The Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)

Structure-Aware Multimodal Sequential Learning for Visual Dialog

Young-Jin Kim[1]*, Min-Jun Kim[1]*, Kyunghwan An[2], Jinwoo Ahn[1],
Jaeseok Kim[3], Yu-Jung Heo[3], Du-Seong Chang[3], Eun-Sol Kim[1,2,4]

[1] Department of Artificial Intelligence Application, Hanyang University, South Korea
[2] Department of Artificial Intelligence, Hanyang University, South Korea
[3] KT Corporation
[4] Department of Computer Science, Hanyang University, South Korea
eunsolkim@hanyang.ac.kr

Abstract

With the ability to collect vast amounts of image and natural language data from the web, there has been a remarkable advancement in Large-scale Language Models (LLMs). This progress has led to the emergence of chatbots and dialogue systems capable of fluent conversations with humans. As the variety of devices enabling interactions between humans and agents expands, and the performance of text-based dialogue systems improves, there has been recently proposed research on visual dialog. However, visual dialog requires understanding sequences of pairs consisting of images and sentences, making it challenging to gather sufficient data for training large-scale models from the web. In this paper, we propose a new multimodal learning method leveraging existing large-scale models designed for each modality, to enable model training for visual dialog with small visual dialog datasets. The key ideas of our approach are: 1) storing the history or context during the progression of visual dialog in the form of spatiotemporal graphs, and 2) introducing small modulation blocks between modality-specific models and the graphs to align the semantic spaces. For implementation, we introduce a novel structure-aware cross-attention method, which retrieves relevant image and text knowledge for utterance generation from the pretrained models. For experiments, we achieved a new state-of-the-art performance on three visual dialog datasets, including the most challenging one COMET.

Introduction

With the emergence of large-scale language and vision models such as GPT-X (Radford et al. 2018, 2019; Brown et al. 2020), LLaMA (Touvron et al. 2023), ViT (Dosovitskiy et al. 2020), pretraining and zero-shot transfer learning paradigm has been proposed where models are trained on massive datasets containing a plethora of parameters and tested across various downstream tasks without additional finetuning. By virtue of the huge number of datasets encompassing diverse topics and concepts collected from the web, large-scale language and vision models have acquired the generalization ability to show remarkable performances on numerous tasks, even unseen problems. Compared to modality-specific large-scale training (*i.e.* image-only and text-only), it is challenging to gather massive amounts of data for well-aligned image-text pairs, and there is a disadvantage of noise in such data, making the learning process difficult. In this paper, we aim to develop a large-scale vision-language multimodal model where image-text pair data are scarce, such as visual dialog tasks. In particular, we propose a new multimodal learning method leveraging existing image-only and text-only models.

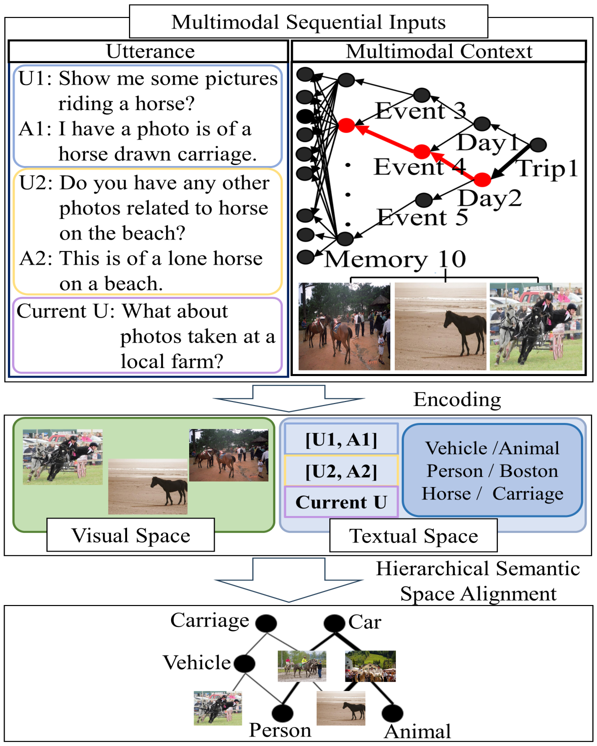
U1: Show me some pictures riding a horse?
A1: I have a photo is of a horse drawn carriage.
U2: Do you have any other photos related to horse on the beach?
A2: This is of a lone horse on a beach.
Current U: What about photos taken at a local farm?

Figure 1: High-level concept of our structure-aware alignment. In the multimodal context, memories having multimodal information compose a structure. By considering this spatiotemporal-hierarchical structure, representations from multimodal context can be aligned and relevent informations for each turn can be retrieved.

*These authors contributed equally.

Recently, pioneering work such as BLIP (Li et al. 2022), and BLIP-2 (Li et al. 2023) have adopted a small-sized transformer as a bridge for modulation between image-only and text-only networks. Those methods learn the correspondence between a single image and text sentence pair through transformers, achieving remarkable performance in tasks such as visual question answering and image captioning. Unlike reasoning tasks for a single image-text pair, multimodal tasks where images and text are given as sequences (such as visual dialog) present a greater challenge. In such tasks, aligning not only the semantic space between images and text but also the temporal relationships within the sequences is necessary. In this paper, we introduce a novel algorithm that leverages large-scale pretrained vision and language models to learn multimodal sequence pairs.

