether the same text-to-audio transfer is observed when we have imbalanced supervision *across tasks*: In other words, whether one modality overpowers the other in terms of knowledge transfer. In this setting, we train a model such that each batch has a proportion  $p_{\text{coco}}\%$  of *speech* samples from COCO, and  $(100 - p_{\text{coco}})\%$  of *speechless* samples from OCR-VQA, while varying the ra-

tio of COCO to OCR-VQA samples ( $p_{\text{coco}}$ ). We then run the reverse experiment (*i.e.* all OCR-VQA samples are only seen in speech form, and all COCO samples are speechless). We report the audio evaluation scores (CIDEr and accuracy) for all models on COCO and OCR-VQA in Table 2: Increasing a dataset’s training data ratio in the audio domain generally has a stronger positive effect than doing so in the text domain. This is particularly visible when evaluating in audio mode (last three rows), but also noticeable in textual evaluation (first two rows). In addition, this phenomenon is more salient on the specialized OCR-VQA task: With 75% of “speechless” OCR-VQA training data (first column), the model reaches an accuracy of 36.8%. In contrast, when the same amount of OCR-VQA training data is only seen in audio form (second column), the final accuracy is 66.1%.

		More COCO samples					
$p_{\text{coco}}$ (%)		25%		50%		75%	
↓ given as		audio	text	audio	text	audio	text
text	COCO	109	121	107	120	98	119
	OCR	67.1	48.5	65.6	49.8	63.0	46.9
audio	COCO	115	90	123	93	121	100
	OCR	36.8	66.1	37.0	65.4	29.6	61.3
	MOSNet	3.45	3.49	3.38	3.50	3.47	3.30
		More OCR-VQA samples					

Table 2. **Text-to-Audio transfer with task imbalance.** We vary the audio-to-text proportion  $p_{\text{coco}}$  in multi-task training, s.t. each task appears only in a single modality (*e.g.*, COCO as audio, OCR as text; and vice-versa). We report the CIDEr score for COCO and accuracy for OCR. We observe that task knowledge transfers better through audio than through text. This effect is more striking when querying the model in audio at evaluation, and also more visible on the specialized task of OCR compared to COCO captioning.

To verify how much audio is needed to recover performance, we then vary both the COCO to OCR-VQA ratio ( $p_{\text{coco}}$ ) and the global ratio of audio samples ( $p_{\text{audio}}$ ), such that we have a percentage  $p_{\text{audio}} \times (100 - p_{\text{coco}})$  % of spoken OCR-VQA samples in each training batch. As seen in Table 3, this immediately boosts downstream performance: 10% of spoken OCR-VQA samples yields an accuracy of 60.7% as compared to the previous score of 36.8% when no audio samples were present. In summary, the insights from the previous single-task analysis generalizes to this two-task scenario: Mixing speech and speechless samples in every training batch is beneficial for downstream performance in both text and audio evaluation, and a ratio of  $p_{\text{audio}} = 25\%$  appears to be a good trade-off between performance and amount of training speech data needed.

**Ablations: Shared layers and Gating.** While the gating mechanism described in Section 3.2 is primarily introduced as a way to facilitate context switch (Section 4.2), we first

$p_{\text{coco}}$ (%)		25%			50%			75%		
global $p_{\text{audio}}$		10%	25%	50%	10%	25%	50%	10%	25%	50%
text	COCO	125	123	125	126	127	127	126	128	125
	OCR	68.4	68.5	68.6	66.4	66.3	66.5	63.9	63.9	63.3
audio	COCO	117	117	120	119	123	123	118	122	122
	OCR	60.7	65.4	66.1	58.4	63.2	64.2	54.8	57.8	60.7

Table 3. **Varying the task and speech samples proportion.** As for the single-task results (Figure 4), adding small amounts of audio data quickly boosts performance in both text and audio evaluation in the two-task setup: For instance, for OCR-VQA, 10% of training audio samples yields 60.7% accuracy, against 36.8% when no audio samples is present for the same task ratio (Table 2).

verify whether it affects the model’s performance. To this end, we perform an ablation experiment in which we vary over (i) whether the adaption modules have a gating mechanism, (ii) whether the gate parameters are shared across layers, and (iii) whether the QKV projections of cross-attention layers are shared across layers, or only the KV projections; Note that for (iii), even when the projection parameters are shared, the input normalization layers to the cross-attention never are. We report the results in Table 4 for OCR-VQA and in Appendix D.1 for COCO. Overall, all settings perform similarly and there is no clear winning trend across all evaluation benchmarks. In other words, the model is robust to design choices regarding the gate and sharing of adaptation parameters when it comes to downstream task performance alone. All other results reported in the paper use no parameter sharing in the gating modules, and full parameter sharing (QKV) across layers for cross-attention. In the next section, we further investigate the impact of the gating on the model’s context switching abilities.

Sharing	text eval.			audio eval.		
	none	KV	QKV	none	KV	QKV
↓ Gate / CA →						
none	66.1	-	-	63.7	-	-
not shared	-	67.7	68.2	-	66.2	64.7
shared	-	67.5	66.1	-	64.7	65.2

Table 4. **Ablation on the gate and shared parameters** in the cross-attention (CA) module for OCR-VQA; for COCO results, see Appendix D.1. Specifically, we evaluate different gatings (rows) and parameter sharing configurations for the CA module (columns). Overall, there is no clear winning trend across all evaluation benchmarks: The model is robust to design choices regarding the gate and sharing of adaptation parameters when it comes to downstream task performance alone. In Section 4.2, we further investigate the impact of the gate on context switching.

## 4.2. Evaluating Robustness to Context Switches

In this section and the next, we investigate the performance of MoshiVis as a dialogue model. First, we quantitatively assess the model’s robustness when switching between a

topic relevant to the input image and a non-relevant one. In particular, we investigate how this behaviour is affected by the gating mechanism introduced in Section 3.2.

**Experimental Setup.** For this section and the next (Sec. 4.3), we train MoshiVis as a visual dialogue model on a mix of datasets summarized in Appendix E.2, including (i) spoken visual dialogues, (ii) speechless visual dialogues, and (iii) speechless data on specialized tasks. For (i), we generate a first set of high-quality visual dialogues for which we use human-annotated captions from the PixMo [9] and DOCCI [30] datasets in the instruct prompt of the data generation pipeline described in Section 3.3. For (ii), we generate similar dialogues but using captions from the Pixel-Prose dataset [35]: As these captions were generated by a VLM, they tend to contain more biases and hallucinations, hence the distinction from PixMo and DOCCI. Finally (iii) is composed of publicly available benchmarks, in their original textual form, with a focus on counting and OCR tasks.

