

# V<sup>2</sup>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

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

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

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

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,

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

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)
Optimization	Dimension of linear layers $\Theta_*$	(1024, 256)
	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
Hardware	Effective batch size	48
	GPU model	A100
	Number of GPUs	8
	Distributed training	DDP

Table 1. Detailed hyperparameter setting of V<sup>2</sup>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, V<sup>2</sup>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-





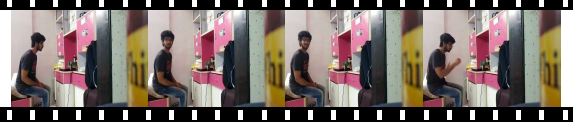
Model	AVSD-DSTC8							AVSD-DSTC7						
	B-1	B-2	B-3	B-4	M	R	C	B-1	B-2	B-3	B-4	M	R	C
<i>Models from the main text</i>														
PDC <sub>JCLR'21</sub> [11]	74.9	62.9	52.8	43.9	28.5	59.2	120.1	77.0	65.3	53.9	44.9	29.2	60.6	129.5
THAM <sub>EMNLP'22</sub> [17]	76.4	64.1	53.8	45.5	30.1	61.0	130.4	77.8	65.4	54.9	46.8	30.8	61.9	133.5
DialogMCF <sub>TASLP'23</sub> [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 <sub>arXiv'24</sub> [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 <sub>ECCV'24</sub> [1]	<b>77.1</b>	<b>65.6</b>	<u>55.7</u>	<u>47.1</u>	<u>30.2</u>	<u>61.8</u>	<u>133.6</u>	<u>78.4</u>	<u>66.0</u>	<u>55.8</u>	<u>47.1</u>	<u>31.0</u>	<u>62.0</u>	<u>136.5</u>
<i>Additional models</i>														
MTN <sub>ACL'19</sub> [9]	—	—	—	—	—	—	—	71.5	58.1	47.6	39.2	26.9	55.9	106.6
JMAN <sub>AAAI'20</sub> [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 <sub>ACL'20</sub> [8]	—	—	—	—	—	—	—	74.9	62.0	52.0	43.6	28.2	58.2	119.4
BiST <sub>EMNLP'20</sub> [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 <sub>AAAI'21</sub> [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 <sub>TASLP'21</sub> [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 <sub>JCASSP'22</sub> [16]	—	—	—	39.4	25.0	54.5	99.7	—	—	—	40.6	26.2	55.4	107.9
VGNMN <sub>NAACL'22</sub> [12]	—	—	—	—	—	—	—	—	—	—	42.9	27.8	57.8	118.8
COST <sub>ECCV'22</sub> [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 <sub>NeurIPS'22</sub> [2]	—	—	—	—	—	—	—	—	59.2	49.3	41.5	26.9	56.9	115.9
<b>V<sup>2</sup>Dial</b> 	<u>76.8</u>	<u>65.5</u>	<u>55.8</u>	<u>47.5</u>	<u>30.4</u>	<u>62.1</u>	<u>135.7</u>	<u>78.9</u>	<u>66.5</u>	<u>56.1</u>	<u>47.4</u>	<u>31.2</u>	<u>62.3</u>	<u>139.8</u>

Table 2. To complement Table 4 of the main text, we compared our V<sup>2</sup>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. Multi-modal 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



**C** A man is sitting in a room facing a stand with a laptop on it smiling at a camera, sneezes twice, grabs medicine and takes it.

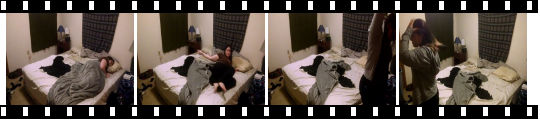
**H<sub>r</sub>** Is the boy in the room when the video starts ?

**A<sub>r</sub>** Yes, the boy is in the room from the start.

**Q<sub>r</sub>** What is he doing on the bed?

**A<sub>r</sub>** He is sitting on the bed.

He is sitting on the bed and looking at the camera.



**C** A woman gets up out of bed and puts a sweatshirt on. She goes over to get a hairbrush and starts brushing her hair while looking in the mirror.

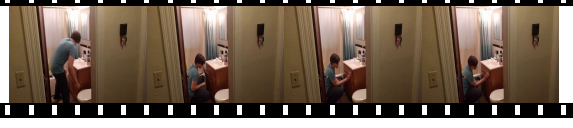
**H<sub>r</sub>** Is the person sleeping to begin with?

**A<sub>r</sub>** Yes, she is sleeping and then gets up from her bed.

**Q<sub>r</sub>** Does she stretch once she gets up?

**A<sub>r</sub>** Yes, she does stretch once she gets up.

No, she never stretches in the video.