The two main ideas of the suggested method are 1) to leverage the representation power of the pretrained large-scale vision and language models and 2) to align the semantic spaces of two models with modulation blocks. To achieve the intricate alignment between multimodal sequences, we adopt transformer architectures with a novel cross-attention method as modulation blocks. We propose a method to represent the hierarchical semantic structure and temporal relationships in multimodal sequences using a graph, and to align the semantic structures between the sequences using the graphs in cross-attention modules.

For experiments, we evaluate our method using three different visual dialog datasets, which are COMET, VisDial, and MNIST Dialog. We achieve two new state-of-the-art performances from these experiments. One thing we should note is that all experimental results do not require any additional dataset and training methods.

Our contribution can be summarized as follows.

- We introduce a new multimodal sequential learning method that can effectively leverage pretrained vision and language models.
- To consider the inherent spatiotemporal semantic structure within the multimodal sequences, we introduce new structure-aware retrieval-augmented modulation blocks.
- For the most challenging multimodal tasks, visual dialog tasks, we achieve new state-of-the-art performances.

Related Work

Large-scale Models for Multimodal Learning

Large-scale multimodal models (Su et al. 2019; Chen et al. 2020b; Yu et al. 2021; Zellers et al. 2022; Kim, Son, and Kim 2021; Lu et al. 2019; Tan and Bansal 2019; Li et al. 2019; Radford et al. 2021; Li et al. 2020) stand out for their proficiency in seamlessly integrating visual and textual data. For instance, ViLBERT (Lu et al. 2019) is a notable example that employs a two-stream architecture, processing visual and textual information separately using co-attention mechanisms to allow for joint reasoning over both modalities. These models leverage huge amounts of unlabeled data to learn joint representations of texts and images or videos. Typically pretrained on vast datasets with paired image-text data, they use attention mechanisms to capture intricate relationships between modalities and can be finetuned for specific tasks.

However, their performance is closely tied to the quality of pretraining data, and their computational intensity, coupled with concerns about generalization and interpretability, presents challenges that remain areas of active exploration. Furthermore, in a sustained reasoning process, like visual dialog tasks, it's crucial to formulate answers by understanding the semantic essence of large-scale image or text models, while also factoring in information from earlier dialog turns. But previous studies (Li et al. 2022, 2023) have primarily focused on processing a single image and its corresponding text sentence using transformers. In our study, our structure-based cross-attention method signifies a shift towards more sophisticated alignment techniques. By representing the sequence of context as a spatiotemporal graph, it ensures effective information retrieval and alignment between modalities.

Visual Dialog

Visual dialog, an emerging research domain, delves into generating conversational responses intricately linked to images. This field, while reminiscent of Visual Question Answering (VQA), places a heightened emphasis on the context extracted from sequential dialog turns. Initial studies, such as the one (Das et al. 2017), predominantly merged CNNs with RNNs, capturing both the image and the sequence of questions. The attention mechanism-based models (Lu et al. 2018; Park et al. 2021; Zhang et al. 2022a; Gan et al. 2019), especially the co-attention model, have been proposed, allowing simultaneous focus on specific image regions and relevant textual parts of the question. This dual focus has enriched the understanding of image-text dynamics. Furthermore, the alignment-based approach model (Chen et al. 2022) has shown promise in explicitly aligning visual concepts with textual semantics via unsupervised and pseudo-supervised vision-language alignment. Another intriguing approach (Chen et al. 2021; Guo et al. 2020; Zhang et al. 2022b; Zheng et al. 2019) is the graph-based representation suitable for the composite scenario of dialog history and image, which offers a structured way to understand relationships within an image. Diverging from these methodologies, our model leverages a large multi-modal hierarchical context. As the dialog progresses, at each turn, a model must retrieve requisite information from this context to answer. Our model's distinctiveness lies in its ability to adaptively reference only the relevant context for answer generation, ensuring efficiency and producing contextually rich and precise answers even as the context expands with each dialog turn.

Structure-Aware Retrieval-Augmented Learning

In this section, we introduce a novel algorithm that leverages large-scale pretrained vision and language models to learn multimodal sequence pairs. The suggested algorithm mainly consists of two parts: 1) a pretrained large-scale vision and language model, and 2) modulation blocks between

