Every Problem, Every Step, All In Focus: Learning to Solve Vision-Language Problems with Integrated Attention

Xianyu Chen, Jinhui Yang, Shi Chen, Louis Wang, Ming Jiang, and Qi Zhao

***Abstract*—Integrating information from vision and language modalities has sparked interesting applications in the fields of computer vision and natural language processing. Existing methods, though promising in tasks like image captioning and visual question answering, face challenges in understanding real-life issues and offering step-by-step solutions. In particular, they typically limit their scope to solutions with a sequential structure, thus ignoring complex inter-step dependencies. To bridge this gap, we propose a graph-based approach to vision-language problem solving. It leverages a novel integrated attention mechanism that jointly considers the importance of features within each step as well as across multiple steps. Together with a graph neural network method, this attention mechanism can be progressively learned to predict sequential and non-sequential solution graphs depending on the characterization of the problem-solving process. To tightly couple attention with the problem-solving procedure, we further design new learning objectives with attention metrics that quantify this integrated attention, which better aligns visual and language information within steps, and more accurately captures information flow between steps. Experimental results on VisualHow, a comprehensive dataset of varying solution structures, show significant improvements in predicting steps and dependencies, demonstrating the effectiveness of our approach in tackling various vision-language problems.**

***Index Terms*—Vision-language problem solving, multimodal attention, graph attention, integrated attention mechanism.**

I. Introduction

RECENT years have witnessed impressive progress in computer vision and natural language processing, enabling intelligent systems to perform a broad range of joint vision-language tasks, such as image captioning [1]–[6], visual storytelling [7], [8], visual question answering [9]–[16], visual dialog [17]–[19], and natural language generation [20]–[22]. However, a major challenge still remains in developing artificial intelligence that can understand vision-language problems and provide procedural solutions with step-by-step instructions. Humans exhibit remarkable ability in visually perceiving problems, comprehending goals, and mapping out plans and procedures to solve them. Developing similar procedural reasoning capabilities in artificial intelligence remains a significant challenge.

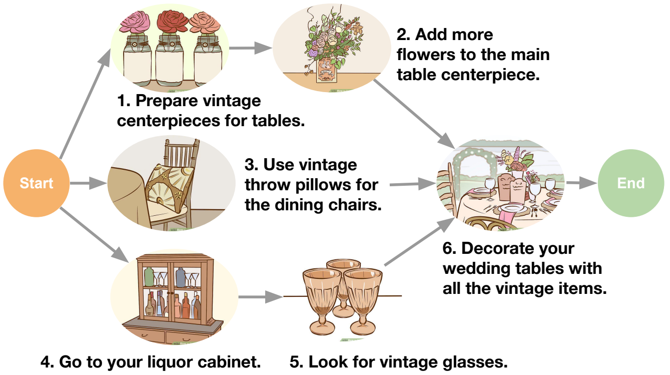
Start
1. Prepare vintage centerpieces for tables.
2. Add more flowers to the main table centerpiece.
3. Use vintage throw pillows for the dining chairs.
4. Go to your liquor cabinet.
5. Look for vintage glasses.
6. Decorate your wedding tables with all the vintage items.
End

Fig. 1. Problem-solving tasks such as "how to decorate the tables for a vintage-themed wedding" often follow a non-sequential procedure. For example, steps 1, 3, and 4 can be completed in no particular order, as long as step 1 takes place before step 2, step 4 happens before step 5, and all of them take place before step 6. Our method represents such problem-solving procedures in a graph structure. Steps are represented as nodes, and dependent steps are directly connected by edges indicating ordering constraints. In this way, our approach can handle various types of step dependencies in free-formed procedures. Attention is optimized end-to-end over the full graph-based solution structure.

Solving vision-language problems requires recognizing important visual details, understanding the multimodal context, and predicting cohesive solutions incorporating visual illustrations and natural language descriptions [23]. Understanding and predicting such multimodal descriptions require an intelligent system to decompose the solution into multiple steps. For example, as shown in Fig. 1, visual illustrations (*e.g.*, flowers, pillows) or natural language descriptions (*e.g.*, "Look for vintage glasses") are used to describe specific steps taken to decorate the tables for a vintage-themed wedding. Existing methods [24]–[34] have approached problem-solving with procedure planning, representing each solution as a linear sequence of steps. Such sequential approaches, while convenient, are unable to model complex dependencies across multiple steps. Vision-language problems often involve multiple dependencies between steps, which might not fit neatly into a linear sequence: (1) a step may depend on multiple steps. As shown in Fig. 1, step 6 must depend on the completion of steps 2, 3, and 5, and (2) certain problem-solving steps (*e.g.*, paths 1-2, 3, 4-5 in Fig. 1) can occur

