V^{2}Dial: Unification of Video and Visual Dialog via Multimodal Experts

Adnen Abdessaied Anna Rohrbach Marcus Rohrbach Andreas Bulling

University of Stuttgart, Germany TU Darmstadt, Germany hessian.AI, Germany

https://www.collaborative-ai.org/publications/abdessaied25_cvpr/

Abstract

*We present **V^{2}Dial** – a novel expert-based model specifically geared towards simultaneously handling image and video input data for multimodal conversational tasks. Current multimodal models primarily focus on simpler tasks (e.g., VQA, VideoQA, video-text retrieval) and often neglect the more challenging conversational counterparts, such as video and visual/image dialog. Moreover, works on both conversational tasks evolved separately from each other despite their apparent similarities, limiting their applicability potential. To this end, we propose to unify both tasks using a single model that for the first time jointly learns the spatial and temporal features of images and videos by routing them through dedicated experts and aligns them using matching and contrastive learning techniques. Furthermore, we systemically study the domain shift between the two tasks by investigating whether and to what extent these seemingly related tasks can mutually benefit from their respective training data. Extensive evaluations on the widely used video and visual dialog datasets of AVSD and VisDial show that our model achieves new state-of-the-art results across four benchmarks both in zero-shot and fine-tuning settings.*

1. Introduction

Enabled by the availability of large-scale training data [10, 14, 41] and advances in model design [12, 24, 37, 48, 52], the field of vision-and-language learning saw unprecedented success in recent years. However, current multimodal foundational models [7, 32, 37, 50, 54] still mainly focus on single-round tasks (e.g., VQA [8], VideoQA [58], video-text and text-video retrieval [59]). In contrast, the significantly more challenging conversational tasks, such as visual [1, 22] and video dialog [5], received considerably less attention. Furthermore, methods for these different tasks have advanced independently of each other despite the apparent structural similarities between them. They both operate on a visual input (i.e. an image or video), a short visual description (caption), and a dialog history composed of previous question-answer pairs. On the one hand, visual dialog models [4, 15, 16, 44, 56] have been primarily trained to rank a list of candidate answers using a Next Sentence Prediction (NSP) head similar to BERT [23] and negative sampling. Thus, they are benchmarked using retrieval met-

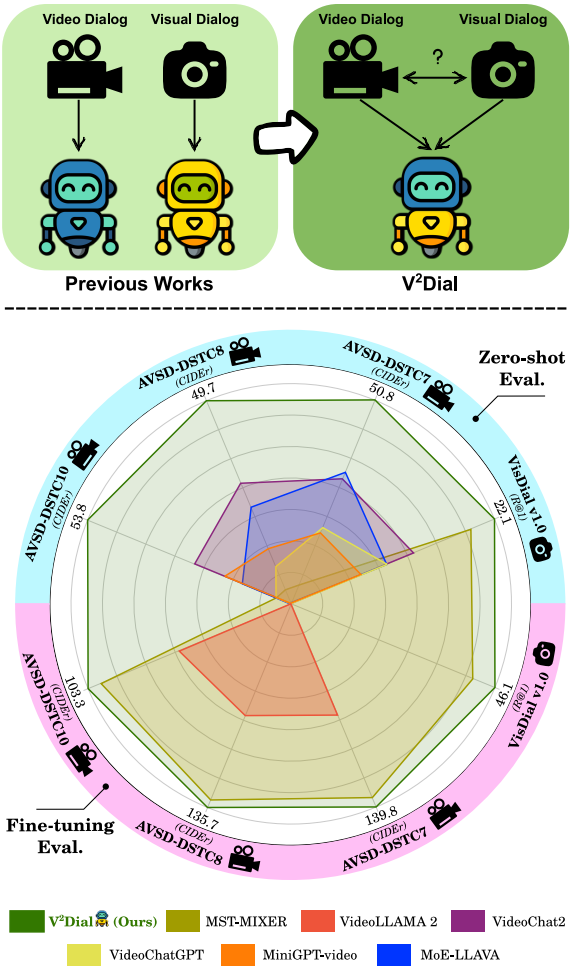

Figure 1. **V^{2}Dial** uses multimodal experts and outperforms state-of-the-art methods on both video and visual dialog in zero-shot and fine-tuning evaluation settings.

rics such as recall (R@k) and normalized discounted cumulative gain (NDCG). In contrast, video dialog models [2, 3, 6, 18, 20, 29, 46, 60] are trained to auto-regressively predict the next answer token using teacher forcing [57] and are evaluated using language generation metrics.

In this work, we mitigate the shortcomings of current dialog systems by proposing **V^{2}Dial** – a novel multimodal expert-based model capable of unifying video and visual dialog tasks without any architectural changes. Specifically, we train dedicated multimodal expert layers that separately process the features of each input modality and learn how to align them using matching and contrastive learning techniques. A key novelty of our approach is that we use dedicated experts to jointly learn the spatial and temporal features of images and videos by routing them through the appropriate experts. Then, we couple these layers with a pre-trained LLM to align their hidden states. Thanks to its modularity, our model can efficiently tackle image and video input data simultaneously and seamlessly train on both data types. In summary, our contributions are three-fold:

- We propose **V^{2}Dial** – a multimodal expert-based model that unifies visual and video dialog by simultaneously learning from image and video data. As a core novelty, it employs two experts to separately learn the spatial and temporal features of images and videos. **V^{2}Dial** outperforms state-of-the-art models in both zero-shot and fine-tuning settings (see Figure 1).
- We are the first to systematically quantify the effect of domain shift between video and visual dialog tasks based on evaluations on the two widely used datasets of AVSD [5] and VisDial [22]. To this end, we propose an alternative ranking scheme that allows computing the VisDial retrieval metric for fully generative models and enables a fair comparison with previous works.
- We are the first to evaluate AVSD in a zero-shot setting, which provides a more solid generalization evaluation of video dialog models compared to the fine-tuned setting. For this, we establish the first benchmark comparison of recent state-of-the-art multimodal models.

2. Related Work

Visual & Video Dialog. Modeled after human-human communication, visual and video dialog involve reasoning about a visual scene in the form of a video or an image through multiple question-answering rounds in natural language. In comparison to their single-round counterparts, VQA [8] and VideoQA [58], dialog models need to additionally reason about the previous dialog history together with the visual grounding and the current question to be able to answer it efficiently. The best performing visual dialog models [4, 42, 56, 61] leverage pre-trained VLMs and are trained using an NSP head, negative sampling, and binary classification loss. At test time, for each question, the candidate answers are ranked based on their respective NSP scores to compute the retrieval metrics. Although some work [15, 16, 55] claim to train generative visual dialog models, they do so by providing a generative mask where each token can only attend to its left tokens. However, they are trained using the NSP head like the discriminative models. However, this training approach is limiting and suboptimal for a unifying model. Thus, we advocate for a fully generative training paradigm and adapt the ranking scheme of VisDial answers to cater to modern generative models.

