Appendix for Vision-Speech Models: Teaching Speech Models to Converse about Images

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 [30] 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 [15] 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 [24] dataset. We report the CIDEr [42] as 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 an off-the-shelf text-to-speech software. 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 an additional 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 a single question. 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, 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?"*

We release the evaluation/test splits of benchmarks alongside this paper.

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 6. **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 tend to be more verbose due to its conversational 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 **(i)** the fact that many text-based evaluation metrics are very sensitive to the output formatting and **(ii)** 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 personally, 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 [42] 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, captions, as illustrated in Table 6.

B.3. Additional Comparison to Omni-Models

As noted in the previous paragraph, many VLM datasets do not lend themselves well to VSMs; *e.g.*, complex structured formats (coding, LaTeX, "`A: D, B: A, C: B. D: E`"-style multi-choice answers, *etc.*) have been explicitly designed with clear turn-based, textual interactions in mind. Adapting such datasets to full–duplex speech models requires tedious re-formatting and is often poorly handled by existing text-to-speech models. As such, evaluating VSMs and omni models is more challenging than VLMs, and current omni models often do not report quantitative results of visual question-answering based on speech inputs.

To mitigate this, we introduce our Babillage dataset and evaluate a subset of omni-models on it in Section 4.3. For the task of image captioning, we found that while open-source omni models (including MoshiVis) achieve comparable CIDEr scores, a gap to text-only VLMs remains. In the following, we additionally evaluate advanced open-source (Qwen2.5-Omni) and commercial (Gemini 2.0) omni-models in Table 7 below. For this, we again use samples of our Babillage dataset and provide a simple text prompt to the model "*Please answer the question in the audio regarding this image*". For the captioning task, as we already discuss in *Sec. 4.3*, conversational omni models tend to generate verbose image captions, which leads to poor CIDEr scores. Similarly, the OCRVQA outputs do not follow the format expected by the benchmark at all, highlighting the difficulty of evaluating dialogue models. We therefore resort to LLM-as-a-judge to evaluate the free form captions: while not directly comparable with existing text-based OCRVQA results, it allows us to compare the impact of the question modality on performance: specifically, our results clearly indicate that a gap between the text and audio capabilities of omni models remains, even for commercial models.

Benchmark	QwenOmni2.5 7B	Gemini 2.0	Gemini 2.5 Lite
COCO CIDEr (audio)	50.30	43.30	52.90
OCR LLM-as-a-judge (audio)	88.67	81.86	76.21
OCR LLM-as-a-judge (text)	90.50	85.21	85.56

Table 7. Additional omni models evaluated on Babillage.

3. Audio Samples Quality

In Section 4.1, we observe that the audio quality of the model strongly improves when adding even a small amount of audio samples during training. In addition to the MOSNet scores reported in the main paper, we illustrate this behaviour through qualitative samples. For simplicity, we provide these audio samples as an HTML file with embedded audio files. This file can be found as `app_7_audio_quality.html` in the supplementary and works best with Chrome and Firefox web browsers.

4. Qualitative Behaviour

4.1. Samples of Conversation

First, we provide qualitative samples of real conversations with MoshiVis trained as a visual dialogue systems. These samples are provided as video files in the supplementary material subdirectory `app_9_qualitative_samples`. Each video contains **(i)** the input image, **(ii)** the audio conversation, **(iii)** the text stream output by the model alongside the audio model. Finally, note that all audios in the video have been slightly altered to preserve anonymity, hence may have slightly lower quality than the original outputs.

Alongside each video, we also provide a visualization of the corresponding gate per-token activation pattern (high values in green and lower values in purple) during the conversation. These are found in the subfolder `app_9_learned_gate_patterns`, and an example is also given in Figure 7.

