$\mathbf{V}^2$Dial: Unification of Video and Visual Dialog via Multimodal Experts (Supplementary Material)

A. Training Details

A.1. Training Objectives

In addition to the proposed spatial-temporal contrastive learning (STC) and spatial-temporal matching (STM), we trained our model with the following established vision-language objectives.

Masked Language Modeling teaches the model to predict masked text tokens given both the visual and textual context. As in [4, 13] we mask 15% of the tokens and minimize the loss

$$\mathcal{L}_{\text{mlm}} = \mathbb{E}_{(\mathbf{V}^{\text{vis}},\bar{\mathbf{T}}^{\text{cap}})}\left[\mathcal{H}(\mathbf{y}^{\text{mlm}},\mathbf{p}^{\text{mlm}})\right], \tag{1}$$

where $\mathbf{y}^{\text{mlm}}$ and $\mathbf{p}^{\text{mlm}}$ denote the ground-truth and predicted probabilities of the masked tokens whereas $\mathbf{V}^{\text{vis}}$ and $\bar{\mathbf{T}}^{\text{cap}}$ are the visual and masked caption token embeddings, respectively.

Vision-Text Contrastive Learning helps the model better align the video/image and the text features and is defined similarly to STC as

$$\mathcal{L}_{\text{vtc}} = \frac{1}{2}\mathbb{E}_{(\mathbf{V}^{\text{vis}},\mathbf{T}^{\text{cap}})}\left[\mathcal{H}\left(\mathbf{y}^{\text{v2t}},\mathbf{p}^{\text{v2t}}\right)+\mathcal{H}\left(\mathbf{y}^{\text{t2v}},\mathbf{p}^{\text{t2v}}\right)\right], \tag{2}$$

where $\mathbf{p}^{\text{v2t}}$ and $\mathbf{p}^{\text{t2v}}$ are the softmax normalized vision-to-text and text-to-vision similarities defined as in Equation 14 and Equation 15 of the main text. $\mathbf{y}^{\text{v2t}}$ and $\mathbf{y}^{\text{t2v}}$ are their respective ground-truth one-hot similarities.

Vision-Text Matching is defined similarly to STM as a binary classification problem and complements the VTC by teaching the model to distinguish between matched and unmatched paired vision-text features. We use a video/image and its corresponding caption as a positive example. The negative examples are constructed via negative sampling of captions from different visual inputs. Formally,

$$\mathcal{L}_{\text{vtm}} = \mathbb{E}_{(\mathbf{V}^{\text{vis}},\mathbf{T}^{\text{cap}})}\left[\mathcal{H}(\mathbf{y}^{\text{vtm}},\mathbf{p}^{\text{vtm}})\right], \tag{3}$$

where $\mathbf{p}^{\text{stm}}$ and $\mathbf{y}^{\text{stm}}$ are the predicted and the ground-truth two-class probabilities, respectively. For completeness, we list the detailed hyperparameters of our model in Table 1.

Category	**Hyperparameter**	
Model	Number of expert-based layers N	12
	Number of multimodal experts layers L	9
	Number of fusion experts layers $(N-L)$	3
	Joint hidden dimension D	1024
	Number of frames F	4
	Number of patches per frame P	64
	Hidden dimension of LLM	1024
	Dimension of LLM linear layer	$(1024, 1024)$
	Dimension of linear layers Θ_*	$(1024, 256)$
Optimization	Optimizer	AdamW
	Learning rate schedule	linear
	Minimum learning rate value	$5e-5$
	Base learning rate value	$1e-4$
	Weight decay	0.01
	Gradient clipping value	1.0
	Effective batch size	48
Hardware	GPU model	`A100`
	Number of GPUs	8
	Distributed training	`DDP`

Table 1. Detailed hyperparameter setting of $\mathbf{V}^2$**Dial**.

B. Additional Model Comparisons

To complement Table 4 of the main text, we compared our model with additional *fine-tuned* baselines on the early two versions of AVSD (i.e. AVSD-DSTC8 and AVSD-DSTC7). As shown in Table 2, $\mathbf{V}^2$**Dial** managed to outperform these baselines as well across all metrics of the dataset.

