Dataset	YTD-18M	MMDialog
Number of Dialog	18M	1M
How *sensible* is the dialog?		
Natural (3)	60.1%	57.7%
Slightly Natural (2)	29.4%	32.6%
Unnatural (1)	10.5%	9.7%
Avg. Score	**2.495**	2.479
How *specific* is the dialog?		
Specific (3)	70.6%	59.7%
Slightly Specific (2)	21.3%	26.9%
Unspecific (1)	8.1%	13.3%
Avg. Score	**2.650***	2.464
Is the data containing *explicit* content?		
Sexually Explicit	**0.5%***	1.6%
Hatespeech	**0.3%***	2.5%
Others	**0.3%***	0.9%

Table 8: Full breakdown for human evaluation results on YTD-18M and MMDialog about the quality of dialogues. * denotes statistically significance after independent two-sample t-test ($p < 0.5$).

Dataset	YTD-18M	MMDialog
Is the interlocutors of dialog *visible*?		
Visible	**61.6%***	11.5%
*If **NOT** visible, then*		
is the dialog *related* to the image(s)?		
Related (3)	71.0%	67.5%
Slightly Related (2)	16.8%	22.9%
Unrelated (1)	12.2%	9.6%
Avg. Score	**2.589**	2.580
is the dialog *grounded* to the image(s)?		
Grounded (3)	62.3%	59.9%
Slightly Grounded (2)	18.2%	26.5%
Not Grounded (1)	19.4%	13.6%
Avg. Score	2.429	**2.463**

Table 9: Full breakdown for human evaluation results on YTD-18M and MMDialog about visual contexts. * denotes statistically significance after independent two-sample t-test ($p < 0.5$).

	Visual Context	#Dialog	Avg. #Turn	Avg. Utt. Length	#Tokens
BST [53]	✗	7K	11.2	13.6	1M
ConvAI2 [13]	✗	20K	13.9	9.9	2.7M
ED [44]	✗	25K	4.3	13.7	1.5M
WOW [14]	✗	22K	9.1	16.4	3.3M
WOI [27]	✗	9.5K	10.9	13.9	1.4M
SODA [25]	✗	1.5M	7.6	16.1	183M
ImageChat [50]	✓	100K	3.0	9.7	2.9M
OVD2.0 [59]	✓	116K	**48.7**	6.3	35.6M
MMD [17]	✓	1M	4.5	15.9	71.5M
YTD-18M	✓	**18M**	3.0	**19.7**	**1.06B**

Table 10: Statistics of YTD-18M compared to other open-domain dialogue and visually grounded dialogue dataset. *Utt.* stands for utterance.

A. Details of Dataset Collection

Safety Filtering. We use Rewire API [1] to filter out unsafe contents from videos. Rewire API identifies abusive, hateful, profane, violent, or sexually explicit content. However, we have discovered that the API is not accurate enough to detect profanities and violent languages in video transcripts. Thus, we only use API to detect abusive, hateful, or sexually explicit content. We set thresholds of 0.99534, 0.83790, 0.99562 to filter out unsafe contents for abuse, hate, and sexually explicit labels, respectively.

Aligning Video and Dialogue. We use Dynamic Time Warping [36] to align the dialogue (text) with the video frames. In particular, we first calculate the distance between the noisy transcript and the converted dialogue using Levenshtein distance. We then employ Dynamic Time Warping to align the words and minimize the distance between the transcript and the dialogue. Following that, using the timing information associated with the transcripts, we estimate the start time of each utterances in the dialogue. We extract the video frame using the start timing of the utterance, resulting in a video-based dialogue with video frames and the dialogue turns $(I_1, T_1, ..., I_n, T_n)$.