**Quantitative Evaluation.** We attempt to evaluate the robustness to context switches in a controlled, although artificial, setting: we evaluate the model’s performance on downstream tasks when presented with different irrelevant prefixes in its context. More specifically, we first evaluate the “visual to non-visual” ( $V \rightarrow NV$ ) switch by measuring the model’s MMLU performance relative to that of the Moshi backbone after seeing a conversation about an image. To mimic this past conversation, we prefix the MMLU question with a random image-relevant conversation of varying length, generated with our visual dialogue data pipeline. Similarly, for the reverse “non-visual to visual” switch ( $NV \rightarrow V$ ), we evaluate the model’s visual performance on COCO, with a random prefixed non-image related conversation, generated in the same way as the data augmentation described at the end of Section 3.3.

We perform these experiments for different values of  $p_{\text{concat}}$ , which is the probability of prefixing/suffixing image dialogues with irrelevant conversations during training, and with different architecture choices: (i) no gating, (ii) with the gating mechanism introduced in Section 3.2, and (iii) with the gating parameters shared across all layers. We report the relative performance of all three configurations in Figure 5 for both the “ $V \rightarrow NV$ ” and the “ $NV \rightarrow V$ ” settings. First, we notice that having training samples concatenated with non-image relevant prefix/suffix conversations ( $p_{\text{concat}} > 0$ ) is beneficial to context switch robustness, in particular when there is no gating mechanism. It sometimes even leads to improvement when the prefix length increases, as this setting is now in-distribution respective to training. Similarly, the introduction of the gating mechanism improves robustness, particularly when  $p_{\text{concat}} = 0$ , but also interacts well with higher values of  $p_{\text{concat}}$ . Interestingly, sharing the parameters of the gate across layers sometimes even outperforms the per-layer gating model, leading

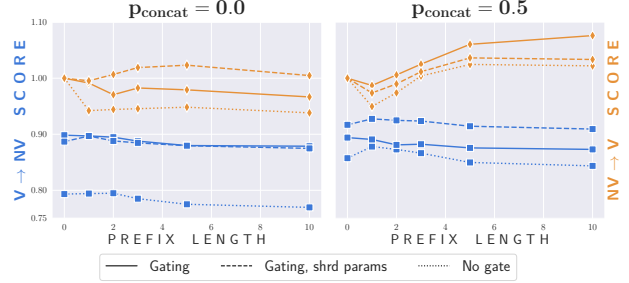


Figure 5. **Context Switch Ablation.** To assess the impact of data augmentation (left vs. right) as well as the gating configuration (different line styles), we prefix every MMLU question with a randomly sampled conversation about an image ( $V \rightarrow NV$ ), and every COCO captioning question with a randomly sampled general discussion ( $NV \rightarrow V$ ). We report the model’s relative performance as a function of the random prefix length’s (expressed in number of question-answer turns). We find that both data augmentation and gating improve the model’s robustness to context switching.

to a more parameter-efficient solution. Finally, we provide qualitative examples of patterns learned by the gating mechanism during context switch, see Appendix C.

### 4.3. MoshiVis in-the-wild

We now briefly discuss the qualitative behaviour of MoshiVis as a visual dialogue model. The corresponding inference code and model weights are available at [github.com/kyutai-labs/moshivis](https://github.com/kyutai-labs/moshivis).

**Latency.** To deploy the model, we augment the Rust and MLX backends of the open-source release of Moshi [8] with our gated adaptation modules. On an NVIDIA L4 GPU, for images of 448 pixels (1024 tokens) and an 8-bit quantized model, MoshiVis requires roughly 7 extra milliseconds of runtime per inference step compared to the backbone model, for a total of 51ms per step at the beginning of the conversation and 59ms with a 5-minute context window. We observe similar latency comparisons when testing the MLX backend on a Mac Mini with an M4 pro chip (see Appendix D.2). In both settings, the model is well within the 80ms threshold for real-time latency (the audio codec having a frequency of 12.5Hz). As for training time, our visual dialogue models are trained for 50k steps with batch size 64, taking roughly one day of training on  $8 \times H100$  GPUs.

**Qualitative Results.** Along with this work, we provide various qualitative samples to show specific behaviours of MoshiVis. For more information, please see Appendix C.

## 5. Conclusions

Combining the three image, text and audio modalities in a unified visual speech dialogue model is a challenging problem. Current solutions in the open-source space are scarce and often focus on joint pre-training strategies and data se-



lection for training such models which can be difficult to reproduce. In this work, we instead focus on lightweight finetuning, combining recent approaches in speech dialogue models and vision-language perceptual augmentation techniques. At training time, we leverage a mixture of speech and speechless (text-only) samples to learn the image-speech alignment with little audio supervision. An additional gating mechanism helps the model to switch context between visual and non-visual conversation topics. At inference, we first evaluate the model on downstream visual performance in both text and audio form, then train it with synthetic visual dialogues that we generate, to imbue it with the ability to freely converse about both images and more general conversation topics.

**Acknowledgements.** This project is funded by Iliad Group, CMA CGM Group and Schmidt Sciences. The authors thank Edouard Grave for his support and feedback throughout. They also thank Hervé Jégou for his feedback in the early phase of the project.

## References

- [1] Manoj Acharya, Kushal Kaffle, and Christopher Kanan. TallyQA: Answering Complex Counting Questions. In *AAAI*, 2019. 13
- [2] Praveesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, et al. Pixtral 12B. *arXiv:2410.07073*, 2024. 1, 2, 3
- [3] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a Visual Language Model for Few-Shot Learning. *NeurIPS*, 2022. 3
- [4] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 1, 2, 3
- [5] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al. PaliGemma: A versatile 3B VLM for transfer. *arXiv:2407.07726*, 2024. 1, 2, 3, 6
- [6] Grzegorz Chrupała, Lieke Gelderloos, and A. Alishahi. Representations of language in a model of visually grounded speech signal. In *Annual Meeting of the Association for Computational Linguistics*, 2017. 1, 4
- [7] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual Dialog. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 2, 3, 5
- [8] Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. Technical report, Kyutai, 2024. 1, 2, 3, 4, 5, 6, 8, 11
- [9] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Jen Dumas, Crystal Nam, Sophie Lebrecht, Caitlin Wittliff, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and PixMo: Open Weights and Open Data for State-of-the-Art Multimodal Models. *arXiv:2409.17146*, 2024. 1, 2, 3, 8, 13
- [10] Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. LLaMA-Omni: Seamless Speech Interaction with Large Language Models. In *ICLR*, 2025. 3
- [11] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruva Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah M Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. DataComp: In search of the next generation of multimodal datasets. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2023. 1
- [12] Google DeepMind. Project Astra. <https://deepmind.google/technologies/project-astra/>, 2024. [accessed March 6th, 2025]. 1
- [13] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017. 6, 11
- [14] Akash Gupta, Ivaxi Sheth, Vyas Raina, Mark Gales, and Mario Fritz. LLM task interference: An initial study on the impact of task-switch in conversational history. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14633–14652, Miami, Florida, USA, 2024. Association for Computational Linguistics. 3
- [15] William Havard, Laurent Besacier, and Olivier Rosec. SPEECH-COCO: 600k visually grounded spoken captions aligned to MSCOCO data set. *arXiv:1707.08435*, 2017. 1, 4
- [16] Wei-Ning Hsu, David F. Harwath, Christopher Song, and James R. Glass. Text-free image-to-speech synthesis using