C. Qualitative Samples

We provide additional qualitative samples comprising of both success and failure cases of our model. Figure 1 and Figure 2 illustrate some zero-shot samples for AVSD and VisDial, respectively. Additional fine-tuning examples for both datasets are shown in Figure 3 and Figure 4.

As defined in Section 3.1 of the main text, we denote with C, H_r, and Q_r the caption, the dialog history, and the current question, respectively. Similar to Figure 5 of the main text, we highlight the caption in green, the dialog history in orange, and the current question-answer pair in blue for zero-shot and pink for fine-tuning evaluation. Furthermore, we use the symbols and to indicate the generated and the golden ground-truth answers, respec-

Model	AVSD-DSTC8							AVSD-DSTC7						
	B-1	**B-2**	**B-3**	**B-4**	**M**	**R**	**C**	**B-1**	**B-2**	**B-3**	**B-4**	**M**	**R**	**C**
Models from the main text														
PDC$_{ICLR'21}$ [11]	74.9	62.9	52.8	43.9	28.5	59.2	120.1	77.0	65.3	53.9	44.9	29.2	60.6	129.5
THAM$_{EMNLP'22}$ [17]	76.4	64.1	53.8	45.5	30.1	61.0	130.4	77.8	65.4	54.9	46.8	30.8	61.9	133.5
DialogMCF$_{TASLP'23}$ [3]	75.6	63.3	53.2	44.9	29.3	60.1	125.3	77.7	65.3	54.7	45.7	30.6	61.3	135.2
♦VideoLLAMA 2$_{arXiv'24}$ [5]	53.3	39.0	29.1	22.2	24.8	46.3	74.0	56.2	41.1	30.7	23.2	26.4	48.5	79.2
MST-MIXER$_{ECCV'24}$ [1]	**77.1**	**65.6**	55.7	47.1	30.2	61.8	133.6	78.4	66.0	55.8	47.1	31.0	62.0	136.5
Additional models														
MTN$_{ACL'19}$ [9]	–	–	–	–	–	–	–	71.5	58.1	47.6	39.2	26.9	55.9	106.6
JMAN$_{AAAI'20}$ [6]	64.5	50.4	40.2	32.4	23.2	52.1	87.5	66.7	52.1	41.3	33.4	23.9	53.3	94.1
VGD$_{ACL'20}$ [8]	–	–	–	–	–	–	–	74.9	62.0	52.0	43.6	28.2	58.2	119.4
BiST$_{EMNLP'20}$ [10]	68.4	54.8	45.7	37.6	27.3	56.3	101.7	75.5	61.9	51.0	42.9	28.4	58.1	119.2
SCGA$_{AAAI'21}$ [7]	71.1	59.3	49.7	41.6	27.6	56.6	112.3	74.5	62.2	51.7	43.0	28.5	57.8	120.1
RLM$_{TASLP'21}$ [14]	74.6	62.6	52.8	44.5	28.6	59.8	124.0	76.5	64.3	54.3	45.9	29.4	60.6	130.8
AV-TRN$_{ICASSP'22}$ [16]	–	–	–	39.4	25.0	54.5	99.7	–	–	–	40.6	26.2	55.4	107.9
VGNMN$_{NAACL'22}$ [12]	–	–	–	–	–	–	–	–	–	–	42.9	27.8	57.8	118.8
COST$_{ECCV'22}$ [15]	69.5	55.9	46.5	3.82	27.8	57.4	105.1	72.3	58.9	48.3	40.0	26.6	56.1	108.5
MRLV$_{NeurIPS'22}$ [2]	–	–	–	–	–	–	–	–	59.2	49.3	41.5	26.9	56.9	115.9
V^{2}Dial	76.8	65.5	**55.8**	**47.5**	**30.4**	**62.1**	**135.7**	**78.9**	**66.5**	**56.1**	**47.4**	**31.2**	**62.3**	**139.8**

Table 2. To complement Table 4 of the main text, we compared our **V^{2}Dial** with additional fine-tuned models on AVSD-DSTC8 and AVSD-DSTC7.

tively. ✅ / ❌ mark success / failure cases. For VisDial, we additionally use 🥇 to show the top ranked candidate answers (i.e. the most similar to the generated responses).