**C** A man is in the bathroom, he takes off his shirt and drops it on the floor [...]


**H<sub>r</sub>** What is the man doing in the beginning of the video?

**A<sub>r</sub>** He's standing in the bathroom looking around. [...]

**Q<sub>r</sub>** After he picks up the items what does he do?

**A<sub>r</sub>** He starts wiping the cabinets.

He sprayed the cabinet and wiped it down.



**C** A man is sitting on a sofa and throwing clothes into a bag. He picks up the bag and walks out an open door.


**H<sub>r</sub>** How many people are in the video?

**A<sub>r</sub>** I can see one but someone else is talking.

**Q<sub>r</sub>** What is the man in the video doing?

**A<sub>r</sub>** Sitting on a couch.

He's gathering clothes and putting them in a bag.



**C** A woman holding a cup bends down to put it in a bucket next to her. She then grabs a broom and starts sweeping the floor .

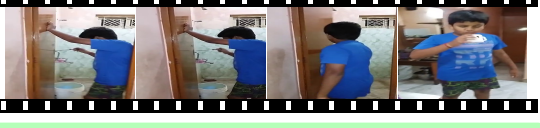
**H<sub>r</sub>** How does the video start?

**A<sub>r</sub>** A lady puts a can in to a bucket. [...]

**Q<sub>r</sub>** Then what does she do?

**A<sub>r</sub>** She is sweeping the floor.

She sweeps the floor with a broom.



**C** A boy stands trying to fix something with a tool. He then comes into the hallway and stretches. Then he walks further to pick up a cup and drink.

**H<sub>r</sub>** No dialog hitory available

**Q<sub>r</sub>** Where does the video happens?


**A<sub>r</sub>** In a home.

This appears to be in a bathroom.

Figure 1. **Zero-shot qualitative examples on AVSD.** We denote with C, H<sub>r</sub>, Q<sub>r</sub>, A<sub>r</sub> 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, 1 = top ranked candidate answers, 🏆 = golden ground-truth answers, ✓ / ✗ = success / failure cases).



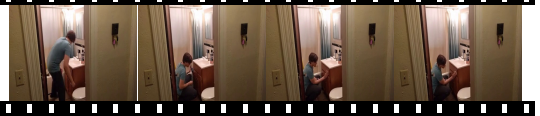
**C** Man watches television on a rug placed on the floor. [...] and he stands to grab a pillow, which he drops to the floor.

**H<sub>r</sub>** How many people are in the video?  
There is only one person, which is the man.  
[...]

**Q<sub>r</sub>** What is the man doing on the floor?

**A<sub>r</sub>** He is sitting on the floor watching tv.   
He was watching tv on the floor. 



✓




**C** The man is in the kitchen making something. [...] The man is holding a book in his hand then walks away.

**H<sub>r</sub>** How many people are in the video?  
One person is in the video.  
[...]

**Q<sub>r</sub>** After he picks up the items what does he do?

**A<sub>r</sub>** He is opening the fridge.   
It seems to me he is prepping something. 



✗



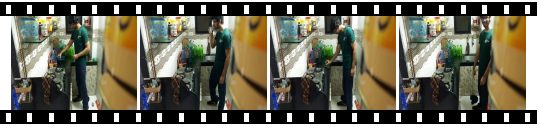
**C** A woman gets up out of bed and puts a sweatshirt on. She goes over to get a hairbrush and starts brushing her hair while looking in the mirror.

**H<sub>r</sub>** Is the person sleeping to begin with?  
Yes, she is sleeping and then gets up from her bed.

**Q<sub>r</sub>** Does she stretch once she gets up?

**A<sub>r</sub>** No, she does not stretch.   
No, she never stretches in the video. 



✓




**C** A guy puts a lid back on a bottle and sets it down. He drinks something. Then he backs up and points to the cat on the floor and smiles.

**H<sub>r</sub>** Is the man wearing one shoe?  
yes he only has one shoe.  
[...]

**Q<sub>r</sub>** What does he do after laughing?

**A<sub>r</sub>** He sets the cup down.   
That is how the video ends. 



✗




**C** A person is working on something at a kitchen counter. He then walks and opens a fridge before closing it to reach for something on a shelf.

**H<sub>r</sub>** How many people are in the video?  
There is just one man in the video.  
[...]

**Q<sub>r</sub>** Is the kitchen clean or messy?

**A<sub>r</sub>** The kitchen appears to be clean.   
The kitchen appears to be very clean. 



✓



**C** This guy is sitting with his phone in the kitchen and reading it. He eats part of a snack then his friend joins him to look at it, too.

**H<sub>r</sub>** Are there just the two people?  
Yeah there are just two people.  
[...]

**Q<sub>r</sub>** Does anything else happen after that?

**A<sub>r</sub>** No that is all that happens.   
He take a photo out of the window. 

✗



Figure 3. **Fine-tuning qualitative examples on AVSD.** We denote with C, H<sub>r</sub>, Q<sub>r</sub>, A<sub>r</sub> 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).

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