In contrast, works on video dialog follow a purely-generative training paradigm and achieved great success building on top of powerful pre-trained LLMs [30, 47]. For example, [28, 34] fine-tuned a LLM on AVSD and obtained performance boosts. More recent works [3, 29] combined LLMs with GNNs and pushed the state-of-the-art results even further. Others [60] introduced a regularization loss to mitigate hallucination. Although video dialog emerged as a natural extension to visual dialog with apparent data structure similarities, research on both tasks evolved separately. To this end, we propose a unifying model that can simultaneously learn both tasks without any architectural modifications and *for the first time*; systemically study the effect of domain shift between both tasks using the AVSD and VisDial v1.0 datasets.

Multimodal Expert-based Training. Enhancing models with expert-based training has shown promising potential in boosting performance while maintaining computational efficiency [25, 63, 64]. Some works [12, 54] explored using single modality specific experts within a multimodal transformer architecture. Specifically, they used *one* vision and *one* language specific FFN after a shared multi-head self-attention block. Other works [35, 43] explored using multiple sparse modality-agnostic experts and trained them using soft-routers. Our work is positioned at the middle ground of the previously mentioned research directions: We propose to use multiple hard-routed experts per modality to be able to capture more fine-grained features compared to a single expert or multiple modality agnostic experts. Specifically, to the best of our knowledge, **V^{2}Dial** is the *first* model that learns disentangled spatial and temporal features using two dedicated experts that jointly learn from image and video data. In addition, we propose to deploy two separate language experts (for caption and context) in order to tackle the unique challenges of multimodal conversational tasks.

3. V^{2}Dial

3.1. Joint Problem Formulation

We use a fully generative formulation to unify both video and visual dialog tasks. Specifically, given visual input $\mathtt{V}$ (video/image), a corresponding visual description (caption

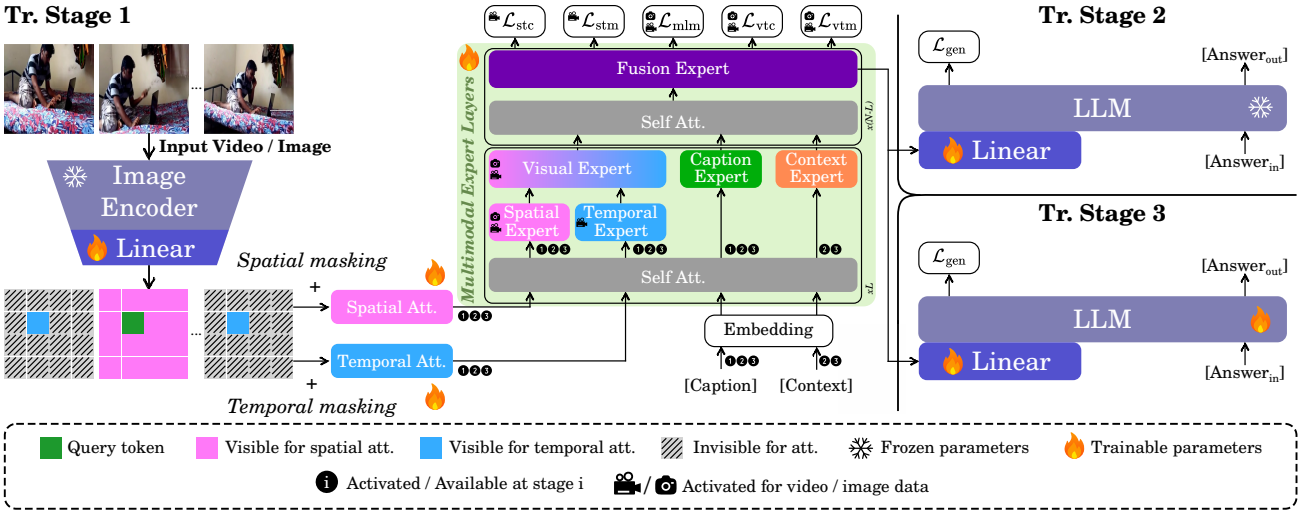

Figure 2. **Architectural overview of V^{2}Dial**. We adopt a training strategy composed of three stages. *First*, we only train the multimodal expert layers using spatial-temporal and video/image text matching losses ($\mathcal{L}_{\text{stm}}$, $\mathcal{L}_{\text{vtm}}$), spatial-temporal and video/image contrastive learning losses ($\mathcal{L}_{\text{stc}}$, $\mathcal{L}_{\text{vtc}}$), and masked language modeling loss ($\mathcal{L}_{\text{mlm}}$). *Second*, we couple the expert layers with a frozen pre-trained LLM end-to-end, using a generative loss $\mathcal{L}_{\text{gen}}$ to align their hidden representations. *Finally*, we additionally fine-tune the LLM weights on the downstream benchmarks. Each expert is a feed-forward network (FFN) composed of two fully connected layers.

C), a dialog history $\mathtt{H}_r = \{(\mathtt{Q}_1, \mathtt{A}_1), ..., (\mathtt{Q}_{r-1}, \mathtt{A}_{r-1})\}$ comprised of the previous question-answer pairs $\{(\mathtt{Q}_i,\mathtt{A}_i)\}_{i=1}^{r-1}$ and the current question $\mathtt{Q}_r$, a model is trained to auto-regressively predict a free-form answer $\mathtt{A}_r$ at round r. Specifically, each answer token $\mathtt{a}_r^i$ satisfies

$$\mathtt{a}_r^i = \arg\max_{\mathtt{a}\in\mathcal{V}} \left[\mathbf{p}\left(\mathtt{a}|\mathtt{V}, \mathtt{C}, \mathtt{H}_r\mathtt{Q}_r, \mathtt{A}_r^{<i}\right)\right], \tag{1}$$

where $\mathtt{A}_r^{<i}$ denotes the previously predicted answer tokens and $\mathcal{V}$ the vocabulary. In the rest, we use *"context"* to refer to the concatenation of the history $\mathtt{H}_r$ and the question $\mathtt{Q}_r$.

3.2. Architecture

Overview. As can be seen from Figure 2, our model takes an image/video $\mathtt{V} \in \mathbb{R}^{F\times 3\times H\times W}$ as input, where F is the number of frames and is set to *one* for images, and (H, W) is the re-sized resolution. Then it processes every frame using a pre-trained EVA-CLIP [51] Image Encoder and concatenates every four spatially adjacent visual patches into a single one. Then, a linear layer maps each visual token into a lower dimensional vector $\mathbf{v}$ of dimension D to obtain

$$\mathbf{V} = \begin{bmatrix} \mathbf{v}_1^1 & \mathbf{v}_1^2 & \cdots & \mathbf{v}_1^F \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{v}_P^1 & \mathbf{v}_P^2 & \cdots & \mathbf{v}_P^F \end{bmatrix} \in \mathbb{R}^{F\times P\times D}, \tag{2}$$

where $P = \frac{1}{4}\frac{H\times W}{14^2}$ and D denote the visual input length and the joint hidden dimension, respectively. Thereafter, in stark contrast to previous works [13, 19] that performed spatial and temporal attention in series, our model *separately* performs these operations using the masks $\mathbf{M}^{\text{spa}}$ and $\mathbf{M}^{\text{tmp}}$ as shown in Figure 2 on the visual features V to obtain

$$\mathbf{V}^{\text{spa}} = \text{SA}(\mathbf{V}, \mathbf{M}^{\text{spa}}) \in \mathbb{R}^{(FP)\times D} \tag{3}$$

$$\mathbf{V}^{\text{tmp}} = \text{SA}(\mathbf{V}, \mathbf{M}^{\text{tmp}}) \in \mathbb{R}^{(FP)\times D} \tag{4}$$

$$\mathbf{M}^{\text{spa}}_{m,n}(\mathbf{v}_i^j) = \delta_{nj},\ \mathbf{M}^{\text{tmp}}_{m,n}(\mathbf{v}_i^j) = \delta_{mi} \tag{5}$$

where SA and δ denote self-attention and Kronecker delta.