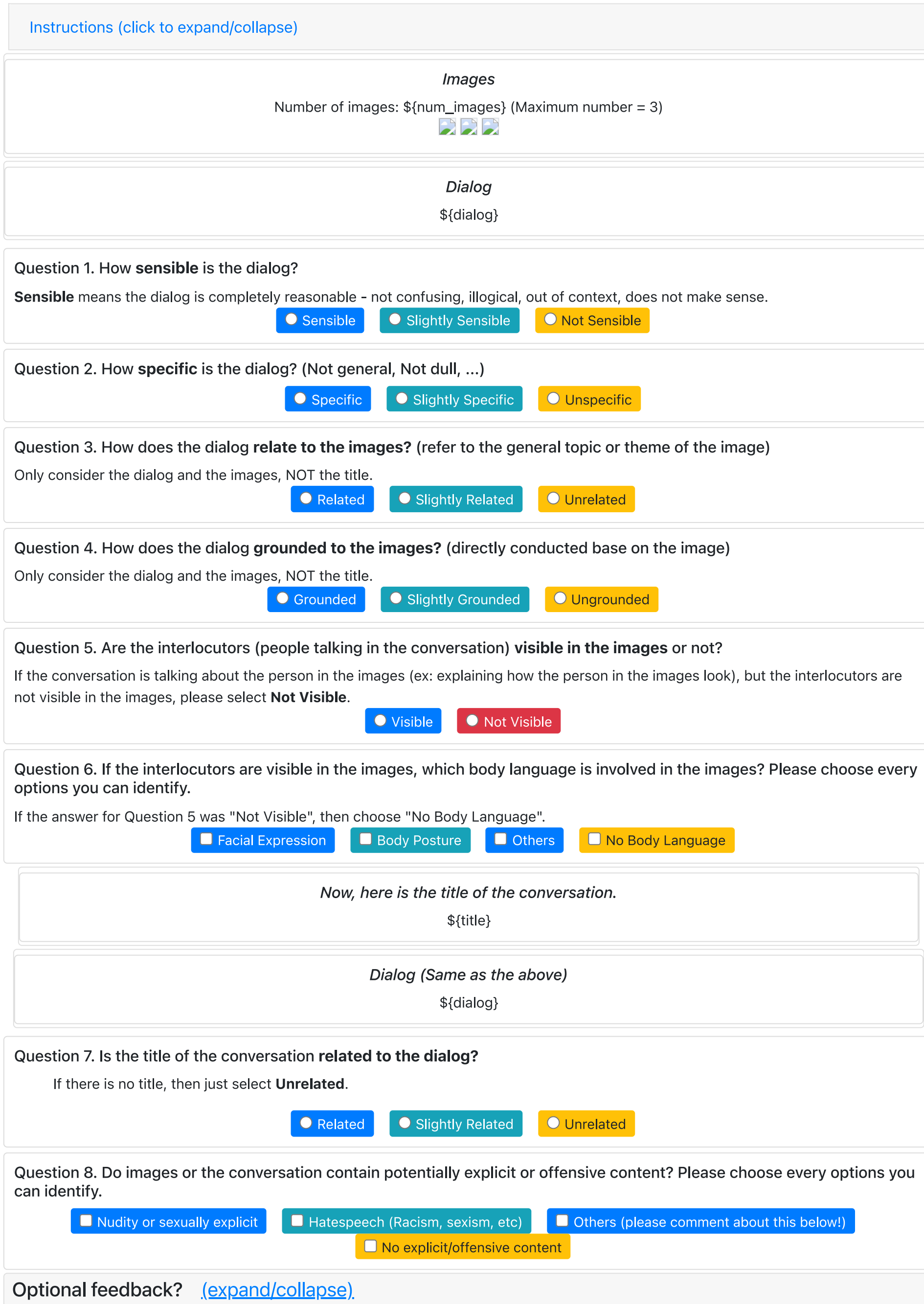
B. Human Evaluation

To provide a more detailed view of the human evaluation results, in Table 8 and Table 9, we report the complete breakdown of human evaluation results. These results complement the summarized results displayed in Table 1. To collect the human annotations, we use Amazon Mechanical Turk (MTurk), a crowdsourcing platform, and ask human workers to annotate for the tasks. We set the qualification tasks to recruit the qualified human workers in MTurk. Figure 8 and Figure 9 show the interface used for human evaluation on MTurk. For human evaluation, we compensate MTurk workers with an hourly wage of $15 for their contributions.

C. Dataset Analysis

Data Statistics. Table 10 shows the statistics about YTD-18M and the other conversational datasets including both text-only and visually-grounded cases.

Details about Visual Feature Distributions. To display visual feature distributions as in Figure 3, we use $n\_neighbors = 15$ and $min\_dist = 0.1$ for UMAP. In Figure 10, we additionally show the clusters created in Figure 3 using HDBSCAN [34] with $min\_samples = 10$ and $min\_cluster\_size = 40$ for HDBSCAN, creating 11 clus-


Instructions (click to expand/collapse)

Images

Number of images: ${num_images} (Maximum number = 3)

Dialog

${dialog}

Question 1. How **sensible** is the dialog?