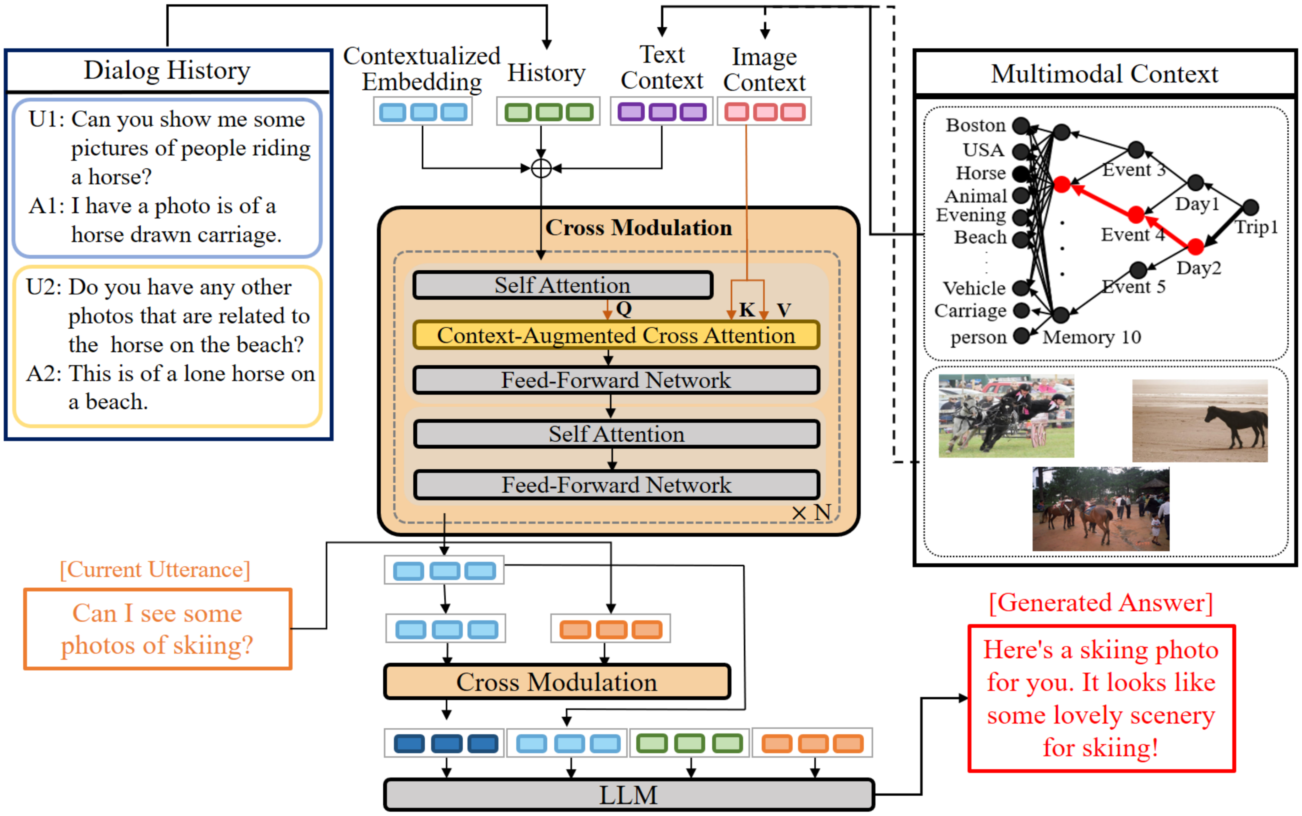
Dialog History
U1: Can you show me some pictures of people riding a horse?
A1: I have a photo is of a horse drawn carriage.
U2: Do you have any other photos that are related to the horse on the beach?
A2: This is of a lone horse on a beach.
Contextualized Embedding
History
Text Context
Image Context
Multimodal Context
Boston
USA
Horse
Animal
Evening
Beach
Vehicle
Carriage
person
Event 3
Event 4
Event 5
Day1
Day2
Trip1
Memory 10
Cross Modulation
Self Attention
Q
K
V
Context-Augmented Cross Attention
Feed-Forward Network
Self Attention
Feed-Forward Network
× N
[Current Utterance]
Can I see some photos of skiing?
Cross Modulation
LLM
[Generated Answer]
Here's a skiing photo for you. It looks like some lovely scenery for skiing!

Figure 2: Overview of our model architecture. Features from multimodal context which consists of sequences of pairs are given to the cross modulation block, along with trainable embedding vector and dialog history. This block is constituted of sequential attention blocks, each having self-attention and cross-attention layers. Passing two consecutive modulation blocks, context and history information based on the current turn's utterance is returned. These representations are passed on to the language model, enabling answer generation with a context.

pretrained models and history or context. The overall architecture is illustrated in Figure 2.

For the sake of clarity in the following discussion, we assume that three types of information are given in visual dialog tasks: a current utterance (text), dialog history (several utterances and answers), and context (image and text sequences). The main purpose of the method is to align the semantic spaces of multiple information using pretrained modality-specific models.

Pretrained Vision and Langauge Models

To get the feature representations for visual modality, we adopted a pretrained vision transformer model (ViT) (Dosovitskiy et al. 2020). Using the pretrained ViT model with frozen weight parameters, images included in context are converted into fixed-size embedding vectors. For the language modality, we adopted a pretrained encoder-decoder-based language model, Flan-T5 (Chung et al. 2022). All textual information from the current utterance, dialog history, and contexts are fed into the Flan-T5. Then, the answer utterances are generated with the Flan-T5 decoder.

Retrieval-augmented Structural Alignment

We introduce a novel method to align semantic spaces between multimodal sequential information using pretrained large-scale models. The two key ideas are 1) visual grounding to language space with a small modulation block and 2) considering the structural relationships within the context.

Basically, for visual grounding, Transformer (Vaswani et al. 2017) blocks between vision and language models are adapted as modulation blocks. The modulation block consists of self-attention and cross-attention layers. By adding additional tokens (illustrated as contextualized embedding tokens in Figure 2) into language models and cross-attention mechanisms between vision and language models, visual features are grounded in language modality.