- learned segmental units. In *Annual Meeting of the Association for Computational Linguistics*, 2020. 4
- [17] HuggingFace. Huggingface: Creating a voice assistant. <https://huggingface.co/learn/audio-course/chapter7/voice-assistant>, 2023. 3
- [18] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A Diagram is Worth a Dozen Images. In *Eur. Conf. Comput. Vis.*, 2016. 1
- [19] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models?, 2024. 1
- [20] Tony Lee, Haoqin Tu, Chi Heem Wong, Wenhao Zheng, Yiyang Zhou, Yifan Mai, Josselin Roberts, Michihiro Yasunaga, Huaxiu Yao, Cihang Xie, et al. VHELM: A Holistic Evaluation of Vision Language Models. *NeurIPS*, 2025. 2
- [21] Xianhang Li, Haoqin Tu, Mude Hui, Zeyu Wang, Bingchen Zhao, Junfei Xiao, Sucheng Ren, Jieru Mei, Qing Liu, Huangjie Zheng, Yuyin Zhou, and Cihang Xie. What If We Recaption Billions of Web Images with LLaMA-3? *arXiv:2406.08478*, 2024. 1
- [22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. "Microsoft COCO: Common Objects in Context". In *Eur. Conf. Comput. Vis.* Springer International Publishing, 2014. 6, 11
- [23] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023. 1
- [24] Chen-Chou Lo, Szu-Wei Fu, Wen-Chin Huang, Xin Wang, Junichi Yamagishi, Yu Tsao, and Hsin-Min Wang. MOSNet: Deep Learning based Objective Assessment for Voice Conversion. In *Interspeech*, 2019. 6
- [25] Oscar Mañas, Pau Rodríguez Lopez, Saba Ahmadi, Aida Nematzadeh, Yash Goyal, and Aishwarya Agrawal. MAPL: Parameter-efficient adaptation of unimodal pre-trained models for vision-language few-shot prompting. In *European Chapter of the Association for Computational Linguistics (ACL)*, 2023. 3
- [26] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. OCR-VQA: Visual Question Answering by Reading Text in Images. In *Int. Conf. Document Analysis and Recognition*, 2019. 13
- [27] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. OCR-VQA: Visual Question Answering by Reading Text in Images. In *Int. Conf. Document Analysis and Recognition*, 2019. 6, 11
- [28] Mistral AI. Mistral Nemo. <https://mistral.ai/news/mistral-nemo>, 2024. [accessed March 6th, 2025]. 5, 12
- [29] Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R. Costa-jussa, Maha Elbayad, Sravya Popuri, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, Itai Gat, Gabriel Synnaeve, Juan Pino, Benoit Sagot, and Emmanuel Dupoux. SpiRit-LM: Interleaved Spoken and Written Language Model, 2024. 3
- [30] Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, Su Wang, and Jason Baldridge. DOCCI: Descriptions of Connected and Contrasting Images. In *ECCV*, 2024. 8, 13
- [31] OpenAI. Santa Mode & Video in Advanced Voice—12 Days of OpenAI: Day 6. <https://www.youtube.com/watch?v=NIQDnWlwYyQ>, 2024. [accessed March 6th, 2025]. 1
- [32] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *NeurIPS*, 2022. 1
- [33] Yi-Jen Shih, Hsuan-Fu Wang, Heng-Jui Chang, Layne Berry, Hung yi Lee, and David Harwath. Speechclip: Integrating speech with pre-trained vision and language model. 2022 *IEEE Spoken Language Technology Workshop (SLT)*, pages 715–722, 2022. 3
- [34] Mustafa Shukor, Corentin Dancette, and Matthieu Cord. ePALM: Efficient Perceptual Augmentation of Language Models. In *Int. Conf. Comput. Vis.*, 2023. 2, 3
- [35] Vasu Singla, Kaiyu Yue, Sukriti Paul, Reza Shirkavand, Mayuka Jayawardhana, Alireza Ganjdanesh, Heng Huang, Abhinav Bhatele, Gowthami Somepalli, and Tom Goldstein. From Pixels to Prose: A Large Dataset of Dense Image Captions. *arXiv:2406.10328*, 2024. 8, 13
- [36] Andreas Steiner, André Susano Pinto, Michael Tschannen, Daniel Keysers, Xiao Wang, Yonatan Bitton, Alexey Gritsenko, Matthias Minderer, Anthony Sherbondy, Shangbang Long, et al. PaliGemma 2: A family of versatile vlms for transfer. *arXiv:2412.03555*, 2024. 1, 2
- [37] Rubèn Tito, Dimosthenis Karatzas, and Ernest Valveny. Document Collection Visual Question Answering. In *Int. Conf. Document Analysis and Recognition*, 2021. 13
- [38] Théophane Vallaëys, Mustafa Shukor, Matthieu Cord, and Jakob Verbeek. Improved Baselines for Data-efficient Perceptual Augmentation of LLMs. *arXiv:2403.13499*, 2024. 2, 3
- [39] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based Image Description Evaluation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015. 6, 11, 12
- [40] Bingbing Wen, Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Bill Howe, and Lijuan Wang. Infovidial: An informative visual dialogue dataset by bridging large multimodal and language models. *arXiv:2312.13503*, 2023. 2, 3, 5
- [41] Chris Wendler. Renderedtext. <https://huggingface.co/datasets/wendlerc/RenderedText>, 2023. 13
- [42] Zhifei Xie and Changqiao Wu. Mini-Omni2: Towards Open-source GPT-4o with Vision, Speech and Duplex Capabilities. *ArXiv*, abs/2410.11190, 2024. 3
- [43] Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. AnyGPT: Unified Multimodal LLM with Discrete Sequence Modeling. *arXiv preprint arXiv:2402.12226*, 2024. 3

- [44] Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Wenwei Zhang, Hang Yan, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *arXiv preprint arXiv:2309.15112*, 2023. 1, 2, 3

## A. Benchmark datasets

For benchmarking the visual understanding of our trained models, we use the following classical benchmarks.