Subsequently, the textual input in the form of a caption and a context is processed by an embedding layer to obtain $\mathbf{T}^{\text{cap/ctx}} \in \mathbb{R}^{N_{\text{cap/ctx}}\times D}$, where N_{cap} and N_{ctx} are the respective lengths of the caption and context. These visual and textual features form the initial input to the multimodal expert layers which are pre-trained using a combination of matching, contrastive learning, and masked language modeling losses. Finally, they are coupled with a pre-trained LLM and are fine-tuned end-to-end using a generative loss.

Multimodal Expert Layers. These consist of N layers of stacked multi-head self-attention with layer normalization (SA), and *several* modality-specific and *one* modality-agnostic feed-forward networks that we refer to as *experts*. As shown in Figure 2, we propose to use a set of *six* experts denoted as $\{\mathcal{E}_*\}$: *three* of which are vision-specific and *two* are language-specific and are activated in the first L layers. The *remaining* expert $\mathcal{E}_{\text{fus}}$ is the fusion expert and is only activated in the last $(N - L)$ layers and operates on the concatenation of all available modalities (Equation 9). To the best of our knowledge, we propose for the first time to learn the spatial and temporal features using dedicated

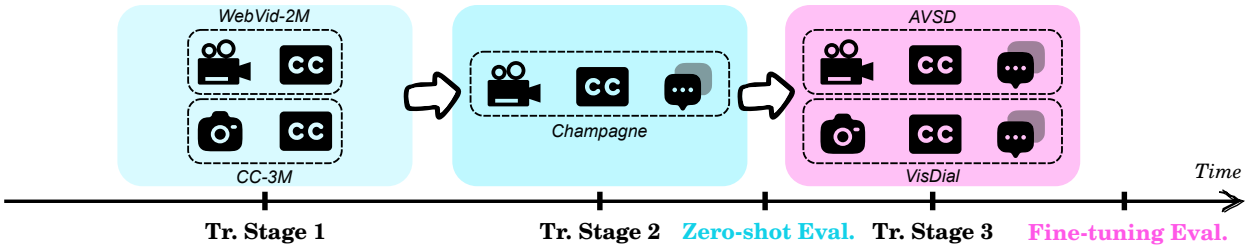

Figure 3. **Overview of the training and evaluation pipeline of V^{2}Dial**. We show the different datasets used to train our model at each stage. Evaluations are conducted on the most popular video and visual dialog datasets of AVSD and VisDial, respectively. (= video data, = image data, = closed / visual captioning data, = dialog data).

experts (i.e., the spatial $\mathcal{E}_{\text{spa}}$ and temporal $\mathcal{E}_{\text{tmp}}$ experts, respectively) as shown in Equation 11. This allows our model to unify video and visual dialog by jointly learning from image and video data. The visual expert $\mathcal{E}_{\text{vis}}$ operates on top of the concatenation of $\mathcal{E}_{\text{spa}}$ and $\mathcal{E}_{\text{tmp}}$ to learn a joint spatial-temporal video representation (Equation 10). Similarly, the textual experts $\mathcal{E}_{\text{cap}}$ and $\mathcal{E}_{\text{ctx}}$ operate on the caption and context embeddings $\mathbf{T}^{\text{cap}}$ and $\mathbf{T}^{\text{ctx}}$ (Equation 12). As seen in Table 1, the availability of the multimodal features depends on the visual input type (i.e., videos vs images) and the training stage. However, without the loss of generality, we can formulate the multimodal expert layers as follows:

$$\mathbf{X}_0 = [\mathbf{V}^{\text{spa}}, \mathbf{V}^{\text{tmp}}, \mathbf{T}^{\text{cap}}, \mathbf{T}^{\text{ctx}}], \tag{6}$$

$$\tilde{\mathbf{X}}_l = [\tilde{\mathbf{V}}_l^{\text{spa}}, \tilde{\mathbf{V}}_l^{\text{tmp}}, \tilde{\mathbf{T}}_l^{\text{cap}}, \tilde{\mathbf{T}}_l^{\text{ctx}}] \tag{7}$$

$$= \text{SA}(\mathbf{X}_{l-1}) + \mathbf{X}_{l-1} \tag{8}$$

$$\mathbf{X}_l = \begin{cases} [\mathbf{V}_l^{\text{vis}}, \mathbf{T}_l^{\text{cap}}, \mathbf{T}_l^{\text{ctx}}] & if\ 1 \le l \le L \\ \mathcal{E}_{\text{fus}}(\tilde{\mathbf{X}}_l) + \tilde{\mathbf{X}}_l & if\ L < l \le N \end{cases}, \tag{9}$$

$$\mathbf{V}_l^{\text{vis}} = \mathcal{E}_{\text{vis}}(\tilde{\mathbf{V}}_l^{\text{vis}}) + \tilde{\mathbf{V}}_l^{\text{vis}},\ \tilde{\mathbf{V}}_l^{\text{vis}} := [\mathbf{V}_l^{\text{spa}}, \mathbf{V}_l^{\text{tmp}}], \tag{10}$$

$$\mathbf{V}_l^{\text{spa}} = \mathcal{E}_{\text{spa}}(\tilde{\mathbf{V}}_l^{\text{spa}}) + \tilde{\mathbf{V}}_l^{\text{spa}},\ \mathbf{V}_l^{\text{tmp}} = \mathcal{E}_{\text{tmp}}(\tilde{\mathbf{V}}_l^{\text{tmp}}) + \tilde{\mathbf{V}}_l^{\text{tmp}}, \tag{11}$$

$$\mathbf{T}_l^{\text{cap}} = \mathcal{E}_{\text{cap}}(\tilde{\mathbf{T}}_l^{\text{cap}}) + \tilde{\mathbf{T}}_l^{\text{cap}},\ \mathbf{T}_l^{\text{ctx}} = \mathcal{E}_{\text{ctx}}(\tilde{\mathbf{T}}_l^{\text{ctx}}) + \tilde{\mathbf{T}}_l^{\text{ctx}}. \tag{12}$$

When dealing with images and non-dialog data, we drop $\mathbf{V}_l^{\text{tmp}}$ and $\mathbf{T}_l^{\text{cap}}$ from the previous equations and deactivate the respective expert.