Through these qualitative samples, we aim to highlight the following behaviours:

- **Long conversations**, in `casual_chat.webm`
- **"Visual → Non-visual"** context switch, in `visual_to_nonvisual_switch.webm`
- **"Non-visual → Visual"** context switch, in `nonvisual_to_visual_switch.webm`
- **Multiple context switches**, in `back_and_forth.webm`
- **Text reading and counting**, in `text_reading.webm`
- **Preserving speech prosody**, in `preserving_prosody.webm`

Finally, in `app_9_bis_multi_image_switch`, we report preliminary results on extending the model to handle multiple images at inference. By nature, cross-attention allows to very easily change the input image by simply recomputing the key-value-pairs used in the cross-attention layers. The only overhead thus comes from processing the image through the image encoder. In our setting, this overhead is minimal as the image encoder is very small compared to the main transformer backbone. As shown in the qualitative samples, this overhead doesn't affect real-time

latency, and the model is able to comprehend the new image as we update the cross-attention inputs on-the-fly. In the future we are hoping to extend this setup to streaming multi-images settings, for instance for live video interaction,

4.2. Latency with MLX backends

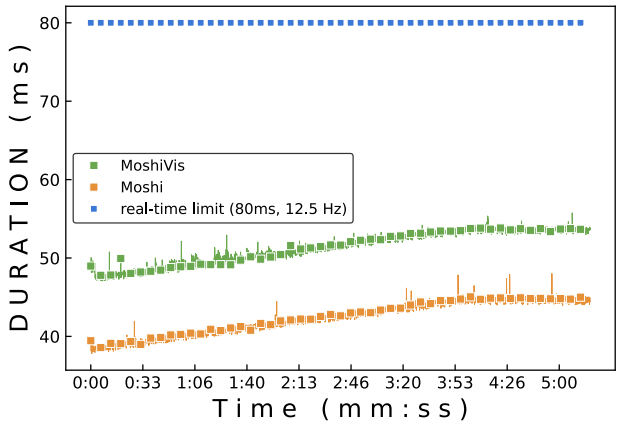

Figure 6. **Latency results with the MLX backend** on a Mac Mini with a M4 Pro chip. Here we report the latency per inference step (time to generate one speech token) for MoshiVis and the original Moshi backbone, both quantized in 8 bits. Both models stay well below the real-time limit of 80ms (12.5Hz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

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 8bits with a block size of 64.

5. Gate Ablation

In Table 8, 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".

6. Synthetic Data Generation Pipeline

6.1. Overview

To generate the synthetic visual dialogues, we use two separate instances of Mistral's Nemo models [31], 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).

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

Table 8. **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.

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, which we refer to as 'SSG' in Table 9. 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 conversations that always start with a high-level description questions (`SSG`, Instruction 1) then

asks 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), or asks misleading questions (`LEAD`, Instruction 7; `NFC`, Instruction 8), or simply generic questions (`TNS`, Instruction 6; `TBS`, Instruction 5).

6.2. Final Datasets Overview

In Table 9, 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 ω_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 [10] and DOCCI [33] datasets in the instruct prompt of the data generation pipeline described in Section 3.3: These are `DOCCI PROP`, `DOCCI LOC`, `PixMo LEAD`, `PixMo SSG`; for the detailed instructions for LEAD, SSG, PROP, and LOC, see Section 7.
- We generate similar dialogues but using captions from the PixelProse dataset [38]: 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 TNS`, `PixelProse TBS`, `PixelProse NFC`; for the detailed instructions for TNS, NFC, and TBS, see Section 7.
- 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	Relative weight w_i	Type
DOCCI PROP	DOCCI [33]	5	Speech
DOCCI LOC	DOCCI [33]	5	Speech
PixMo LEAD	PixMo [10]	5	Speech
PixMo SSG	PixMo [10]	15	Speech
PixelProse TNS	PixelProse [38]	15	Text
PixelProse TBS	PixelProse [38]	15	Text
PixelProse NFC	PixelProse [38]	15	Text
TallyQA	TallyQA [1]	1	Text
OCR-VQA	OCR-VQA [29]	5	Text
RENDERED TEXT	TODO	1	Text
DocVQA	DocVQA [40]	2	Text

Table 9. **Datasets used for training MoshiVis.** We list the combination of datasets we used along with the respective source datasets that were used to create them, the relative frequency (relative weight w_i) with which we sampled them, and whether the dataset contained audio ('Type'). In particular, the datasets were sampled with a probability given by $p_{\text{sample}} = w_i / \sum_i w_i$. The respective splits (*e.g.* TNS, LOC, LEAD) were created according to the generation scripts discussed in Appendix 6.1.