Sensible means the dialog is completely reasonable - not confusing, illogical, out of context, does not make sense.

Sensible | Slightly Sensible | Not Sensible

Question 2. How **specific** is the dialog? (Not general, Not dull, ...)

Specific | Slightly Specific | Unspecific

Question 3. How does the dialog **relate to the images?** (refer to the general topic or theme of the image)

Only consider the dialog and the images, NOT the title.

Related | Slightly Related | Unrelated

Question 4. How does the dialog **grounded to the images?** (directly conducted base on the image)

Only consider the dialog and the images, NOT the title.

Grounded | Slightly Grounded | Ungrounded

Question 5. Are the interlocutors (people talking in the conversation) **visible in the images** or not?

If the conversation is talking about the person in the images (ex: explaining how the person in the images look), but the interlocutors are not visible in the images, please select **Not Visible**.

Visible | Not Visible

Question 6. If the interlocutors are visible in the images, which body language is involved in the images? Please choose every options you can identify.

If the answer for Question 5 was "Not Visible", then choose "No Body Language".

Facial Expression | Body Posture | Others | No Body Language

Now, here is the title of the conversation.

${title}

Dialog (Same as the above)

${dialog}

Question 7. Is the title of the conversation **related to the dialog?**

If there is no title, then just select **Unrelated**.

Related | Slightly Related | Unrelated

Question 8. Do images or the conversation contain potentially explicit or offensive content? Please choose every options you can identify.

Nudity or sexually explicit | Hatespeech (Racism, sexism, etc) | Others (please comment about this below!) | No explicit/offensive content

Optional feedback? (expand/collapse)

Figure 8: An interface for evaluating datasets on Amazon Mechanical Turk.

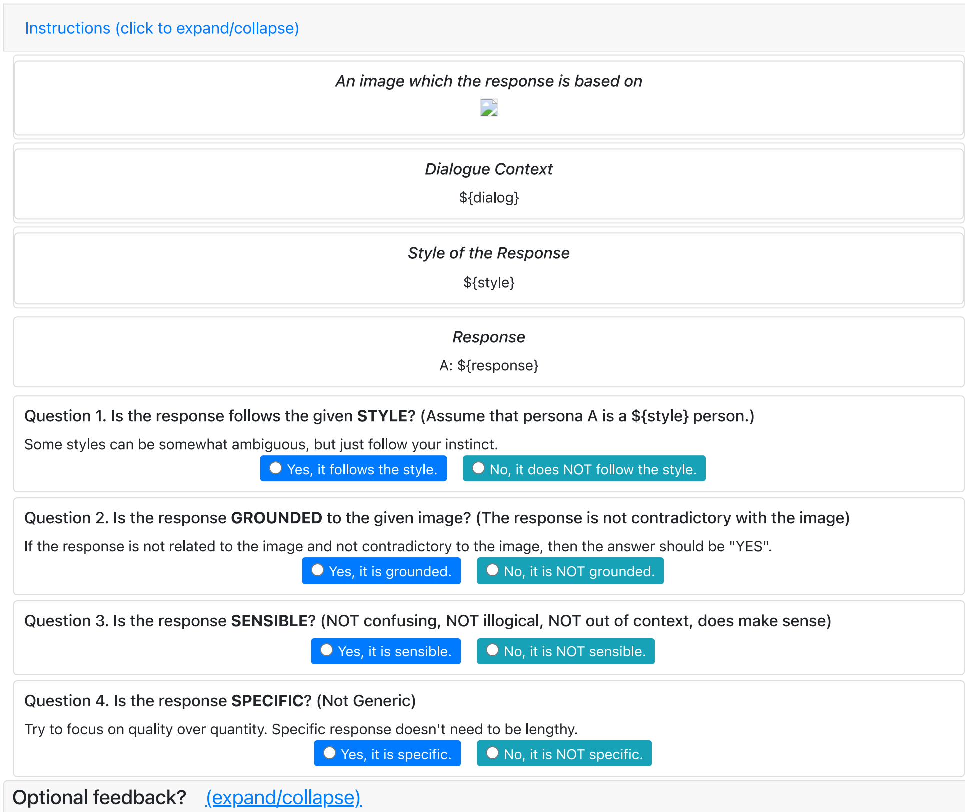


Figure 9: An interface for evaluating dialogue responses on Amazon Mechanical Turk.