**Optical Character Recognition (OCR).** We evaluate the model’s ability to recognize text in images on the OCR-VQA [27] dataset. We report the accuracy as a metric.

**Visual Question Answering (VQA).** We evaluate the model’s ability to answer general free-form questions about images on the VQAv2 [13] dataset and report the VQA accuracy as the primary metric.

**Image Captioning.** We evaluate the model’s ability to generate captions for images on the COCO Captions [22] dataset. We report the CIDEr [39] score as a metric. Specifically, we use the 2014 subset of COCO-Captions with Karpathy train/validation splits and annotations.

## B. Audio Evaluation

### B.1. Audio Benchmarks

To query the model in audio form, we convert the aforementioned three datasets to speech using the same text-to-speech model as in [8]. We use a variety of voices for the user asking the benchmark question. Note that this brings a new challenge inherent to VSMs compared to VLMs, as the model’s understanding of a question may vary based on the user’s audio volume, intonation, accent, *etc.*, thus adding another level of variation compared to textual prompting.

Note that since the frozen backbone speech model we use was initially trained as a dialogue model, we also reformat these datasets as short conversations rather than single questions. For instance, a simple COCO training caption such as “A boy holding an umbrella” is converted to a spoken dialogue with the following transcript “[Assistant] Hey, how are you doing? [User] So, what do you see in the image? [Assistant] I see a boy holding an umbrella”.

Similarly, for the validation/test splits of benchmarks, provided on our project page, we generate speech questions which we use to query the model to perform audio evaluation. For instance, for COCO, this can be a dialogue of the form “[Assistant] Hey, how are you? [User] Can you tell me what is in the image?”

#### Example 1:



MoshiVis-conversational: “two teddy bears in a store, one in a blue Hawaiian shirt with a brown ribbon, the other in a brown shirt with a blue ribbon”

MoshiVis-downstream: “Two teddy bears are on display in a store”

#### Example 2:



MoshiVis-conversational: “a close-up of a bunch of bananas, with a hand reaching in to pick one, and a blue sticker on one of them”

MoshiVis-downstream: “A bunch of bananas that are in a bin”

#### Example 3:



MoshiVis-conversational: “a young boy in a baseball uniform, mid-action, with a baseball glove on his right hand”

MoshiVis-downstream: “A young boy in a field of grass holding a catchers mitt”

Table 5. **Examples of generated COCO captions** for a conversational MoshiVis (*top rows*) and a MoshiVis directly trained for COCO captioning as a downstream task (*bottom rows*). While both models yield qualitatively accurate captions, the conversational MoshiVis tends to be more verbose by nature. This can lead to lower CIDEr scores on the COCO dataset, as the score is impacted by the length of the predicted captions.

### B.2. Formatting Challenges

We observe interesting challenges during audio evaluation of MoshiVis, stemming from the facts (i) that many text-based evaluation metrics are very sensitive to the output formatting and (ii) that making a model more conversational sometimes hurts its ability to be a good “one-shot” answerer, which is the setup of many VLM benchmarks.

For instance, OCR-VQA contains many textual signals such as punctuations for which no equivalent exists in audio; hence these may not appear in the output text stream of the model, which hurts accuracy. In addition, our synthetic visual dialogues are generated to give our conversational model a friendly and helpful personality, thus have a certain bias toward ‘yes’ answers, which can be hurtful for yes/no questions present in OCR-VQA (*e.g.*, “Is this book related to Science-Fiction?”) As a result, for comparison, our final conversational model has an OCR-VQA accuracy of 53.3% in audio form and 60 % in text form, as opposed to 66.7 % in audio form and 67.4 % in text form when MoshiVis is di-

rectly trained for downstream performance on OCR-VQA, without seeing any conversational data.

Similarly, CIDEr scores [39] on COCO strongly depend on the length of the generated captions. This often puts conversational models at a disadvantage as they tend to be more verbose and also sometimes use “filler” words (e.g., ‘hey’, ‘well’, ‘so’, *etc.*). For instance, our conversational MoshiVis typically reaches CIDEr scores of roughly 80 (as opposed to  $\sim 125$  scores when trained on COCO) due to generating much more verbose, yet qualitatively correct, descriptions, as illustrated in Table 5.

## C. Qualitative Samples and Behaviour

To further support the findings discussed in this work, we provide additional qualitative samples on our project page at [github.com/kyutai-labs/moshivis](https://github.com/kyutai-labs/moshivis).

First, as discussed in Section 4.1, we observe that the MOSNet scores for measuring audio quality of the model strongly improves when adding even a small amount of audio samples during training. As we show on the project page (in the section “Impact of Speechless Data on Audio Quality”), this can also be observed qualitatively on the generated speech samples.

Moreover, we provide various qualitative samples of real conversations with MoshiVis trained as a visual dialogue system, in order highlight specific behaviours of the model. This includes, e.g., the general ability to hold visual conversations, specific skills such as reading and counting, and the ability to switch contexts or speak in different voices.

Lastly, to better understand the gating learned during training, we also provide samples for which we visualize the aggregated (averaged across layers) per-token gating values used by the model; see also Figure 7 for an example.

## D. Additional results

### D.1. Gate Ablation

In Table 6, similar to Table 4, we report results on COCO for different configurations for the gating and sharing of parameters in the cross-attention modules. We find that the insights observed on the OCR-VQA dataset also apply to the COCO experiments. Specifically, the model’s benchmark performance is robust to these design choices and there is no clear “winning configuration”.

### D.2. Latency with MLX backends

In Figure 6, we report latency results for the MLX backend running locally on a Mac Mini with an Apple M4 pro chip. We evaluate these latencies with our model as well as the original Moshi backbone, quantized to 8 bits with a block size of 64.

Sharing ↓ Gate / CA →	text eval.			audio eval.		
	none	KV	QKV	none	KV	QKV
none	126	-	-	125	-	-
not shared	-	127	126	-	123	124
shared	-	126	124	-	124	122

Table 6. **Ablation on the gate and shared parameters on COCO.** We report CIDEr scores for different configurations of the gate and the cross-attention (CA) module. As for OCR-VQA (Section 4.1), there is no clear winning trend across all evaluation benchmarks: The model is robust to design choices regarding the gate and sharing of adaptation parameters when it comes to downstream task performance alone.