3.3. Training

3.3.1 Stage 1

In the first stage, we only pre-train the multimodal expert layers, the vision encoder linear layer, and the spatial-temporal attention modules. Since we are the first to suggest learning the spatial and temporal features of videos and images using dedicated experts, we propose to train our model using spatial-temporal contrastive learning (STC) and spatial-temporal matching (STM). In addition, we use the established masked language modeling (MLM), vision-text[1] contrastive learning (VTC), and vision-text matching (VTM) similar to [19, 31, 32].

[1] Vision can either be video or image depending on the dataset.

	Tr. Stage 1	Tr. Stage 2	Tr. Stage 3
Videos	$\mathbf{V}^{\text{spa}}, \mathbf{V}^{\text{tmp}}, \mathbf{T}^{\text{cap}}$	$\mathbf{V}^{\text{spa}}, \mathbf{V}^{\text{tmp}}, \mathbf{T}^{\text{cap}}, \mathbf{T}^{\text{ctx}}$	$\mathbf{V}^{\text{spa}}, \mathbf{V}^{\text{tmp}}, \mathbf{T}^{\text{cap}}, \mathbf{T}^{\text{ctx}}$
Images	$\mathbf{V}^{\text{spa}}, \mathbf{T}^{\text{cap}}$	-	$\mathbf{V}^{\text{spa}}, \mathbf{T}^{\text{cap}}, \mathbf{T}^{\text{ctx}}$

Table 1. Overview of the available features for each training stage and visual input type.

Spatial-Temporal Contrastive Learning aims to better align the spatial and temporal features of video data. To this end, we use output features of the last multi-modal exert layer[2] and learn a cosine similarity function

$$\text{s}(\mathbf{V}^{\text{spa}}, \mathbf{V}^{\text{tmp}}) = \Theta_{\text{spa}}(\mathbf{V}^{\text{spa}})^\top \Theta_{\text{tmp}}(\mathbf{V}^{\text{tmp}}), \tag{13}$$

so that aligned spatial-temporal features result in higher similarity scores, where Θ_* are linear layers that map the features to a normalized lower dimensional vector space. Then, given spatial and temporal feature pairs, we compute the softmax normalized spatial-to-temporal and temporal-to-spatial similarities as

$$p_i^{\text{s2t}}(\mathbf{V}^{\text{spa}}) = \frac{\exp(\tilde{\text{s}}(\mathbf{V}^{\text{spa}}, \mathbf{V}_i^{\text{tmp}})/\tau)}{\sum_{k=1}^{K} \exp(\tilde{\text{s}}(\mathbf{V}^{\text{spa}}, \mathbf{V}_k^{\text{tmp}})/\tau)}, \tag{14}$$

$$p_i^{\text{t2s}}(\mathbf{V}^{\text{tmp}}) = \frac{\exp(\tilde{\text{s}}(\mathbf{V}^{\text{tmp}}, \mathbf{V}_i^{\text{spa}})/\tau)}{\sum_{k=1}^{K} \exp(\tilde{\text{s}}(\mathbf{V}^{\text{tmp}}, \mathbf{V}_k^{\text{spa}})/\tau)}, \tag{15}$$

where τ is learnable temperature parameters, and $\tilde{\text{s}}$ is the maximum value of s as in [32]. Finally, we can compute the loss as the cross-entropy $\mathcal{H}$ between $\mathbf{p}$ and $\mathbf{y}$:

$$\mathcal{L}_{\text{stc}} = \frac{1}{2}\mathbb{E}_{(\mathbf{V}^{\text{spa}}, \mathbf{V}^{\text{tmp}})}\left[\mathcal{H}\left(\mathbf{y}^{\text{s2t}}, \mathbf{p}^{\text{s2t}}\right) + \mathcal{H}\left(\mathbf{y}^{\text{t2s}}, \mathbf{p}^{\text{t2s}}\right)\right], \tag{16}$$

where $\mathbf{y}^{\text{s2t}}$ and $\mathbf{y}^{\text{t2s}}$ are the golden one-hot similarities.

Spatial-Temporal Matching complements STC and teaches the model to distinguish between positive and negative spatial-temporal feature pairs. Specifically, a matched feature pair originates from the same video, whereas an unmatched pair is constructed using negative sampling from a

[2] Index dropped for clarity.

different video. We use a classification token as a proxy of the joint spatial-temporal representations to learn a binary classification problem using the STM loss

$$\mathcal{L}_{\text{stm}} = \mathbb{E}_{(\mathbf{V}^{\text{spa}}, \mathbf{V}^{\text{tmp}})} \left[\mathcal{H}(\mathbf{y}^{\text{stm}}, \mathbf{p}^{\text{stm}}) \right], \tag{17}$$

where $\mathbf{p}^{\text{stm}}$ and $\mathbf{y}^{\text{stm}}$ are the predicted and the ground-truth two-class probabilities, respectively.

We provide more details about the remaining established objectives (i.e., MLM, VTC, VTM) in the supplementary.

3.3.2 Stages 2 & 3

In the subsequent stages, we couple the multimodal expert layers with a pre-trained Flan-T5$_{\texttt{large}}$ [21] via a linear layer. Specifically, Stage 2 aims to align the hidden states of the proposed layers with those of the pre-trained LLM. To this end, we keep the LLM weights frozen and train the whole architecture end-to-end using the generative loss (i.e., next token prediction) on large scale video dialog data [3], i.e.,

$$\mathcal{L}_{\text{gen}} = \mathbb{E}_{\mathbf{X}^{\text{gen}}} \left[\mathcal{H}(\mathbf{y}^{\text{gen}}_{\rightarrow}, \mathbf{p}^{\text{gen}}) \right], \tag{18}$$

$$\mathbf{X}^{\text{gen}} = \Theta_{\text{gen}} \left(\text{LLM}_{\text{dec}}([\mathbf{X}^{\text{enc}}, \mathbf{T}^{\text{ans}}]) \right), \tag{19}$$

where $\mathbf{X}^{\text{enc}}$, $\mathbf{T}^{\text{ans}}$ and Θ_{gen} are the LLM encoder output, the answer token embeddings, and a linear layer that maps the features to the vocabulary space, respectively. $\mathbf{y}^{\text{gen}}_{\rightarrow}$ and $\mathbf{p}^{\text{gen}}$ denote the right-shifted ground-truth answer tokens and the predicted text token probabilities. Finally, in Stage 3, we unfreeze the LLM weights and fine-tune our model end-to-end on the downstream tasks of video and visual dialog using the same generative loss.

4. Experiments

4.1. Datasets

As shown in Figure 3, we simultaneously use the video and image captioning datasets of WebVid-2M [10] and CC-3M [49] to pre-train the multimodal expert layers in Stage 1. Then in the *second* stage, we use 25% of the recent large-scale video dialog dataset Champagne [26] before performing *zero-shot* evaluation on the widely used video and visual dialog datasets of AVSD [5] and VisDial [22], respectively. Finally, in the *third* stage, we perform a domain shift evaluation based on different combinations of AVSD and VisDial to quantify whether and to what extent these seemingly similar benchmarks benefit from each other in both *zero-shot* and *fine-tuning* evaluation settings.