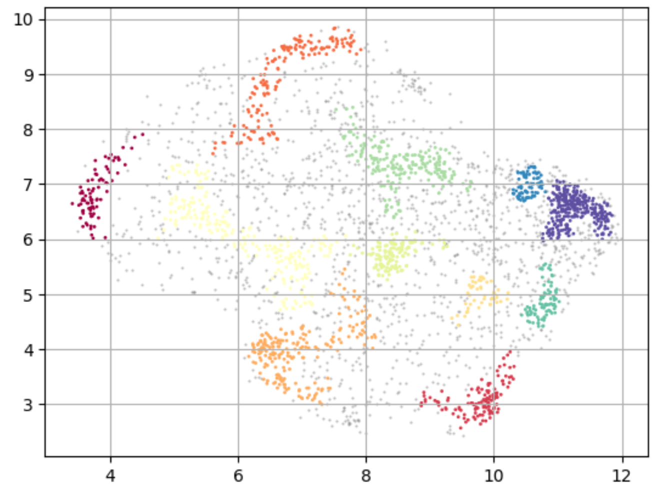


Figure 10: Visual feature distributions of visually grounded dialogue datasets with clusters.

ters in total.

D. Training and Fine-tuning 🍾CHAMPAGNE

When training CHAMPAGNE on YTD-18M, we train the model for 3 epochs with a learning rate of 3e-4, an input text sequence length of 256, a target text sequence length of 128, an input image sequence length of 576, and a batch size of 256. For fine-tuning, we also use an input text sequence length of 256, a target text sequence length of 256, and an input image sequence length of 576. In Table 11, we report other important hyper-parameters when fine-tuning CHAMPAGNE on downstream tasks.

E. Benchmarks and Evaluation Details

CMU-MOSEI. CMU-MOSEI [62] is the multimodal dataset for studying sentiments and emotions in videos. It has 16K examples in the dataset, and we use the sentiment label in our experiments. The task uses binary classification accuracy and F1 score to measure the performance. For the task, we use the template `"context: {{transcript}}, question: Is the person positive?"` to turn transcript to the input and the model produces the output from the given input.

Visual Comet. Visual Comet [38] is the benchmark for visual commonsense reasoning where the event from a still image is given. The dataset contains 59K examples, and the task uses generative evaluation so that the model generates five results and compares these results with the references using CIDEr-D[57] and BLEU-4 [37]. For the task, we use the template `"Event: {{event}} Before, what the person needed to do ?"` to turn given event to the input.

Visual Dialog. Visual Dialog [12] is a visual conversational QA dataset, consisted of 150K dialogue examples. In particular, for each example, an image, a dialogue history, and a follow-up question about the image is given, and model should answer the question. The task reports Normalized Discounted Cumulative Gain (NDCG) [23] for evaluation, where each answer has 100 candidate options and four human workers annotated relevance for each candidate option. Each given image has an caption from COCO challenge and a dialogue history, and we use the template `"<extra_id_1> {{image_caption}} <extra_id_0> {{dialogue_turn_1}} <extra_id_0> ... {{dialogue_turn_n}}"` to format the given inputs.

Image Chat. Image Chat [50] is the dataset containing 200K dialogues and each dialogue is grounded to the image. Specifically, for each conversation, an image is given and two different styles (*e.g.* "Happy", "Sad") are assigned to speakers and the speakers conduct a conversation based on the image and the styles. For the task, we use the template `"<extra_id_1> Conversation with {{style}} person <extra_id_0> {{dialogue_turn_1}} <extra_id_0> ... {{dialogue_turn_n}}"` to format the given inputs.

Open-domain Text-only Conversations. We use five open-domain text-only conversation benchmarks in this study: Blended Skill Talk (BST), ConvAI2, Empathetic Dialogue (ED), Wizard Of Wikipedia (WOW), and Wizard Of Internet. Statistics about the dataset are depicted in Table 10. These benchmarks have meta information about the conversation (*e.g.* for BST, the persona information for the speaker is given as a meta information), and we use the template `"<extra_id_1> {{meta}} <extra_id_0> {{dialogue_turn_1}} <extra_id_0> ... {{dialogue_turn_n}}"` to format the given inputs.