References

[1] Adnen Abdessaied, Lei Shi, and Andreas Bulling. Multi-Modal Video Dialog State Tracking in the Wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024. 2

[2] Huda Alamri, Anthony Bilic, Michael Hu, Apoorva Beedu, and Irfan Essa. End-to-end multimodal representation learning for video dialog. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[3] Zhe Chen, Hongcheng Liu, and Yu Wang. DialogMCF: Multimodal Context Flow for Audio Visual Scene-Aware Dialog. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023. 2

[4] Feng Cheng, Xizi Wang, Jie Lei, David Crandall, Mohit Bansal, and Gedas Bertasius. VindLU: A Recipe for Effective Video-and-Language Pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1

[5] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, and Lidong Bing. VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs. *arXiv preprint arXiv:2406.07476*, 2024. 2

[6] Yun-Wei Chu, Kuan-Yen Lin, Chao-Chun Hsu, and Lun-Wei Ku. Multi-step joint-modality attention network for scene-aware dialogue system. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) Workshops*, 2020. 2

[7] Junyeong Kim, Sunjae Yoon, Dahyun Kim, and Chang D. Yoo. Structured co-reference graph attention for video-grounded dialogue. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021. 2

[8] Hung Le and Steven C.H. Hoi. Video-Grounded Dialogues with Pretrained Generation Language Models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020. 2

[9] Hung Le, Doyen Sahoo, Nancy Chen, and Steven Hoi. Multimodal transformer networks for end-to-end video-grounded dialogue systems. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019. 2

[10] Hung Le, Doyen Sahoo, Nancy Chen, and Steven C.H. Hoi. BiST: Bi-directional Spatio-Temporal Reasoning for Video-Grounded Dialogues. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. 2

[11] Hung Le, Nancy F. Chen, and Steven Hoi. Learning reasoning paths over semantic graphs for video-grounded dialogues. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 2

[12] Hung Le, Nancy F. Chen, and Steven C. H. Hoi. VGNMN: video-grounded neural module network to video-grounded language tasks. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2022. 2

[13] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1

[14] Zekang Li, Zongjia Li, Jinchao Zhang, Yang Feng, and Jie Zhou. Bridging text and video: A universal multimodal