[3] The weights of $\mathcal{E}_{\text{ctx}}$ are initialized with those of $\mathcal{E}_{\text{cap}}$ from Stage 1.

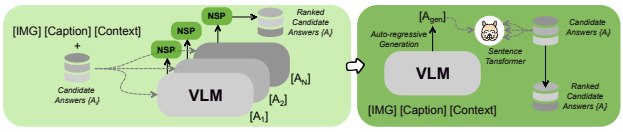

Figure 4. Instead of training a dedicated NSP head, we propose a ranking scheme based on the cosine similarity of the candidate answers' embeddings with the respect to those of the generated ones. We used RoBERTa$_{\texttt{large}}$ [38] and OpenAI Text Embedding-3$_{\texttt{large}}$ to generate these embeddings.

4.2. Evaluation Metrics

We use the established official metrics for each dataset to fairly benchmark **V^{2}Dial** with previous works. Specifically, for all *three* AVSD datasets, we use BLEU (**B-n**) [45], ROUGE-L (**R**) [36], METEOR (**M**) [11], and CIDEr (**C**) [53]. Whereas for VisDial, we use the retrieval metrics of recall (**R@k**), mean reciprocal rank (**MRR**), and normalized discounted cumulative gain (**NDCG**). However, since we are jointly tackling both tasks with a fully generative model, we propose to rank the VisDial candidate answers by means of cosine similarity with respect to the generated answer using the embdeddings of a pre-trained sentence transformer (i.e. RoBERTa [38] and OpenAI Text Embedding-3). We posit that this approach is more natural, caters to the current advances in generative models, and appropriately captures the semantic similarities between the generated and the candidate answers. In addition, it allows for a seamless unification of AVSD and VisDial without any training or architectural modifications. As shown in Figure 4, our proposed adaptation does *not* alter the computation of the sparse metrics itself and *only rethinks* the ranking of the candidate answers allowing for a fair comparison with previous works.

4.3. Experimental Setup

In the first stage, we trained our model for a maximum of *ten* epochs and applied early stopping based on a validation split to select the best checkpoint. In the subsequent stages, we trained it for up to *three* and *twelve* epochs, respectively. In all stages, we used the AdamW [39] optimizer with the default parameters and a weight decay value of 0.01. Furthermore, we applied a linear learning rate schedule with warm-up and minimum and base values of $5e{-}5$ and $1e{-}4$, respectively. We conducted our experiments on a cluster consisting of *eight* A100 GPUs.

4.4. Zero-shot Evaluation

AVSD. We first assessed **V^{2}Dial** in a zero-shot setting on AVSD. This is in stark contrast to previous models that were exclusively evaluated in a fine-tuning setting. We instead advocate for complementing the fine-tuning evaluation setting with a zero-shot one, as it results in a more rigorous

Model	AVSD-DSTC10							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	B-1	B-2	B-3	B-4	M	R	C
♦MoE-LLAVA$_{arXiv'24}$ [35]	35.8	18.9	10.1	5.9	15.4	27.1	12.8	39.8	23.9	15.2	10.1	18.7	32.2	23.7	44.7	29.1	19.6	13.8	21.8	37.3	33.2
♦MiniGPT4-video$_{CVPR'24}$ [9]	37.9	19.9	11.3	6.8	16.2	28.7	17.7	34.8	17.6	9.7	5.8	15.8	26.3	13.3	37.8	21.2	12.7	8.2	18.4	30.2	17.7
♦Video-ChatGPT$_{ACL'24}$ [40]	24.5	14.7	8.8	5.4	16.7	25.2	3.9	25.5	16.0	10.1	6.4	18.4	27.1	9.1	28.5	18.5	11.8	7.6	20.4	32.1	19.1
♦MST-MIXER$_{ECCV'24}$ [3]	0.1	0.0	0.0	0.0	3.1	6.8	3.0	0.2	0.1	0.1	0.0	3.3	7.1	4.3	0.2	0.1	0.0	0.0	3.4	6.9	4.6
♦VideoChat2$_{CVPR'24}$ [33]	42.5	25.9	16.0	10.3	18.7	33.1	25.4	43.9	28.1	18.5	12.6	**20.8**	34.5	29.2	46.7	31.1	20.9	14.4	**22.9**	37.6	31.4
V^{2}Dial	**54.6**	**34.8**	**24.0**	**17.2**	**19.7**	**38.3**	**53.8**	**53.2**	**33.8**	**23.5**	**16.7**	18.8	**37.7**	**49.7**	**55.5**	**36.7**	**26.2**	**18.7**	20.0	**39.2**	**50.8**

Table 2. Zero-shot performance comparison on AVSD-DSTC10, AVSD-DSTC8 and AVSD-DSTC7. Best and second-best performances are in **bold** and underlined. ♦ indicates that we evaluated the model. (**B-n** = BLEU-n, **M** = METEOR, **R** = ROUGE-L, **C** = CIDEr).

Model	Sent. Trans.	R@1	R@5	R@10	MRR	NDCG
FROMAGe$_{ICML'23}$ [27]	*n/a*	17.6	20.1	25.1	22.0	16.5
ESPER$_{CVPR'23}$ [62]		14.6	–	–	25.7	22.3
Champagne$_{ICCV'23}$ [26]		–	–	–	–	25.5
♦MoE-LLAVA$_{arXiv'24}$ [35]	*RoBERTa*	10.6	25.4	36.4	19.6	26.7
♦MiniGPT-video$_{CVPR'24}$ [9]		7.4	17.4	26.5	14.6	23.2
♦Video-ChatGPT$_{ACL'24}$ [40]		10.0	22.5	31.5	18.1	24.8
♦MST-MIXER$_{ECCV'24}$ [3]		18.2	22.1	25.7	21.9	24.6
♦VideoChat2$_{CVPR'24}$ [33]		12.7	29.0	39.9	22.3	30.9
V^{2}Dial	*RoBERTa*	20.0	30.2	39.3	26.9	**33.3**
	OpenAI TE-3	**22.1**	**41.2**	**48.1**	**32.7**	32.0

Table 3. Zero-shot performance comparison on the VisDial v1.0 val split. OpenAI TE-3 = OpenAI Text Embedding-3$_{\texttt{large}}$.

and challenging testbed for the proposed models. To this end, we establish; to the best of our knowledge; the *first* zero-shot benchmark comparison on AVSD comprised of recent capable multimodal models. As can be seen from Table 2, our model outperforms all baselines by a considerable margin across 6/7 metrics of AVSD-DSTC8 and AVSD-DSTC7. On the more recent and challenging version of the benchmark (i.e. AVSD-DSTC10), **V^{2}Dial** ranks first across all metrics. For instance, it more than doubles the CIDEr score compared to VideoChat2 [33].

VisDial. Additionally, we assessed the same model checkpoint on VisDial v1.0. As can be seen from Table 3, **V^{2}Dial** managed to outperform previous models such as FROMAGe [27] by a considerable margin across all metrics of the dataset. In addition, it outperformed Champagne [26] that was trained on x4 more dialog data by 7.8 absolute NDCG points. Furthermore, our model outperformed the more recent baselines of the previous section on 4/5 metrics, underlining it capability of dealing with both video and image input data types. Finally, replacing the sentence embeddings generated by RoBERTa$_{\texttt{large}}$ with those from OpenAI Text Emedding-3 improved the external ranking of the candidate answers and resulted in higher scores across all metrics, as can be seen in the last row of Table 3.