F. Decoding for Model Inference

In this section, we describe the decoding strategy for model inference in different benchmarks. To decode the results for Visual Comet, we use beam decoding with a beam size of 10. For Image Chat and other open-domain text-only conversation, we follow same decoding strategy from [51] for a fair comparison. Specifically, we apply beam decoding with a beam size of 10, a minimum beam length of 20. We also use a subsequence blocking of 3-grams to prevent

Model	CHAMPAGNE-BASE	CHAMPAGNE-LARGE	CHAMPAGNE-XL
CMU-MOSEI			
Epochs	2	2	2
Learning Rate	3e-4	3e-4	3e-4
Batch Size	256	256	256
Visual Comet			
Epochs	6	6	2
Learning Rate	3e-4	3e-4	3e-4
Batch Size	256	256	64
Visual Dialog			
Epochs	2	2	2
Learning Rate	3e-4	3e-4	3e-4
Batch Size	256	256	64
Mixture of Conversation Benchmarks			
Epochs	1	1	1
Learning Rate	3e-4	3e-4	1e-4
Batch Size	256	256	256

Table 11: Hyperparameters for fine-tuning CHAMPAGNE on CMU-MOSEI, Visual Comet, Visual Dialog, and mixture of conversation benchmarks. We use the same hyperparameters for fine-tuning same sized Unified-IO$_{PT}$.

Figure 11: Examples of conversation based on unseen images from COCO [31] and prompts (colored in blue) between person and CHAMPAGNE-XL finetuned on mixture of dialog benchmarks. The utterances with the prefix *Bot:* are generated by the model. Bot in (a) pretends to be a photographer and describes the picture as a photographer, whereas in (b), bot responds with specific knowledge about the picture as an ornithologist.

model from generating repeated 3-grams of the input context and repeating within the generated response. To obtain qualitative results in Figure 7, we use the minimum beam length of 10 instead of 20 since large number of minimum beam length causes a degeneration, and use temperature sampling [18] with $temperature = 0.3$ and $topk = 5$. In Figure 11, we provide additional examples of conversations between humans and CHAMPAGNE-XL that has been fine-tuned on a mixture of dialogue benchmarks.

G. Additional Evaluation Results on Visual Dialog

In the main paper, we followed the recommendations of the official Visual Dialog challenge[4], which only use ranking-basd metrics like Recall@K and MRR as supplementary measures rather than primary metrics. Visual Dialog dataset contains dense annotations per each candidate and measures performance based on NDCG to account for the nuanced evaluation, and the fact that dialogue is one-to-many task.

In contrast, ranking-based metrics assume the existence of a single correct response for a given context, which is not an accurate assumption for dialogue. For example, ranking metrics can be penalize models arbitrarily for their scoring candidate semantically equivalent options (*e.g.*, 'yes' and 'yes it is'). Even if the model selects a viable answer, the rank-based score may be low. Moreover, dialogue tasks are inherently one-to-many problems, where multiple possible responses exist for a given dialogue context. Nonetheless, for comparison purposes we add Recall@K and MRR for comparison purposes in Table 12.

[4] https://visualdialog.org/challenge/2019

Metric	NDCG ($\times$100) ($\uparrow$)	MRR	Recall@1	Recall@5	Recall@10	Mean Rank
Zero-shot						
Flamingo-80B	52.0	-	-	-	-	-
ESPER [61]	22.3	25.7	14.6	-	-	-
FROMAGe [26]	16.5	22.0	17.6	20.1	25.1	-
🍾CHAMPAGNE-XL	25.5	16.7	9.14	20.9	30.2	-
Fine-tuned						
Flamingo-80B	61.8	-	-	-	-	-
AlignVD [8]	67.2	70.5	57.6	87.1	94.2	3.05
Unified-IO$_{PT}$ BASE	58.9	49.1	38.8	59.4	71.2	9.95
Unified-IO$_{PT}$ LARGE	60.3	49.6	39.2	59.9	72.3	9.50
Unified-IO$_{PT}$ XL	65.6	54.0	43.4	65.0	77.1	7.76
🍾CHAMPAGNE-BASE	60.0	50.1	39.9	59.9	71.5	9.81
🍾CHAMPAGNE-LARGE	62.5	51.6	41.0	62.1	74.1	8.84
🍾CHAMPAGNE-XL	68.2	56.1	45.1	67.8	78.3	7.48

Table 12: Evaluation results on Visual Dialog valid set in finetuned and zero-shot settings. For fair comparison, we report baselines that do not use additional dense annotations to finetune the model. All the results are evaluated using the official server.