Furthermore, it is assumed that there is an inherent spatiotemporal semantic structure in history and multimodal context as can be seen in Figure 1. Specifically, the most challenging dataset, COMET, provides context with complex multimodal graphs. Each node in the graph represents a memory unit, which consists of an image with textual attributes. According to the attributes of each memory unit, we can define a hierarchical structure with three layers: *trip, day* and *event*. To consider the structural information in the visual grounding step, we introduce a structure-aware cross-attention mechanism.

In detail, the cross-modulation block consists of self-attention, cross-attention, and a feed-forward layer. Every attention block executes basic scaled-dot product attention:

$$Attention(Q, K, V) = \text{softmax}(\frac{QK^\top}{\sqrt{d_k}})V \quad (1)$$

where Q = query, K = key, V = value, and d_k is the dimension of the key and query.

To include the spatiotemporal semantic structure of history and context into the cross-modulation block, an adjacency matrix A is introduced. The nodes of A represent each image or attribute, indicating their respective meanings. Edges of A are added between consecutive days or trips, also attributes having hierarchical relationships.

$$\begin{gathered} Attention(Q, K^*, V) = \text{softmax}(\frac{QK^*}{\sqrt{d_k}})V, \\ K^* = K^\top \odot A \end{gathered} \quad (2)$$

where $\odot$ is the element-wise product.

In our setting, the query stands for newly added tokens (*i.e.* contextualized embedding tokens). The representations for the contextualized embedding tokens will be trained to ground the visual information into the language semantic space constructed from the history and utterance.

In the 3, using a matrix to encode the structural information brings another advantage when it comes to multimodal data. If the context consists of images along with a text sequence, the correlation of the image-text pair can be lost when each modality is encoded with the modality-specific pretrained models. By encoding the relationships between the image-text pair in the matrix, this can be prevented and the model can match the true pair.

Outputs of cross-modulation blocks are contextualized embedding vectors. These vectors are trained and can be viewed as a contextualized representation of retrieved information. By using these, the language model can generate the answer by viewing the proper context at each turn.

Training Method

By leveraging the ability of the pretrained image encoder, the weight parameters of ViT are frozen. To generate answer sentences with contextualized visual information, the parameters of language models are trained with modulation blocks.

Our model is trained in an end-to-end manner by minimizing an answer generation loss:

$$L_\theta^{GEN} = -\sum_{i=1}^{|y|} \log P_\theta(y_i | y_{<i}, U, H, C, A) \quad (3)$$

where the decoder in the language model attends to previously generated tokens $y_{<i}$, utterance U, history H, and context C with structural information matrix A.

Experiments

Dataset

We mainly evaluate our algorithm on the most challenging visual dialog dataset, COMET (Kottur et al. 2022) and use two commonly used visual dialog datasets: VisDial 1.0 (VisDial) (Das et al. 2017), and MNIST Dialog (Seo et al. 2017).

In COMET, unlike the other two datasets having a static image as a context for each dialog, a model must retrieve relevant information from a memory graph at each turn, which is much more challenging. A dialog is grounded in this memory graph, which contains 100 memories. These memories are a set of pairs of images and attributes ((I_1, $attr_1$), . . . , (I_m, $attr_m$)). This memory graph has a spatiotemporal-hierarchical structure, where memories are grouped into *events*, *days*, and *trips*. A visualization of the structure of the memory graph is illustrated in figure 3.

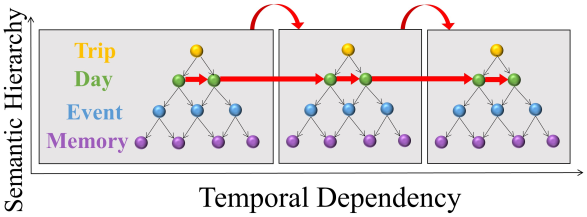

Figure 3: Visualization of spatiotemporal-hierarchical structure in the multimodal memory context.

In VisDial and MNIST Dialog datasets, at each turn *t* in a dialog, history H_t = ((Q_1, A_1), . . . , (Q_{t-1}, A_{t-1})), question Q_t and a single image I are given and the model is asked to answer. Images in VisDial and COMET are from MS-COCO (Lin et al. 2014) dataset. Images in MNIST Dialog are generated using MNIST digits. Statistics of three datasets are as follows:

	VisDial	MNIST Dialog	COMET
Total # dialogs	125.3k	150k	11.5k
Avg # turns / dialog	10	10	4.4
Avg # images in dialog	1	1	100

Table 1: Statistics of Visual Dialog datasets. In VisDial and MNIST Dialog datasets, every dialog consists of ten turns and one image. In COMET, the number of turns and images (memories) varies.

Evaluation Metrics

As each dataset has different tasks and different settings, we summarized evaluation metrics for clear comparison.

COMET To evaluate, four different subtasks are proposed: API Call Prediction, Multimodal Coreference Resolution (MM-Coref), Multimodal Dialog State Tracking (MM-DST), and Response Generation. In detail,

1. API Call Prediction predicts the right API call to execute the query. it consists of five classes.
   - SEARCH: Search based on current query.
   - REFINE_SEARCH: Enhancing the search over the existing query.
   - GET_INFO: Retrieve information about current or previously viewed memories.
   - GET_RELATED: Retrieve other memories similar to the current/prior memories.
   - SHARE: Share it with others.
2. Multimodal Coreference Resolution (MM-Coref) retrieves relevant memories from the context.
3. Multimodal Dialog State Tracking (MM-DST) is tracking user belief states across multiple turns.
4. Response Generation generates appropriate responses to user questions.