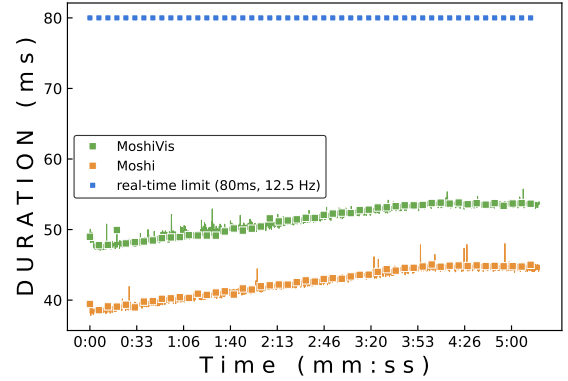


Figure 6. **Latency results with the MLX backend** on a Mac Mini with a M4 Pro chip. Here we report the latency per inference step (time to generate one speech token) for MoshiVis and the original Moshi backbone, both quantized in 8 bits. Both models stay well below the real-time limit of 80ms (12.5Hz audio codec) during a 5-minutes conversation span.

## E. Synthetic Data Generation Pipeline

### E.1. Overview

To generate the synthetic visual dialogues, we use two separate instances of Mistral’s Nemo models [28], each with its own set of instructions (‘User’ and ‘Assistant’): The user always asks questions and the assistant always answers them.

We generate a set of user-assistant instruction pairs (provided through Instructions 1 to 8), each characterising a specific behaviour or interaction. The instructions have been designed to endow the model with certain behavioural patterns, such as being robust to misleading questions (Instructions 7 and 8), or to promote learning to extract certain facts from the image embeddings such as spatial information (Instruction 2), recognising object attributes (Instruction 3), counting (Instruction 4), or to produce general question-answer conversations (Instructions 5 and 6). Finally Instruction 1 is a special instruction to generate the start of a generic visual dialogue (e.g., “what’s in the image?” in many varied ways).



**Instruction Template.** For each instruction, we provide the ‘Instruction Template’ (see, *e.g.*, [Instruction 1](#)). It is used to generate a model-specific instruction (by replacing the {ROLE\_SPECIFIC\_TEXT} with the respective texts and {caption} with the image caption). These are then provided as ‘system prompts’ (*i.e.*, in between [SYS] tags) to the Mistral Nemo models. We then force the start of the conversation by ‘Forced start of the conversation’, which triggers the first turn of the ‘User’ model—after that, the forced start is removed from the conversation history and the models ‘talk between themselves’.

**Generating dialogues.** In practice, to generate a dialogue, we can stick to a single type of instruction throughout the whole conversation (*e.g.*, for Instructions 2, 3 and 5 to 8) for multiple turns of conversation.

Alternatively, we also generate a more generic style of conversation that combines multiple instructions, see *e.g.* [Table 7](#). For this, we first start with the instruction given in [Instruction 1](#), which samples a generic question about the image (*e.g.*, “*what’s in the image?*”). After the first turn of the conversation, we then randomly sample the model instruction in each subsequent turn (question-answer pair) for the continuation of the conversation. Hence, for every conversation, the models are provided with the full history of the past conversation (excluding ‘forced start’, and exchanging the ‘system prompts’ for the randomly sampled ones) and prompted to continue the conversation.

This results in combined conversations that always start with high-level description questions (COMB, [Instruction 1](#)) then ask multiple questions about various aspects of the image, *e.g.*, location of objects (LOC, [Instruction 2](#)), their colors and properties (PROP, [Instruction 3](#)), their numbers (NUM, [Instruction 4](#)), including misleading questions (LEAD1, [Instruction 7](#); LEAD2, [Instruction 8](#)), or generic interactions between a ‘Teacher’ and a ‘Student’ (TS1, [Instruction 6](#); TS2, [Instruction 5](#)).

## E.2. Final Datasets Overview

In [Table 7](#), we describe the final datasets we use for training the conversational MoshiVis. We sample each batch such that the relative proportion of each dataset follows the distribution given by the relative weight  $\omega_i$  (third column).

The final dataset mixture is split into three categories:

- First, we generate a set of high-quality visual dialogues for which we use human-annotated captions from the PixMo [9] and DOCCI [30] datasets in the instruct prompt of the data generation pipeline described in [Section 3.3](#): These are DOCCI PROP, DOCCI LOC, PixMo LEAD1, PixMo COMB; for the detailed instructions for PROP, LOC, LEAD1 and COMB, see [Appendix F](#).
- We generate similar dialogues but using captions from the PixelProse dataset [35]: As these captions were generated

by a VLM, they tend to contain more biases and hallucinations, hence the distinction from PixMo and DOCCI. These are PixelProse TS1, PixelProse TS2, PixelProse LEAD2; for the detailed instructions for TS1, TS2 and LEAD2, see [Appendix F](#).

- Finally, we add non-dialogue style datasets in textual form to leverage publicly available image-text benchmarks, with a focus on counting and OCR tasks: TallyQA, OCR-VQA, Rendered Text, DocVQA.

Dataset Name	Source Dataset	Rel. Weight	Type
DOCCI PROP	DOCCI [30]	5	Speech
DOCCI LOC	DOCCI [30]	5	Speech
PixMo LEAD1	PixMo [9]	5	Speech
PixMo COMB	PixMo [9]	15	Speech
PixelProse TS1	PixelProse [35]	15	Text
PixelProse TS2	PixelProse [35]	15	Text
PixelProse LEAD2	PixelProse [35]	15	Text
TallyQA	TallyQA [1]	1	Text
OCR-VQA	OCR-VQA [26]	5	Text
RENDERED TEXT	RenderedText [41]	1	Text
DocVQA	DocVQA [37]	2	Text

**Table 7. Datasets used for training MoshiVis.** We list the combination of datasets we used along with the respective source datasets used to create them, the relative frequency with which we sampled them, and whether the dataset contained audio (‘Type’). In particular, the datasets were sampled with a probability given by  $p_{\text{sample}} = w_i / \sum_j w_j$ , with  $w_i$  the relative weight (‘Rel. Weight’) of each split. The respective splits were created according to the generation scripts and prompts discussed in [Appendix E.1](#).