4.5. Fine-tuning Evaluation

AVSD. Similar to almost all previous works on AVSD, we assessed **V^{2}Dial** in a fine-tuning setting on all *three* benchmarks of the dataset. As can be seen from Table 4, our model managed to maintain it competitiveness ahead of recent models and outperformed them on the latest and most challenging AVSD-DSTC10 benchmarks across all evaluation metrics. For instance, it lifted CIDEr by over 6 absolute points compared to the second-best model. Furthermore, our model managed to maintain an on par performance with the state of the art on AVSD-DSTC8 and AVSD-DSTC7. As shown in Table 4, **V^{2}Dial** increased their respective CIDEr scores by over 2 and 3 absolute points compared to the second-best model.

VisDial. Finally, we fine-tuned our model and MST-MIXER [3] that had the closest AVSD performance on Visdial v1.0 using the same fully-generative approach. As can be seen from Table 5, **V^{2}Dial** managed to outperform all previous models on the strictest metric of the dataset by achieving a R@1 score of 44.2. However, when using OpenAI Text Embedding-3 our model managed to increase the R@1 and MRR scores to 44.9 and 52.4, respectively, thereby setting new state-of-the-art results. As expected and due to the more challenging aspect of a tackling VisDial as a fully generative task, our model performed slightly worse than the previous fine-tuned models on the remaining metrics of the dataset. However, when comparing our model with MST-MIXER that was trained using the same paradigm (i.e. the last two rows of Table 5), we can see that our model outperformed it across 4/5 metrics of the task and scored almost equally on NDCG.

4.6. Domain Shift Evaluation

Zero-shot setting. First, we fine-tuned our model's checkpoint from Stage 2 on AVSD and zero-shot evaluated it on VisDial. As can be seen from the second section of Table 6, our model's performance was lifted by a considerable margin across most metrics. Notably, the NDCG score improved by 9 absolute points compared to the results of Table 3. Then, we replicated the same experiment on AVSD after having fine-tuned the model on VisDial. Interestingly, our model's performance deteriorated across all metrics of the benchmark. This behavior could be explained by the nature of both datasets. Whereas AVSD encourages the model

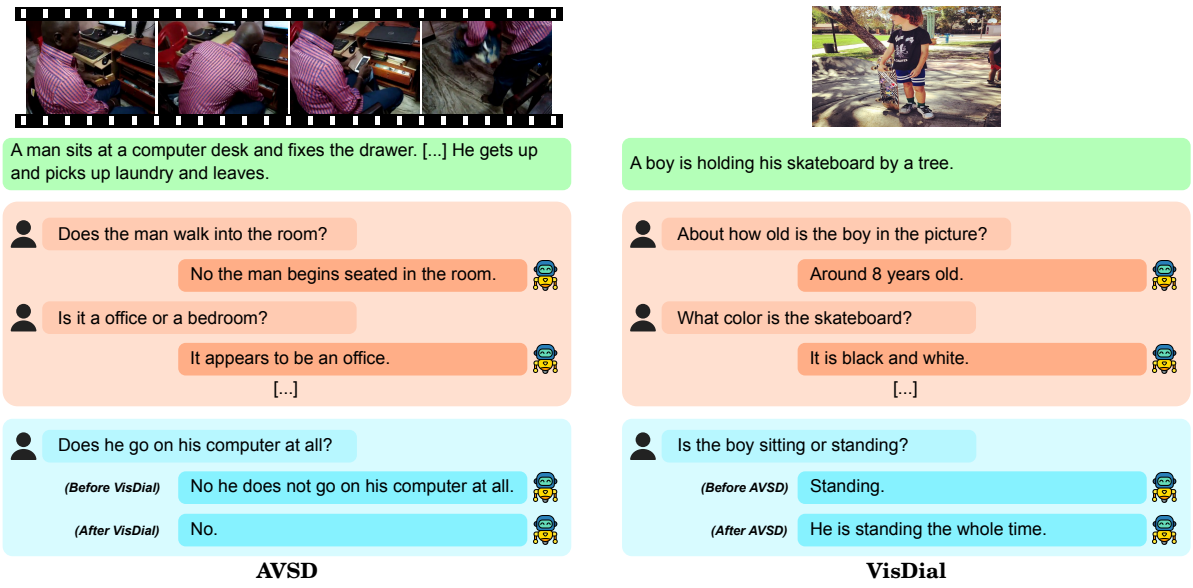

Figure 5. Zero-shot qualitative examples of **V^{2}Dial** before and after fine-tuning on VisDial and AVSD. The former teaches the model to answer question with brief responses whereas the latter teaches it to produce longer and more elaborate answers.

Model	AVSD-DSTC10							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	B-1	B-2	B-3	B-4	M	R	C
PDC$_{ICLR'21}$ [29]	–	–	–	–	–	–	–	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$_{EMNLP'22}$ [60]	–	–	–	–	–	–	–	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$_{TASLP'23}$ [18]	69.3	55.6	45.0	36.9	24.9	53.6	91.2	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$_{arXiv'24}$ [20]	50.2	35.0	24.9	18.1	21.8	42.8	57.5	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$_{ECCV'24}$ [3]	69.7	57.1	47.2	39.5	25.1	54.0	96.9	**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
V^{2}Dial	**70.7**	**58.2**	**48.2**	**40.3**	**26.0**	**55.4**	**103.3**	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 4. Fine-tuning performance comparison on AVSD-DSTC10, AVSD-DSTC8 and AVSD-DSTC7. VideoLLAMA 2 [20] was trained on AVSD amongst other datasets. Additional model comparisons can be found in the supplementary material.

Model	Sent. Trans.	R@1	R@5	R@10	MRR	NDCG
LTMI$_{ECCV'20}$ [44]		40.4	61.6	**69.7**	50.7	**63.5**
LTMI-LG$_{EMNLP'21}$ [17]		41.3	61.6	69.0	51.3	63.2
GoG$_{ACL'21}$ [16]	*n/a*	41.2	**61.8**	69.4	51.3	62.6
UTC$_{CVPR'22}$ [15]		41.3	59.8	66.3	50.6	61.0
Champagne$_{ICCV'23}$ [26]		–	–	–	–	62.5
♣MST-MIXER$_{ECCV'24}$ [3]	*RoBERTa*	42.2	51.6	57.8	47.7	52.5
V^{2}Dial	*RoBERTa*	45.4	54.7	61.1	50.9	54.0
	OpenAI TE-3	**46.1**	59.3	65.7	**53.2**	53.1

Table 5. Fine-tuning performance comparison on the VisDial v1.0 val split. ♣ indicates that we trained and evaluated the model.

to produce long and elaborate responses, VisDial teaches it to produce brief answers instead, which diminishes its performance on the language generation metrics. The qualitative examples of Figure 5 clearly illustrate this phenomenon on both datasets.