These subtasks are solved at once, by parsing the generated answer and comparing each part to the ground-truth answer. Evaluation metrics for each subtask are summarized in Table 2.

Task	Metric
API Call Prediction	Accuracy
MM-Coref	Precision / Recall / F1
MM-DST	Precision / Recall / F1
Response Generation	***Generation***: BLEU ***Retrieval***: Accuracy, Mean Reciprocal Rank, Mean Rank

Table 2: Subtasks and corresponding metrics of COMET dataset.

VisDial As 100 candidate answers are given at each turn t, both generative and discriminate settings can be considered. As our model is tackling generative setting, by calculating log-likelihood scores of each answer candidate given question and history, we can define the score by every possible answer. Based on this score, the ranking of 100 candidates is sorted. For evaluation, retrieval-based metrics are used: NDCG, Mean Reciprocal Rank (MRR), Recall@k, and Mean Rank with respect to human response. This generative setting, which ranks based on generative score and uses retrieval-based metrics, is broadly used in prior works (Kang et al. 2023; Wang et al. 2020)

MNIST Dialog In this dataset, a model must generate an answer for each turn, which is in a set of 38 possible single words. This can be treated as a classification task, having 38 classes. So the evaluation metric is classification accuracy. Answers contain information about visual attributes, *i.e.*, $\{$color, background color, number, style, count$\}$.

Baselines

We compare our model against various baselines on each dataset.

COMET GPT-2 model (Radford et al. 2019) based approach was proposed. In terms of using a memory graph, two approaches were considered: *text-only* and *multimodal*. In *text-only* setting, instead of image features, memory attributes are used as flattened strings. In contrast, in *multimodal* setting, the GPT-2 model uses BUTD (Anderson et al. 2018) and CLIP (Radford et al. 2021) image features as input image tokens.

VisDial Similar to the previous work (Kang et al. 2023), we compare the performance of our method with 10 baselines: 1) Attention-based models: CoAtt (Wu et al. 2018), HCIAE (Lu et al. 2017), Primary (Guo, Xu, and Tao 2019), ReDAN (Gan et al. 2019), DMRM (Chen et al. 2020a), DAM (Jiang et al. 2020b) 2) Graph-based models: KBGN (Jiang et al. 2020a), LTMI (Nguyen, Suganuma, and Okatani 2020), LTMI-GoG (Chen et al. 2021) 3) Semi-supervised learning model: GST (Kang et al. 2023).

MNIST Dialog To retrieve relevant information along with a question, associative attention memory was proposed (Seo et al. 2017). Another approach focused on coreference resolution and was implemented at a word level by utilizing two different modules, with each targeting coreference resolution and description separately (Kottur et al. 2018).

Implementation Details

For experiments with the COMET dataset, we use pretrained models with ViT-base for the image model and Flan-T5-base for the text model. The two sequential cross-modulation blocks of the proposed method are trained from scratch. The image model and text model have 86M and 250M parameters, respectively. Our modulation block contains 66M parameters. In total, our overall model has 468M parameters. It takes 5 hours for 20 epoch training with 64 batch size on a 4-A100 machine. During training, the pretrained image model is frozen, text model and cross-modulation models are optimized. Each cross-modulation block contains 6 cross-attention layers and 12 self-attention layers. The dimension of each vector is 768 and 32 vectors for the output embedding.

The same set of training hyperparameters is used for all experiments except for feature extraction with the ViT-base model. For experiments with the COMET dataset, representations of the special [CLS] token from the ViT-base models are used for the image features. For experiments with VisDial and MNIST Dialog datasets, the patch-wised representation from the ViT model is used as an image feature in the cross-modulation blocks. We use the AdamW optimizer (Loshchilov and Hutter 2017) with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and weight decay of 0.05. We use a piecewise linear scheduler with a linear warmup of 2K steps starting from a learning rate of 1e-4 and a peak learning rate of 1e-3.

Quantitative Results

COMET Table 3 shows that our model demonstrated superior results than the baseline in all tasks. Results were slightly different by cross-modulation architectures, but all exceeded the prior scores. In the multimodal baseline, despite image features from the memory graph being used,

Model	1. API	2. Coref	3. DST.		4. Gen.	
	Acc ↑	Coref F1 ↑	Slot F1 ↑	Joint Acc. ↑	BLEU ↑	BERTS. ↑
Text-only (Kottur et al. 2022)	88.6	78.2	91.5	72.9	0.205	0.895
Multimodal (Kottur et al. 2022)	89.4	84.8	92.6	77.5	0.251	0.905
Ours (Concatenation)	90.9	87.3	93.4	80.5	0.294	0.905
Ours (utterance → history)	90.6	87.2	93.7	80.5	0.296	0.907
Ours (history → utterance)	90.9	87.3	93.7	80.6	0.298	0.907

Table 3: Overview of evaluation results on the COMET test set. Our algorithm achieves higher performance than baselines, in all four tasks. Scores by different cross-modulation block structures are presented.

Model	1. API	2. Coref	3. DST.		4. Gen.	
	Acc	Coref F1	Slot F1	Joint Acc.	BLEU	BERTS.
w/o Hierarchy	90.8	87.1	93.5	80.2	0.291	0.906
Event	90.8	87.7	94.1	81.1	0.296	0.906
Day	90.9	87.3	93.7	80.6	0.298	0.907
Trip	91.0	87.3	93.7	80.8	0.299	0.907

Table 4: Archies on the COMET test set.