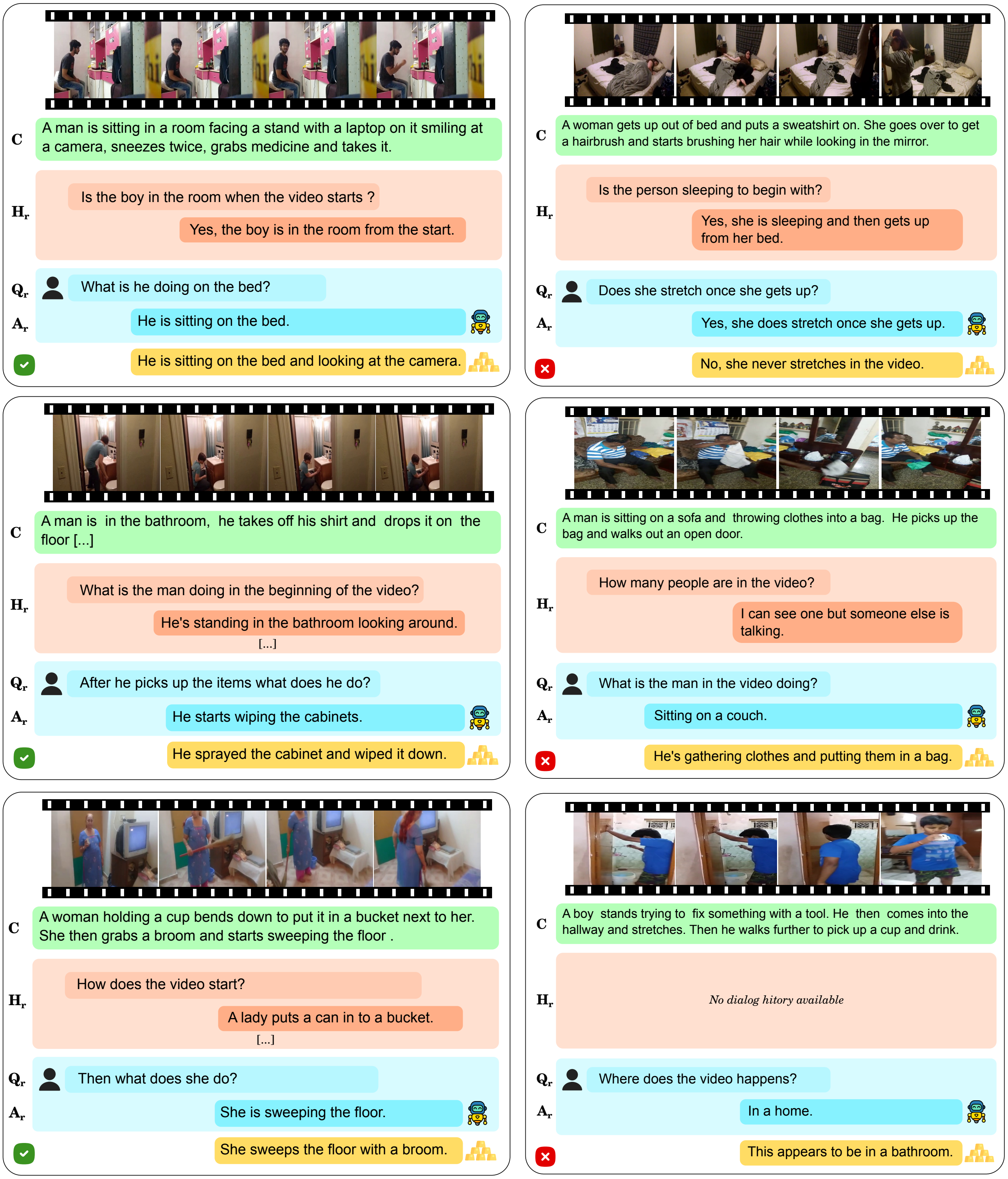

Figure 1. **Zero-shot qualitative examples on AVSD.** We denote with C, H_r, Q_r, A_r the caption, the dialog history, the current question, and its response as generated from our model, respectively. (= generated answers, = golden ground-truth answers, / = success / failure cases).

Figure 2. **Zero-shot qualitative examples on VisDial.** We denote with C, H_r, Q_r, A_r the caption, the dialog history, the current question, and its response as generated from our model, respectively. (= generated answers, = top ranked candidate answers, = golden ground-truth answers, / = success / failure cases).

transformer for audio-visual scene-aware dialog. *Transactions on Audio, Speech, and Language Processing*, 2021. 2

[15] Hoang-Anh Pham, Thao Minh Le, Vuong Le, Tu Minh Phuong, and Truyen Tran. Video Dialog as Conversation