Fine-tuning Setting. We first experimented with a curriculum learning strategy where we used one dataset for pre-training before finally fine-tuning on the other. As can be seen from the last section of Table 6, this training paradigm resulted in performance drops on both datasets compared to Table 4 and Table 5 where the model was only trained on the data of the respective benchmark. This indicates that the converged model's weights on one dataset do not offer a good initialization for training on the remaining one. Allowed by our model design that can jointly handle video and image input data, we finally fine-tuned one single model on both datasets simultaneously. As seen from the last row of Table 6, this resulted in the best joint performance of our model across the two datasets. Although the results on AVSD slightly dropped compared to Table 4, our model lifted its performance on VisDial by a considerable margin. This could largely be attributed to the same previous observation, as training on VisDial incentivizes our model to shorten its responses on AVSD.

4.7. Expert Swapping

In order to validate the specialization of each expert, we conducted a swapping experiment where we routed some features through *inadequate* experts. We first swapped experts of the same modality (i.e., experts operating on vision

Fine-tuning data			AVSD-DSTC10				AVSD-DSTC8				AVSD-DSTC7				VisDial			
AVSD		VisDial	B-1	M	R	C	B-1	M	R	C	B-1	M	R	C	R@1	R@5	R@10	NDCG
✗		✗	*Zero-shot (from Table 2 and Table 3)*															
			54.6	19.7	38.3	53.8	53.2	18.8	37.7	49.7	55.5	20.0	39.2	50.8	20.0	30.2	39.3	33.3
✓		✗	*Fine-tuning (from Table 4)*												*Zero-shot*			
			70.7	26.0	55.4	103.3	76.8	30.4	62.1	135.7	78.9	31.2	62.3	139.8	12.8	36.7	50.8	42.3
✗		✓	*Zero-shot*												*Fine-tuning (from Table 5)*			
			11.5	6.8	20.1	14.6	11.5	7.3	20.7	20.9	7.9	6.2	17.4	18.2	44.2	53.3	59.5	52.3
✓	→	✓	*Fine-tuning*															
			–	–	–	–	–	–	–	–	–	–	–	–	42.2	50.1	56.3	51.3
✓	←	✓	69.6	25.7	55.0	100.5	75.9	29.8	61.4	132.1	77.6	30.4	61.5	134.5	–	–	–	–
✓	&	✓	69.3	25.4	54.8	99.9	75.1	29.3	61.1	130.0	77.3	30.0	61.7	134.5	45.4	54.7	61.1	54.0

Table 6. Domain shift evaluation between the respective *most prominent* video and visual dialog datasets of AVSD and VisDial. □ → △ means that the model was pre-trained on dataset □ before fine-tuning on dataset △.

Expert Swapping			AVSD-DSTC7				VisDial	
			B-1	M	R	C	R@1	NDCG
Original			78.9	31.2	62.3	139.8	44.2	52.3
Swapping experts of the same modality (vision / language)								
$\mathcal{E}_{\text{spa}}$	↔	$\mathcal{E}_{\text{tmp}}$	77.0	29.5	61.2	133.7	–	–
$\mathcal{E}_{\text{cap}}$	↔	$\mathcal{E}_{\text{ctx}}$	76.1	29.6	60.3	131.1	42.8	51.9
Swapping experts of different modalities								
$\mathcal{E}_{\text{spa}}$	↔	$\mathcal{E}_{\text{cap}}$	28.5	10.7	22.0	10.1	35.7	45.3
$\mathcal{E}_{\text{tmp}}$	↔	$\mathcal{E}_{\text{ctx}}$					–	–
$\mathcal{E}_{\text{spa}}$	↔	$\mathcal{E}_{\text{ctx}}$	34.4	12	25.4	11.7	32.4	42.9
$\mathcal{E}_{\text{tmp}}$	↔	$\mathcal{E}_{\text{cap}}$					–	–

Table 7. Expert swapping results. $\mathcal{E}_{□} \leftrightarrow \mathcal{E}_{△}$ means that the □ features are *inadequately* routed *at test time* through $\mathcal{E}_{△}$ and vice versa. The other experts remain unchanged.

or language data). As shown in Table 7, this resulted in performance drops across all metrics of both datasets, indicating that experts of the same modality are able to capture the semantic nuances of the data they specialize on. More interestingly, the performance of our model dropped more significantly when swapping experts of different modalities, as seen from the last section of Table 7. This showcases their ability to adjust to the nature of the data they process and to capture its modality specific features.

4.8. Ablation Study

Effect of Pre-training Data. We trained two versions of our model, where one was only pre-trained on Stage 1 using WebVid-2M & CC-3M and the other only on Stage 2 with a subset of Champagne. As can be seen from the middle section of Table 8, our model witnessed a comparable drop in performance compared to the full model. This underlines the equal importance of these proposed training stages to the joint down-stream performance on AVSD and VisDial. We did not conduct ablations using either WebVid-2M or CC-3M in Stage 1 as this was sufficiently explored by other recent works [19] that showed the benefit of pre-training on both image and video data.

Effect of Pre-training Objectives & Model Design. Then, we trained a version of our model without $\mathcal{L}_{\text{stc}}$ and $\mathcal{L}_{\text{stm}}$ in Stage 1 using the same schedule and training data as our full model. As shown in the *fourth* row of Table 8, this ablated version suffered a performance drop not only in AVSD but also in VisDial. This indicates that these losses improve not only the temporal capabilities of our model but also its spatial ones. Then, we trained a version that sequentially applies spatial and temporal attention, as in [13, 19]. Since this version does not have separate spatial-temporal experts, we also omitted the previous two objectives. As seen in the penultimate row of Table 8, this version underperformed our full model on both datasets, showcasing the effectiveness of our approach. Finally, we trained a version without all the expert layers. As shown in the last row, its performance dropped compared to our full model and performed the worst on VisDial.

Model Ablations	AVSD-DSTC7				VisDial	
	B-1	M	R	C	R@1	NDCG
Full	78.9	31.2	62.3	139.8	44.2	52.3
w/o Tr. Stage 1	76.9	30.0	61.4	134.0	34.5	44.6
w/o Tr. Stage 2	77.8	30.7	61.7	134.7	32.6	44.1
w/o $\mathcal{L}_{\text{stc}}$ & $\mathcal{L}_{\text{stm}}$	77.2	29.9	61.1	133.2	33.1	44.6
w/o separate $\mathcal{E}_{\text{spa}}$ & $\mathcal{E}_{\text{tmp}}$	77.0	30.1	61.1	133.8	32.9	43.8
w/o experts $\{\mathcal{E}_*\}$	77.5	30.0	61.4	134.8	30.6	42.2

Table 8. Ablation results.

5. Conclusion

We presented **V^{2}Dial** – a model that can jointly tackle video and visual conversational tasks using a multimodal expert-based approach that; *for the first time*, disentangles the learning of the spatial and temporal features of images and videos using two separate experts. Extensive evaluation on the respective widely used video and visual dialog datasets of AVSD and VisDial show that our model achieves new state-of-the-art zero-shot and fine-tuning performance. Finally, we conducted the *first* domain shift evaluation of AVSD and VisDial and provided insights on how to optimally leverage their respective training data.