	NDCG↑	MRR↑	R@1↑	R@5↑	R@10↑	Mean↓
CoAtt	59.24	49.64	40.09	59.37	65.92	17.86
HCIAE	59.70	49.07	39.72	58.23	64.73	18.43
Primary	-	49.01	38.54	59.82	66.94	16.60
ReDAN	60.47	50.02	40.27	59.93	66.78	17.41
DMRM	-	50.16	40.15	60.02	67.21	15.19
DAM	60.93	50.51	40.53	60.84	67.94	16.65
KBGN	60.42	50.05	40.40	60.11	66.82	17.54
LTMI	61.61	50.38	40.30	60.72	68.44	15.73
LTMI-GoG	62.63	51.32	41.25	61.83	69.44	15.32
GST	64.50	52.06	42.04	62.92	71.06	14.54
Ours	63.27	52.59	42.90	62.51	69.57	14.96

Table 5: Comparison with the state-of-the-art generative models on VisDial v1.0 validation set. Compared baseline results are from (Kang et al. 2023).

Model	Accuracy ↑
I (Seo et al. 2017)	20.18
Q (Seo et al. 2017)	36.58
ATT (Seo et al. 2017)	62.62
ATT\H (Seo et al. 2017)	79.72
AMEM (Seo et al. 2017)	87.53
AMEM\H+SEQ (Seo et al. 2017)	96.39
CorefNMN\SEQ (Kottur et al. 2018)	88.7
CorefNMN (Kottur et al. 2018)	99.3
Ours	98.6

Table 6: Answer accuracy on MNIST Dialog test set.

these features are treated as flattened tokens, losing intrinsic structural information of the memory graph. But our algorithm takes account of *trip-day-event* hierarchical structure. Using image features exclusively disregards valuable information that cannot be found in the image only, such as time, location, and person's name. Ablation results without hierarchical information and considering various levels of hierarchy are presented in table 4. The overall performance without the hierarchical structure of context was the lowest, proving the necessity of using hierarchical information.

VisDial As presented in table 5, our model exhibits remarkable results on overall metrics, compared to the baselines on VisDial v1.0 validation set. We compared our result with the attention-based approach (top 6 rows), graph-based approach (middle 3 rows), and self-training-based approach (bottom 1 row). Among six evaluation metrics, we demonstrate state-of-the-art scores in two: MRR and R@1. In the other four metrics: R@5, R@10, Mean Rank, and NDCG, the score gap is little. One thing we want to emphasize is in the previous work, an additional training phase and generated dataset were used.

In GST (Kang et al. 2023), a generative self-training-based approach was used, which trains the teacher & questioner model first, and the main student model successively. Conceptual 12M dataset (Changpinyo et al. 2021) is used as an image and dialog data is generated by teacher & questioner model using these images. The student model is trained on the generated dialog dataset additionally. To train with this generated dataset, perplexity-based data selection was used and multimodal consistency regularization loss term was introduced to improve the generalization capability of the model. Our algorithm is trained end-to-end with a single model and uses a VisDial train set only, which is a much simpler yet harder setting.

MNIST Dialog Our model surpassed every other baseline, as shown in table 6, except one (Kottur et al. 2018). This algorithm takes advantage of separate neural modules specialized in coreference resolution and description. And $\Delta_i t$, which is the absolute difference between the current turn and the turn when each candidate from the reference pool was first mentioned, was used. This information is critical, as removing this drops accuracy significantly. But our model assumes a more general setting, not saving every previously seen entity with its visual groundings in a reference pool, which is memory inefficient. By using multimodal context,

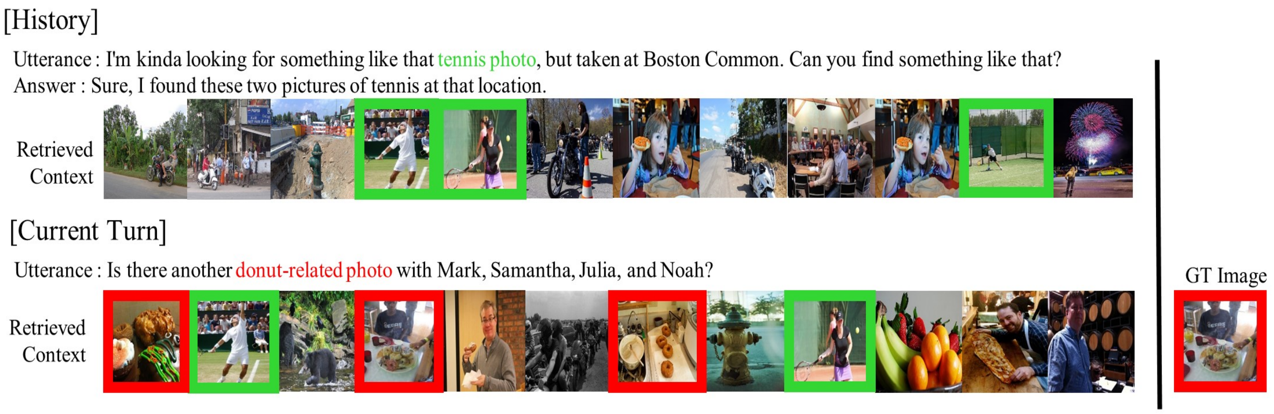
[History]
Utterance : I'm kinda looking for something like that tennis photo, but taken at Boston Common. Can you find something like that?
Answer : Sure, I found these two pictures of tennis at that location.
Retrieved Context
[Current Turn]
Utterance : Is there another donut-related photo with Mark, Samantha, Julia, and Noah?
Retrieved Context
GT Image