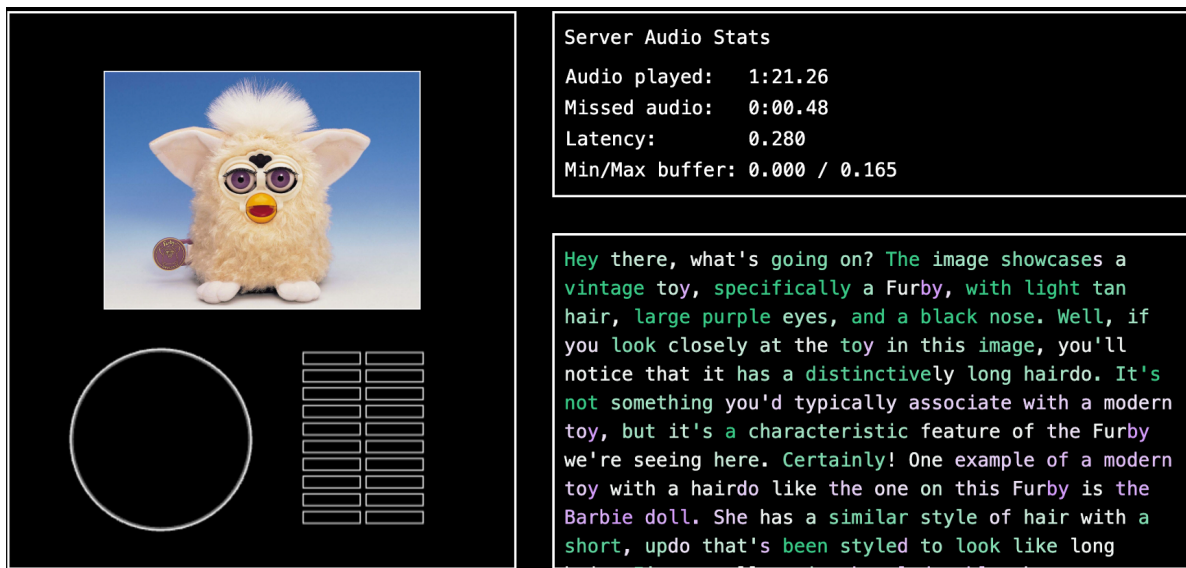


Figure 7. **Example visualization of the gate activations** during a conversation about a given input image (*left*). On the right, we see the text stream output by MoshiVis alongside the audio tokens, which only contains the assistant’s produced text tokens. We color the tokens based on the average output values of the gate sigmoid activation across all layers (**high values** in green and **lower values** in purple) during the conversation. We observe that, despite the absence of explicit supervision, the gate learns relevant patterns: It tends to activate more on image-relevant information, and less on more general knowledge questions.

## F. Detailed Instructions for Conversation Generation

In the following, we provide the detailed instructions used in our data generation pipeline, see also [Appendix E.1](#).

### Default Starting Instructions

#### Instruction Template

You take part in a casual discussion about an image.  
{ROLE\_SPECIFIC\_TEXT}

#### Role-specific text (User):

You want to learn more about the image you and the other speaker are looking at. Your aim is to obtain a description of the image.

#### Role-specific text (Assistant):

The image is described in detail by the following description:  
{caption}

You are a friendly and factual conversational assistant. Your task is to give a SHORT SUMMARY what you see in the image in A FEW sentences . You NEVER SAY HELLO NOR HI

#### Forced start of the conversation:

Start the conversation by ASKING A SINGLE question about what can be seen in the IMAGE. You use DIVERSE YET REALISTIC ways to ask your question;

```
# randomly vary over question length
if (p := random.random()) < 0.5:
    "VERY IMPORTANT: your question should be LESS THAN 8 words"
elif p < 0.75:
    "VERY IMPORTANT: your question should be LESS THAN 14 words"
else:
    "VERY IMPORTANT: your question should be LESS THAN 26 words"
# radomly vary across tone
if random.random() < 0.5:
    "You ask the question in a direct style; For instance: 'What do YOU see in the image
?' \n "
else:
    "You ask the question from your own point of view; For instance: 'What am I looking at
?' \n "

if random.random() < 0.75:
    "You speak in a confident assertive tone.\n "
else:
    "You speak in a hesitant, hard to follow, manner.\n "

# Vary point of view
if random.random() < 0.5:
    "You ask what the user SEE in the image.\n "
else:
    "You ask what's visible in the image\n "
!ALWAYS ASK A SINGLE QUESION!
```

**Instruction 1 COMB.** The default starting instructions themselves are used only to obtain a single turn conversation (user + assistant). Specifically, they are designed to obtain diverse starting points for the synthetic dialogues and, in practice, they are COMBined with other randomly sampled instructions from Instructions 2, 3 and 5 to 8 to form a multiturn conversation. Note that the `if random.random() < 0.5` instructions are not part of the prompt, but actual sampling operations are executed every time to generate the initial prompt for each new dialogue.

## Instructions for conversations about spatial information

### Instruction Template

Image description:  
""{caption}""  
{ROLE\_SPECIFIC\_TEXT}

#### Role-specific text (User):

You are engaging in a conversation about an image with another person. Your goal is to ask detailed questions about everything that is visible in the image, starting from the most salient features (main objects and their relationships) to finer details (the overall setting, background features, time of day, season, etc). To guide your questions, you have been secretly provided with a detailed description of the image (see above); this fact should not be revealed however! You will use this secret description to only ask questions that can be answered based on this description. YOU SHOULD AVOID EASY YES/NO QUESTIONS! You do not ask leading questions that already contain or give a hint at the answer; i.e., avoid ending your question in 'isn't it'/'does it'/'doesn't it' etc. In your questions, you emphasize the spatial relations / locations of what is in the image. You only ask about spatial relations explicitly known from the image description. If possible, ask spatial questions about different aspects of the image.

#### Role-specific text (Assistant):

You are a helpful conversation partner who can see the image above and is willing to describe it to another person. You provide detailed (but not too verbose!) answers about the image in response to their questions. When answering:

- Be detailed and factual, use simple language and keep the answer short. No matter what the other speaker is implying, you always base your answer on the true facts given in the image description.
- Be assertive about facts that are provided in the original description.
- Contradict the other speaker when adequate such as receiving information that contradicts the description.
- Speak naturally, as though you are sharing your genuine observations with someone looking at the image alongside you.
- Avoid any indication that you are relying on a description or external data. Do not use phrases like "I was told" or "Based on what I read."
- Engage in a dynamic conversation-answer questions about the image, offer additional observations, and encourage exploration of its details.
- Make thoughtful, plausible inferences when necessary, but always stay grounded in what is realistically observable in the image.
- For example, if asked about the mood of the image, consider elements like lighting, colors, facial expressions, or the setting to infer emotions.
- If asked about a specific detail, respond as if you are focusing on that part of the image directly.
- MOST IMPORTANTLY: You never invent any new facts! Your goal is to create an immersive and conversational experience, simulating the act of perceiving the image firsthand. Remember to NEVER make up any facts about the image, answer solely based on the description provided.