Acknowledgment

This work was partially funded by a LOEWE-Start-Professur (LOEWE/4b//519/05.01.002-(0006)/94), LOEWE-Spitzen-Professur (LOEWE/4a//519/05.00.002-(0010)/93), and an Alexander von Humboldt Professorship in Multimodal Reliable AI, sponsored by Germany's Federal Ministry for Education and Research.

References

[1] Adnen Abdessaied, Mihai Bâce, and Andreas Bulling. Neuro-Symbolic Visual Dialog. In *Proceedings of the International Conference on Computational Linguistics (COLING)*, 2022. 1

[2] Adnen Abdessaied, Manuel Hochmeister, and Andreas Bulling. OLViT: Multi-Modal State Tracking via Attention-Based Embeddings for Video-Grounded Dialog. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)*, 2024. 2

[3] 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, 6, 7

[4] Adnen Abdessaied, Lei Shi, and Andreas Bulling. VD-GR: Boosting Visual Dialog With Cascaded Spatial-Temporal Multi-Modal Graphs. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024. 1, 2

[5] Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. Audio visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 1, 2, 5

[6] 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

[7] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1

[8] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2015. 1, 2

[9] Kirolos Ataallah, Xiaoqian Shen, Eslam Abdelrahman, Essam Sleiman, Deyao Zhu, Jian Ding, and Mohamed Elhoseiny. MiniGPT4-video: Advancing multimodal llms for video understanding with interleaved visual-textual tokens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2024. 6

[10] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1, 5

[11] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 2005. 5

[12] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. VLMo: Unified vision-language pre-training with mixture-of-modality-experts. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1, 2

[13] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is Space-Time Attention All You Need for Video Understanding? In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021. 3, 8

[14] Soravit Changpinyo, Piyush Sharma, and Nan Ding andRadu Soricut. Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[15] Cheng Chen, Yudong Zhu, Zhenshan Tan, Qingrong Cheng, Xin Jiang, Qun Liu, and Xiaodong Gu. UTC: A Unified Transformer with Inter-Task Contrastive Learning for Visual Dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 7

[16] Feilong Chen, Xiuyi Chen, Fandong Meng, Peng Li, and Jie Zhou. GoG: Relation-aware graph-over-graph network for visual dialog. In *Findings of the Association for Computational Linguistics: ACL*, 2021. 1, 2, 7

[17] Feilong Chen, Xiuyi Chen, Can Xu, and Daxin Jiang. Learning to ground visual objects for visual dialog. In *Findings of the Association for Computational Linguistics: EMNLP*, 2021. 7

[18] 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, 7

[19] 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. 3, 4, 8

[20] 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, 7

[21] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *Journal of Machine Learning Research (JMLR)*, 2024. 5

[22] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. Visual Dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 2, 5

[23] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019. 1

[24] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020. 1

[25] Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. GLaM: Efficient scaling of language models with mixture-of-experts. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2022. 2

[26] Seungju Han, Jack Hessel, Nouha Dziri, Yejin Choi, and Youngjae Yu. Champagne: Learning real-world conversation from large-scale web videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 5, 6, 7

[27] Jing Yu Koh, Ruslan Salakhutdinov, and Daniel Fried. Grounding language models to images for multimodal inputs and outputs. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023. 6

[28] 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

[29] 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, 7

[30] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020. 2

[31] 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. 4

[32] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023. 1, 4

[33] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. MVBench: A Comprehensive Multi-modal Video Understanding Benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 6

[34] Zekang Li, Zongjia Li, Jinchao Zhang, Yang Feng, and Jie Zhou. Bridging text and video: A universal multimodal transformer for audio-visual scene-aware dialog. *Transactions on Audio, Speech, and Language Processing*, 2021. 2

[35] Bin Lin, Zhenyu Tang, Yang Ye, Jiaxi Cui, Bin Zhu, Peng Jin, Jinfa Huang, Junwu Zhang, Yatian Pang, Munan Ning, and Li Yuan. MoE-LLaVA: Mixture of Experts for Large Vision-Language Models. *arXiv preprint arXiv:2401.15947*, 2024. 2, 6

[36] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 2004. 5

[37] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 1

[38] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv 1907.11692*, 2019. 5

[39] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019. 5

[40] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2024. 6

[41] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 1

[42] Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[43] Basil Mustafa, Carlos Riquelme, Joan Puigcerver, Rodolphe Jenatton, and Neil Houlsby. Multimodal Contrastive Learning with LIMoE: the Language-Image Mixture of Experts. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[44] Van Quang Nguyen, Masanori Suganuma, and Takayuki Okatani. Efficient Attention Mechanism for Visual Dialog that Can Handle All the Interactions Between Multiple Inputs. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1, 7

[45] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine

translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002. 5

[46] Hoang-Anh Pham, Thao Minh Le, Vuong Le, Tu Minh Phuong, and Truyen Tran. Video Dialog as Conversation about Objects Living in Space-Time. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 2

[47] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 2019. 2

[48] Ankit P. 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. 1

[49] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018. 5

[50] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. FLAVA: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[51] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. EVA-CLIP: Improved Training Techniques for CLIP at Scale. *arXiv preprint arXiv:2303.15389*, 2023. 3

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 1

[53] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based Image Description Evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 5

[54] Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for vision and vision-language tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2

[55] Yue Wang, Shafiq R. Joty, Michael R. Lyu, Irwin King, Caiming Xiong, and Steven C. H. Hoi. VD-BERT: A Unified Vision and Dialog Transformer with BERT. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020. 2

[56] Zihao Wang, Junli Wang, and Changjun Jiang. Unified multimodal model with unlikelihood training for visual dialog. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2022. 1, 2

[57] Ronald J Williams and David Zipser. A learning algorithm for continually running fully recurrent neural networks. *Neural Comput.*, 1989. 2

[58] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the ACM International Conference on Multimedia (ACM MM)*, 2017. 1, 2

[59] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[60] 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, 7

[61] Xintong Yu, Hongming Zhang, Ruixin Hong, Yangqiu Song, and Changshui Zhang. VD-PCR: Improving visual dialog with pronoun coreference resolution. *Pattern Recognition*, 2022. 2

[62] Youngjae Yu, Jiwan Chung, Heeseung Yun, Jack Hessel, Jae Sung Park, Ximing Lu, Rowan Zellers, Prithviraj Ammanabrolu, Ronan Le Bras, Gunhee Kim, et al. Fusing pretrained language models with multimodal prompts through reinforcement learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 6

[63] Zexuan Zhong, Mengzhou Xia, Danqi Chen, and Mike Lewis. Lory: Fully Differentiable Mixture-of-Experts for Autoregressive Language Model Pre-training. In *Proceedings of the Conference on Language Modeling (COLM)*, 2024. 2

[64] Tong Zhu, Xiaoye Qu, Daize Dong, Jiacheng Ruan, Jingqi Tong, Conghui He, and Yu Cheng. Llama-moe: Building mixture-of-experts from llama with continual pre-training. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024. 2