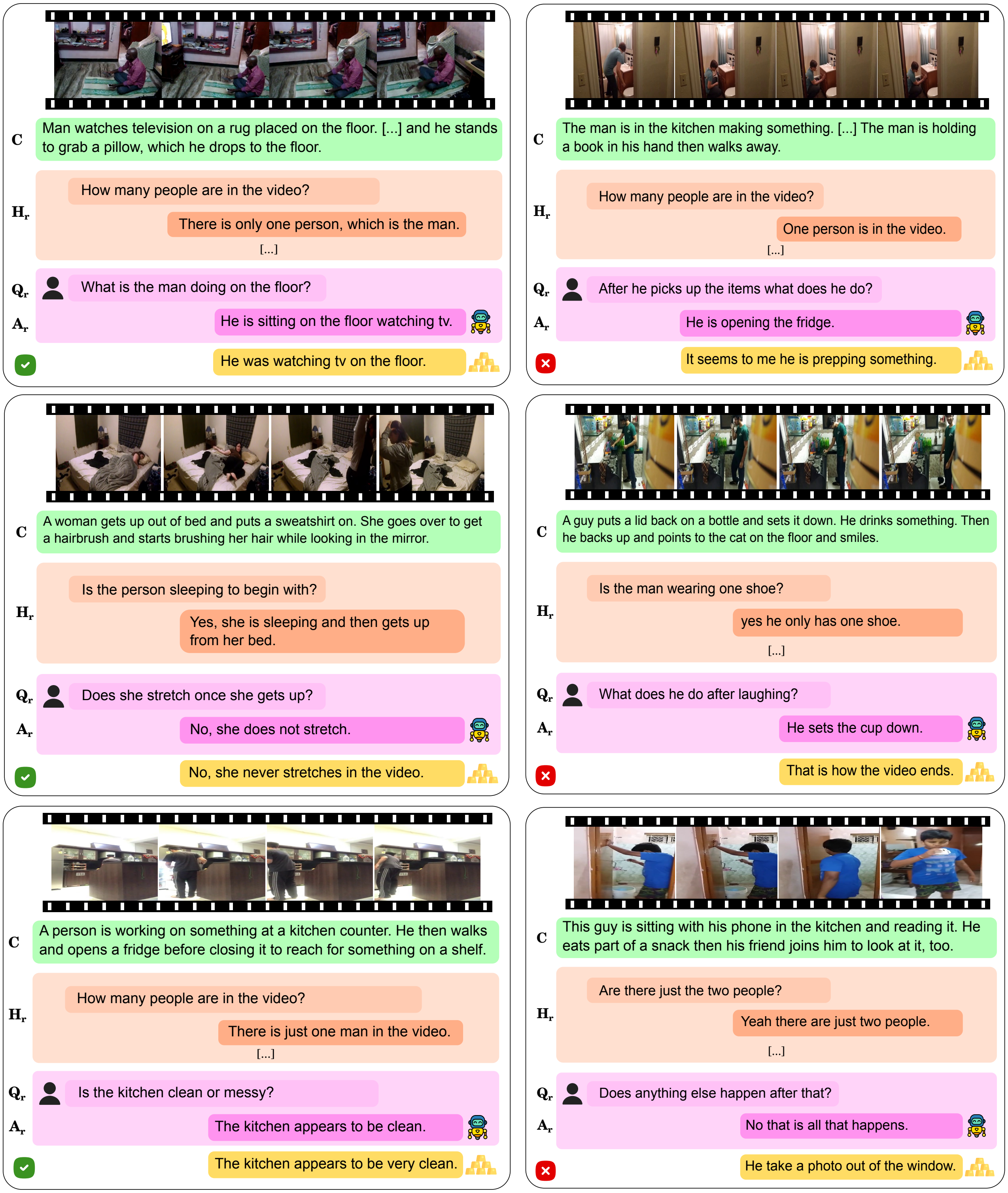

Figure 3. **Fine-tuning qualitative examples on AVSD.** We denote with C, H$_r$, Q$_r$, A$_r$ the caption, the dialog history, the current question, and its response as generated from our model, respectively. (= generated answers, = golden ground-truth answers, / = success / failure cases).

Figure 4. **Fine-tuning qualitative examples on VisDial.** We denote with C, H$_r$, Q$_r$, A$_r$ the caption, the dialog history, the current question, and its response as generated from our model, respectively. (= generated answers, = top ranked candidate answers, = golden ground-truth answers, / = success / failure cases).

about Objects Living in Space-Time. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2

[16] Ankit Shah, Shijie Geng, Peng Gao, Anoop Cherian, Takaaki Hori, Tim K Marks, Jonathan Le Roux, and Chiori Hori.

Audio-visual scene-aware dialog and reasoning using audio-visual transformers with joint student-teacher learning. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022. 2

[17] Sunjae Yoon, Eunseop Yoon, Hee Suk Yoon, Junyeong Kim, and Chang Yoo. Information-theoretic text hallucination reduction for video-grounded dialogue. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2022. 2