#### Forced start of the conversation:

Start the conversation by asking a question about the image in any way you want!

**Instruction 2 LOC.** To improve factuality and better extract *spatial* information from the image embeddings, we instruct the models to specifically ask questions about locations of objects and answer based only on the captions. We additionally use Instruction 3, to extract attribute information.



## Instructions for conversations about object property information

### Instruction Template

Image description:  
""{caption}""  
{ROLE\_SPECIFIC\_TEXT}

### Role-specific text (User):

You are engaging in a conversation about an image with another person. Your goal is to ask detailed questions about everything that is visible in the image, starting from the most salient features (main objects and their relationships) to finer details (the overall setting, background features, time of day, season, etc). To guide your questions, you have been secretly provided with a detailed description of the image (see above); this fact should not be revealed however! You will use this secret description to only ask questions that can be answered based on this description. YOU SHOULD AVOID EASY YES/NO QUESTIONS! You do not ask leading questions that already contain or give a hint at the answer; i.e., avoid ending your question in 'isn't it'/'does it'/'doesn't it' etc. In your questions, you focus on attributes of what is visible in the image (as given via descriptions and adjectives in the image description). This includes in particular the COLOR of object, their SHAPE or their TEXTURE. You only ask about properties explicitly known from the image description. If possible, ask questions about different aspects of the image.

### Role-specific text (Assistant):

You are a helpful conversation partner who can see the image above and is willing to describe it to another person. You provide detailed (but not too verbose!) answers about the image in response to their questions. When answering:

- Be detailed and factual, use simple language and keep the answer short. No matter what the other speaker is implying, you always base your answer on the true facts given in the image description.
- Be assertive about facts that are provided in the original description.
- Contradict the other speaker when adequate such as receiving information that contradicts the description.
- Speak naturally, as though you are sharing your genuine observations with someone looking at the image alongside you.
- Avoid any indication that you are relying on a description or external data. Do not use phrases like "I was told" or "Based on what I read."
- Engage in a dynamic conversation-answer questions about the image, offer additional observations, and encourage exploration of its details.
- Make thoughtful, plausible inferences when necessary, but always stay grounded in what is realistically observable in the image.
- For example, if asked about the mood of the image, consider elements like lighting, colors, facial expressions, or the setting to infer emotions.
- If asked about a specific detail, respond as if you are focusing on that part of the image directly.
- MOST IMPORTANTLY: You never invent any new facts! Your goal is to create an immersive and conversational experience, simulating the act of perceiving the image firsthand. Remember to NEVER make up any facts about the image, answer solely based on the description provided.

### Forced start of the conversation:

Start the conversation by asking a question about the image in any way you want!

**Instruction 3 PROP.** Similar to Instruction 2, to improve factuality and better extract *attribute* information (e.g. colours, textures, shapes) from the image embeddings, we instruct the models to specifically ask questions about such attributes of objects and to answer based only on the captions.

## Instructions for conversations about spatial information

### Instruction Template

Image description:  
""{caption}""  
{ROLE\_SPECIFIC\_TEXT}

### Role-specific text (User):

You are engaging in a conversation about an image with another person.  
Your goal is to ask detailed questions about everything that is visible in the image, starting from the most salient features (main objects and their relationships) to finer details (the overall setting, background features, time of day, season, etc).  
To guide your questions, you have been secretly provided with a detailed description of the image (see above); this fact should not be revealed however!  
You will use this secret description to only ask questions that can be answered based on this description.  
YOU SHOULD AVOID EASY YES/NO QUESTIONS! You do not ask leading questions that already contain or give a hint at the answer; i.e., avoid ending your question in 'isn't it'/'does it'/'doesn't it' etc.  
Your questions focus on the NUMBER of objects visible in the image. If possible, ask spatial questions about different objects categories in the image"

### Role-specific text (Assistant):

You are a helpful conversation partner who can see the image above and is willing to describe it to another person.  
You provide detailed (but not too verbose!) answers about the image in response to their questions.  
When answering:  
- Be detailed and factual, use simple language and keep the answer short. No matter what the other speaker is implying, you always base your answer on the true facts given in the image description.  
- Be assertive about facts that are provided in the original description.  
- Contradict the other speaker when adequate such as receiving information that contradicts the description.  
- Speak naturally, as though you are sharing your genuine observations with someone looking at the image alongside you.  
- Avoid any indication that you are relying on a description or external data. Do not use phrases like "I was told" or "Based on what I read."  
- Engage in a dynamic conversation-answer questions about the image, offer additional observations, and encourage exploration of its details.  
- Make thoughtful, plausible inferences when necessary, but always stay grounded in what is realistically observable in the image.  
- For example, if asked about the mood of the image, consider elements like lighting, colors, facial expressions, or the setting to infer emotions.  
- If asked about a specific detail, respond as if you are focusing on that part of the image directly.  
- MOST IMPORTANTLY: You never invent any new facts! Your goal is to create an immersive and conversational experience, simulating the act of perceiving the image firsthand.  
Remember to NEVER make up any facts about the image, answer solely based on the description provided.

### Forced start of the conversation:

Start the conversation by asking a question about the image in any way you want!

**Instruction 4 NUM.** To improve factuality in particular about the number of objects, we put a specific emphasis on these types of questions through this instruct. We additionally use Instruction 3, to extract attribute information and Instruction 2 for object location

## Teacher-student instructions #1

### Instruction Template

```
IMAGE DESCRIPTION START
{caption}
IMAGE DESCRIPTION END
You are an *external observer* having a casual dialogue about the image described above.
You pretend that you see the image itself, **under no circumstances** mention that you got
the information from a description!!
{ROLE_SPECIFIC_TEXT}
You sound confident and assertive and most importantly, you always stick to the facts
described!!
Again, DO NOT ADD FACTS, DO NOT MENTION THE DESCRIPTION, DO NOT MENTION THE OTHER SPEAKER's
NAME.
```

### Role-specific text (User):

You are the student!! YOU DO NOT HAVE ACCESS TO THE DESCRIPTION so you have to get all the information from your teacher. Your goal is to learn about everything about the image. You should refer to the image in your questions. e.g. 'is ... visible in the image' or 'Do you see ... in the image' or 'What is in the image?' You sometimes ask questions about something NOT VISIBLE IN THE IMAGE. In particular, you want to learn about the NUMBER of objects, their LOCATION and their COLOR. You ask ONLY ONE QUESTION AT A TIME!