Figure 4: Examples of retrieved images from multimodal context. We visualize the focus of each multi-head in a cross-modulation block by attention scores. The green-colored boxes represent images related to history, while the red-colored boxes represent images related to the current utterance. As selected memories containing ground-truth memories, this result indicates that our cross-modulation block can retrieve effectively necessary information from the context.

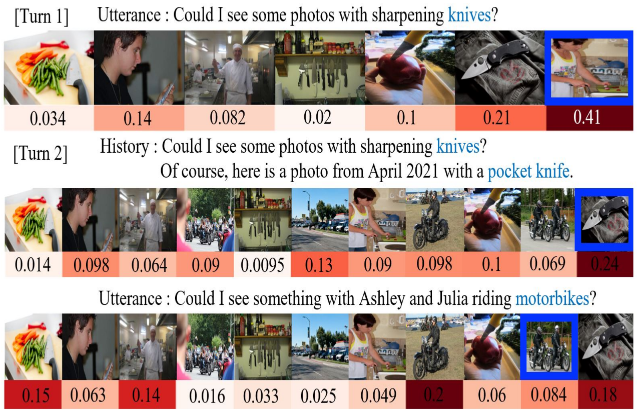
[Turn 1]
Utterance : Could I see some photos with sharpening knives?
0.034
0.14
0.082
0.02
0.1
0.21
0.41
[Turn 2]
History : Could I see some photos with sharpening knives?
Of course, here is a photo from April 2021 with a pocket knife.
0.014
0.098
0.064
0.09
0.0095
0.13
0.09
0.098
0.1
0.069
0.24
Utterance : Could I see something with Ashley and Julia riding motorbikes?
0.15
0.063
0.14
0.016
0.033
0.025
0.049
0.2
0.06
0.084
0.18

Figure 5: Visualization of attention scores with retrieved images. The blue box represents the ground truth images associated with the given dialog history and the below boxes with numbers represent attention scores from the suggested method.

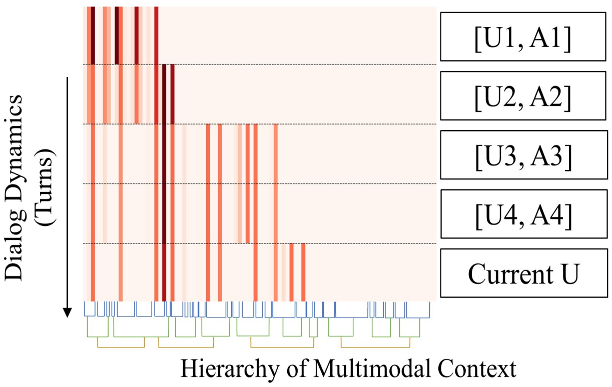

Figure 6: Visualization of attention scores for context and history.

textual information that cannot be found in the image (*e.g.*, time, name of the person, city) can be given to the model, flourishing the reasoning ability.

Qualitative Results

We investigated retrieved contexts by each attention head in Figure 4 and visualize with the attention scores in Figure 5. By sequentially passing history and current utterance, not only the current turn's context but also history's context are retrieved. These results clearly demonstrate our sequential cross-modulation blocks.

In Figure 6, we visualize which information in the entire dialog history (past utterances and answers) and the entire memory context receives high attention scores as the dialog's turn progresses. This can demonstrate how the referenced memory changes with each turn. One important point to emphasize is that the suggested method gives less attention scores on unrelated previous history as the dialog progresses.

Conclusion

We propose a novel cross-modulation-based method for Visual Dialog. By introducing a modulation block, pretrained language-only, and vision-only models can be aligned without joint training. Thus, the model can enjoy the rich representational ability from the two modalities. As multiple cross-attention layers, relevant history and context information to the current utterance can be retrieved. By using a matrix representing structural information, hierarchical structure in context can be preserved. These two processes, align and retrieve, are the key components of using pretrained models for visual dialog tasks. We demonstrated impressive results in three visual dialog datasets.

Acknowledgements

Authors from Hanyang University were supported by IITP grant funded by MSIT (Grant No. 2022-0-00264/40%, 2022-0-00612/20%, 2022-0-00951/20%,) and IITP Artificial Intelligence Graduate School Program for Hanyang University funded by MSIT (Grant No. 2020-0-01373/20%).

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 6077–6086.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Changpinyo, S.; Sharma, P.; Ding, N.; and Soricut, R. 2021. 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*, 3558–3568.

Chen, F.; Chen, X.; Meng, F.; Li, P.; and Zhou, J. 2021. GoG: Relation-aware graph-over-graph network for visual dialog. *arXiv preprint arXiv:2109.08475*.

Chen, F.; Meng, F.; Xu, J.; Li, P.; Xu, B.; and Zhou, J. 2020a. Dmrm: A dual-channel multi-hop reasoning model for visual dialog. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 7504–7511.