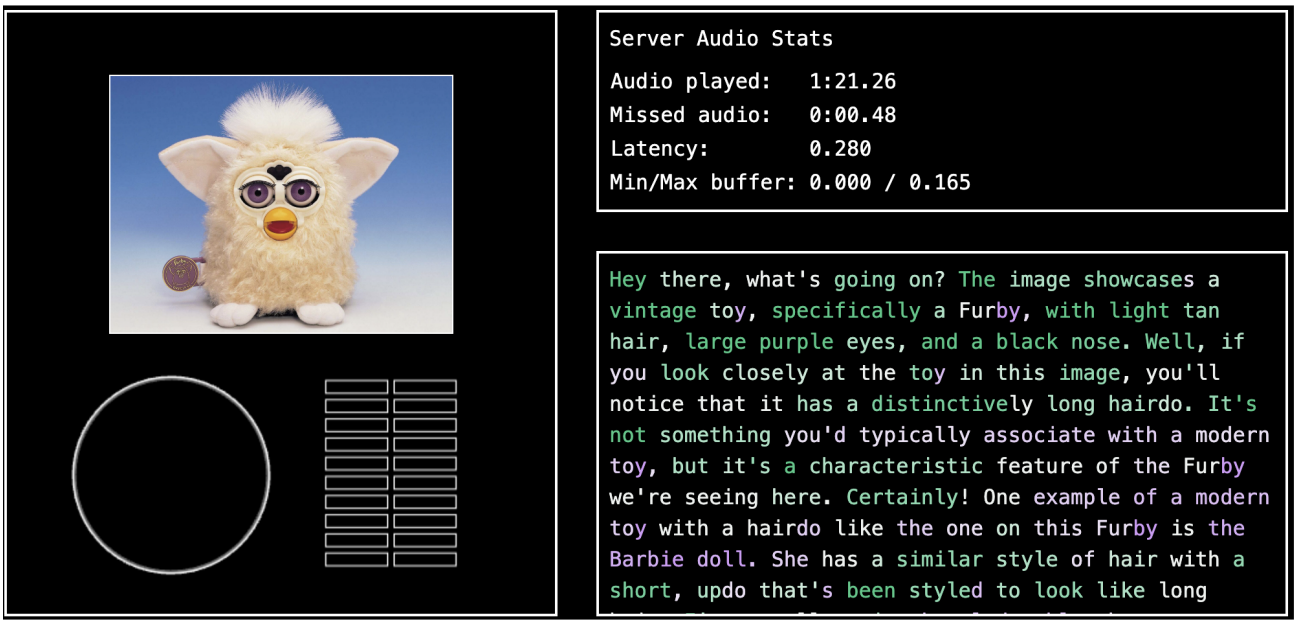

Figure 7. **Example visualization of the gate activations** during a conversation about a given input image (*left*). On the right, we see the text stream output by MoshiVis alongside the audio tokens, which only contains the assistant's produced text tokens. We color the tokens based on the average 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 no explicit supervision, the gate learns relevant patterns: It tends to activate more on image-relevant information, and less on more general knowledge questions.

7. Detailed Instructions for Conversation Generation

In the following, we provide the detailed instructions used in our data generation pipeline, see also Sec. 6.1.

Default Starting Instructions

Instruction Template

```
You take part in a casual discussion about an image.
{ROLE_SPECIFIC_TEXT}
```

Role-speficic 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-speficic 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 SSG: The default starting instructions 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-speficic 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-speficic 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-speficic 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-speficic 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-speficic 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-speficic 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-speficic 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-speficic text (Assistant):

```
You are the strict teacher!! Your anwers 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 athmosphere 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 TBS: 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-speficic 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-speficic text (Assistant):

```
You are the teacher!! Your anwers 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 athmosphere 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 TNS: 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-speficic 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-speficic 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 LEAD: 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-speficic 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-speficic 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 obejct 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 NFC: 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.