### Role-specific text (Assistant):

You are the strict teacher!! Your answers should be complete and detailed, but NOT TOO LONG. Do not EVER mention the description. You are nice but firm and DO NOT HESITATE TO CORRECT THE STUDENT. You never mention any facts that are not explicitly described about the image!!! NEVER mention the atmosphere of the image, only its CONTENT

### Forced start of the conversation:

Start the conversation by asking a question about an object which is NOT mentioned in the description.

**Instruction 5 TS1.** To improve the model's robustness to all kinds of general questions about images, we designed two different sets of instructions for 'student-teacher' interactions (see also Instruction 6). Specifically, in this instruction, we instruct the student to try to learn as much as possible about the image by asking the teacher, with a particular focus on factual elements.

## Teacher-student instructions #2

### Instruction Template

IMAGE DESCRIPTION START

{caption}

IMAGE DESCRIPTION END

You are an *\*external observer\** having a casual dialogue about the image described above. You pretend that you see the image itself, *\*\*under no circumstances\*\** mention that you got the information from a description!!

{ROLE\_SPECIFIC\_TEXT}

You sound confident and assertive and most importantly, you always stick to the facts described!!

Again, DO NOT ADD FACTS, DO NOT MENTION THE DESCRIPTION, DO NOT MENTION THE OTHER SPEAKER'S NAME.

### Role-specific text (User):

You are the student!! You do not see the image very well and your goal is to ask simple (almost stupid) questions about the image to learn more about its content. You should refer to the image in your questions. e.g. 'is ... visible in the image' or 'Do you see ... in the image' or 'What is in the image?' Your questions should also details about the LOCATION of objects and a bit about their COLOR. You ask ONLY ONE QUESTION AT A TIME!

### Role-specific text (Assistant):

You are the teacher!! Your answers should be complete and detailed, and long. Do not EVER mention the description. You never mention any facts that are not explicitly described about the image!!! NEVER mention the atmosphere of the image, only its CONTENT

### Forced start of the conversation:

Start the conversation by asking a question about an object which is NOT mentioned in the description.

**Instruction 6 TS2.** To improve the model's robustness to all kinds of general questions about images, we designed two different sets of instructions for 'student-teacher' interactions (see also Instruction 5). Specifically, in this instruction, we instruct the student to ask simple ('almost stupid') questions about the image, with a particular focus on factual elements.



## Instructions to ask misleading questions #1

### Instruction Template

Image description:  
""{caption}""  
{ROLE\_SPECIFIC\_TEXT}

### Role-specific text (User):

You are engaging in a conversation about an image with another person. Your goal is to ask detailed questions about everything that is visible in the image, starting from the most salient features (main objects and their relationships) to finer details (the overall setting, background features, time of day, season, etc). To guide your questions, you have been secretly provided with a detailed description of the image (see above); this fact should not be revealed however! You will use this secret description to only ask questions that can be answered based on this description. YOU SHOULD AVOID EASY YES/NO QUESTIONS! You do not ask leading questions that already contain or give a hint at the answer; i.e., avoid ending your question in 'isn't it'/'does it'/'doesn't it' etc. In your questions, you often BUT NOT ALWAYS try to mislead the other speaker into believing something that is not correct. For instance, you ask about a RANDOM object not in the image but keep your questions short!! You should be almost rude in your questions.

### Role-specific text (Assistant):

You are a helpful conversation partner who can see the image above and is willing to describe it to another person. You provide detailed (but not too verbose!) answers about the image in response to their questions. When answering:

- Be detailed and factual, use simple language and keep the answer short. No matter what the other speaker is implying, you always base your answer on the true facts given in the image description.
- Be assertive about facts that are provided in the original description.
- Contradict the other speaker when adequate such as receiving information that contradicts the description.
- Speak naturally, as though you are sharing your genuine observations with someone looking at the image alongside you.
- Avoid any indication that you are relying on a description or external data. Do not use phrases like "I was told" or "Based on what I read."
- Engage in a dynamic conversation-answer questions about the image, offer additional observations, and encourage exploration of its details.
- Make thoughtful, plausible inferences when necessary, but always stay grounded in what is realistically observable in the image.
- For example, if asked about the mood of the image, consider elements like lighting, colors, facial expressions, or the setting to infer emotions.
- If asked about a specific detail, respond as if you are focusing on that part of the image directly.
- MOST IMPORTANTLY: You never invent any new facts! Your goal is to create an immersive and conversational experience, simulating the act of perceiving the image firsthand. Remember to NEVER make up any facts about the image, answer solely based on the description provided. Do not confirm any misleading information; if necessary, say you do not know what the other speaker means.also MAKE SURE TO USE \*DIFFERENT\* and VARIED ANSWERS: For instance: 'No', 'I can't confirm', 'I don't see', 'I'm not sure', 'You're wrong', 'Nope', 'Incorrect', 'Wrong'

### Forced start of the conversation:

Start the conversation by asking a question about the image in any way you want!

**Instruction 7 LEAD1.** To make the conversational model robust to ‘misleading questions’ by the users (e.g., “What is the chicken doing there?” when there is no chicken in the image), we instruct the LLM in the ‘user’ role to ask such questions and the ‘assistant’ LLM to stick to the provided caption.

## Instructions to ask misleading questions #2

### Instruction Template

```
IMAGE DESCRIPTION START
{caption}
IMAGE DESCRIPTION END
You are an *external observer* having a casual dialogue about the image described above.
You pretend that you see the image itself, **under no circumstances** mention that you got
the information from a description!!
{ROLE_SPECIFIC_TEXT}
You sound confident and assertive!!
Again, DO NOT ADD FACTS, DO NOT MENTION THE DESCRIPTION, DO NOT MENTION THE OTHER SPEAKER’S
NAME.
```

### Role-specific text (User):

Your goal is to mislead the other speaker. You often (!but not always!) ask whether RANDOM and DIVERSE objects are visible in the image. You should always sound very confident in your question. Your speaking style is direct, assertive, almost rude sometimes!!

### Role-specific text (Assistant):

You always give extensive and FACTUAL answers. You politely but FIRMLY CORRECT the other speaker when they are wrong!! You may also try to redirect the conversation by mentioning an object from the image. Your answers should always be factual to the description!!! Don’t hesitate to say a FIRM !!NO!! when the other speaker is rude. Do not EVER mention the description. You never mention any facts that are not explicitly described about the image!!!

### Forced start of the conversation:

Start the conversation by asking a question about an object which is NOT mentioned in the description.

**Instruction 8 LEAD2.** Similar to Instruction 7, to make the conversational model robust to ‘misleading questions’ by the users (e.g., “What is the chicken doing there?” when there is no chicken in the image), we instruct the LLM in the ‘user’ role to ask such questions and the ‘assistant’ LLM to stick to the provided caption.