Chen, F.; Zhang, D.; Chen, X.; Shi, J.; Xu, S.; and Xu, B. 2022. Unsupervised and Pseudo-Supervised Vision-Language Alignment in Visual Dialog. In *Proceedings of the 30th ACM International Conference on Multimedia*, 4142–4153.

Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020b. Uniter: Universal image-text representation learning. In *European conference on computer vision*, 104–120. Springer.

Chung, H. W.; Hou, L.; Longpre, S.; Zoph, B.; Tay, Y.; Fedus, W.; Li, E.; Wang, X.; Dehghani, M.; Brahma, S.; et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Das, A.; Kottur, S.; Gupta, K.; Singh, A.; Yadav, D.; Moura, J. M.; Parikh, D.; and Batra, D. 2017. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 326–335.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Gan, Z.; Cheng, Y.; Kholy, A. E.; Li, L.; Liu, J.; and Gao, J. 2019. Multi-step reasoning via recurrent dual attention for visual dialog. *arXiv preprint arXiv:1902.00579*.

Guo, D.; Wang, H.; Zhang, H.; Zha, Z.-J.; and Wang, M. 2020. Iterative context-aware graph inference for visual dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10055–10064.

Guo, D.; Xu, C.; and Tao, D. 2019. Image-question-answer synergistic network for visual dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10434–10443.

Jiang, X.; Du, S.; Qin, Z.; Sun, Y.; and Yu, J. 2020a. KBGN: Knowledge-bridge graph network for adaptive vision-text reasoning in visual dialogue. In *Proceedings of the 28th ACM international conference on multimedia*, 1265–1273.

Jiang, X.; Yu, J.; Sun, Y.; Qin, Z.; Zhu, Z.; Hu, Y.; and Wu, Q. 2020b. DAM: Deliberation, abandon and memory networks for generating detailed and non-repetitive responses in visual dialogue. *arXiv preprint arXiv:2007.03310*.

Kang, G.-C.; Kim, S.; Kim, J.-H.; Kwak, D.; and Zhang, B.-T. 2023. The Dialog Must Go On: Improving Visual Dialog via Generative Self-Training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6746–6756.

Kim, W.; Son, B.; and Kim, I. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, 5583–5594. PMLR.

Kottur, S.; Moon, S.; Geramifard, A.; and Damavandi, B. 2022. Navigating Connected Memories with a Task-oriented Dialog System. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2495–2507.

Kottur, S.; Moura, J. M.; Parikh, D.; Batra, D.; and Rohrbach, M. 2018. Visual coreference resolution in visual dialog using neural module networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 153–169.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 12888–12900. PMLR.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, 121–137. Springer.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzer-*

land, September 6-12, 2014, Proceedings, Part V 13, 740–755. Springer.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32.

Lu, J.; Kannan, A.; Yang, J.; Parikh, D.; and Batra, D. 2017. Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model. *Advances in Neural Information Processing Systems*, 30.

Lu, P.; Li, H.; Zhang, W.; Wang, J.; and Wang, X. 2018. Co-attending free-form regions and detections with multi-modal multiplicative feature embedding for visual question answering. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Nguyen, V.-Q.; Suganuma, M.; and Okatani, T. 2020. Efficient attention mechanism for visual dialog that can handle all the interactions between multiple inputs. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, 223–240. Springer.

Park, S.; Whang, T.; Yoon, Y.; and Lim, H. 2021. Multi-view attention network for visual dialog. *Applied Sciences*, 11(7): 3009.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Radford, A.; Narasimhan, K.; Salimans, T.; Sutskever, I.; et al. 2018. Improving language understanding by generative pre-training.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Seo, P. H.; Lehrmann, A.; Han, B.; and Sigal, L. 2017. Visual reference resolution using attention memory for visual dialog. *Advances in neural information processing systems*, 30.

Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; and Dai, J. 2019. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv prepLewiearXiv:1908.08530*.

Tan, H.; and Bansal, M. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, Y.; Joty, S.; Lyu, M. R.; King, I.; Xiong, C.; and Hoi, S. C. 2020. Vd-bert: A unified vision and dialog transformer with bert. *arXiv preprint arXiv:2004.13278*.

Wu, Q.; Wang, P.; Shen, C.; Reid, I.; and Van Den Hengel, A. 2018. Are you talking to me? reasoned visual dialog generation through adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 6106–6115.

Yu, F.; Tang, J.; Yin, W.; Sun, Y.; Tian, H.; Wu, H.; and Wang, H. 2021. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 3208–3216.

Zellers, R.; Lu, J.; Lu, X.; Yu, Y.; Zhao, Y.; Salehi, M.; Kusupati, A.; Hessel, J.; Farhadi, A.; and Choi, Y. 2022. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16375–16387.

Zhang, H.; Wang, X.; Jiang, S.; and Li, X. 2022a. Multi-Granularity Semantic Collaborative Reasoning Network for Visual Dialog. *Applied Sciences*, 12(18): 8947.

Zhang, S.; Jiang, X.; Yang, Z.; Wan, T.; and Qin, Z. 2022b. Reasoning with Multi-Structure Commonsense Knowledge in Visual Dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4600–4609.

Zheng, Z.; Wang, W.; Qi, S.; and Zhu, S.-C. 2019. Reasoning visual dialogs with structural and partial observations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6669–6678.