Xianyu Chen, Jinhui Yang, Shi Chen, Louis Wang, Ming Jiang, and Qi Zhao are with the Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN, 55455.
Source code: https://github.com/chenxy99/SGAN
E-mail: see https://www-users.cs.umn.edu/~qzhao

simultaneously. A sequential model might oversimplify the relationships and struggle to represent these cases effectively, facing challenges in the following aspects: First, sequential models inherently follow a linear *structure*, processing information in a step-by-step fashion. This linear nature becomes a constraint when dealing with multiple dependencies that don't conform to a straightforward sequence. Second, the *efficiency* of sequential models is compromised when confronted with interdependencies across multiple steps. Directly converting parallel processes into a fixed-order sequence regardless of variations can lead to suboptimal and inefficient solutions. Third, sequential models often lack the *interpretability* required to understand complex dependencies between different steps of the problem-solving process, diminishing the trust and transparency essential for real-world applications. Therefore, in light of these challenges, our work is motivated by the need for a more flexible and structured approach to vision-language problem-solving. Therefore, in light of these challenges, our work is motivated by the need for a more flexible and nuanced approach to vision-language problem-solving.

In this work, to enable more general and flexible problem solving, we propose a graph neural network approach that represents solutions as graphs. This structured representation allows graph-based models to overcome the limitations of sequential models, providing a more general and effective approach to handling complex problem-solving scenarios. Our method leverages an integrated attention mechanism that jointly models intra-step attention and inter-step attention. This provides a more holistic view compared to isolated step-based attention. To jointly and progressively supervise the integrated attention, we further introduce quantitative metrics that consider attention propagation across the entire graph of solution steps. This graph-based approach with the novel integrated attention mechanism aims to provide an effective framework for modeling complex dependencies across multiple steps and solving real-world problems, such as those in autonomous driving, medical diagnosis, and various other applications.

To summarize, the main contributions of this paper are as follows:

1) We propose a graph neural network approach to represent procedural solutions as graphs, capturing complex step dependencies and enabling an integral understanding of the entire problem-solving procedure.
2) We design an integrated attention mechanism that jointly models the importance of multimodal features within each step as well as across interdependent steps.
3) We introduce quantitative attention metrics to optimize attention propagation across the full solution graph, enabling supervised learning of attention for complex vision-language problem solving.

The remainder of this paper is structured as follows. In Section II, we provide a concise overview of related research pertaining to vision-language problem solving and attention mechanisms in vision-language tasks. Section III outlines the problem statement, introducing the formulation of the vision-language problem solving task that we aim to address. The details of our proposed method, designed to tackle the aforementioned task, are elaborated in Section IV. Extensive experiments are presented in Section V, where we report quantitative and qualitative results, along with comprehensive analyses of our approach's performance. We conclude this paper and discuss its limitations in Section VI, while also providing directions for future research and improvements.

II. Related Works

Our work is relevant to previous efforts on visual problem-solving, attention mechanisms in vision-and-language tasks, and supervision of attention.

A. Problem Solving Methods

Procedural problem solving with instructional solutions has gained increasing research attention. Several studies [23]–[26], [28]–[30], [34], [35] have curated datasets of images or videos demonstrating procedures for daily tasks like cooking, maintenance, sports, and healthcare. These efforts have enabled data-driven approaches to generate solutions for diverse problems. A series of previous methods focus on developing captioning models to summarize instructional text describing procedures [23], [24], [33]. Other works emphasize aligning textual and visual modalities [23], [25], [26], [28], [36]. They retrieve images given instruction text or localize described activities. Alternative approaches factorize solutions into discrete steps and predict structured representations [27], [31]–[33], [35], [37]–[39]. However, these studies oversimplify real-world solution procedures as sequential activities. Solutions often have complex, free-formed structures with inter-dependencies between steps. Thus, while demonstrating feasibility for varied tasks, existing methods are limited in generalizing across problems regardless of solution structure. They also do not perform joint reasoning over steps and their relationships. Our work addresses these gaps by representing solutions as graphs to capture step dependencies and provide a comprehensive framework for complex problem solving.

B. Attention in Vision-Language Tasks

Attention mechanisms have become critical components in vision-language models to effectively couple modalities and identify salient features for various tasks. Prior studies have focused on designing attention for input feature prioritization [1], [13], [40]–[42], cross-modal alignment [23], [43], and concept-dependency modeling [44]–[46]. Early attention approaches operated on grid-structured inputs like images or text, using convolutional neural networks [47] or Transformers [48], [49], while recent graph-based methods [44], [46] allow modeling attention in structured inputs [50]–[53]. However, capturing the complex dependencies across steps in procedural solutions requires structured representations that consider attention shifts across multiple modalities and multiple steps. We advance existing techniques with a novel integrated attention mechanism that enables joint attention modeling for both aspects and leverage this new attention mechanism to progressively construct structured solutions for various problems.

C. Supervised Learning of Attention

Instead of implicitly learning the attention mechanism with the end objectives of different tasks, prior works have explored explicitly supervising attention mechanisms to improve alignment with regions of interest. Various approaches have been proposed to construct the ground truth attention based on task annotation [13], [42], human attention [4], [5], [41], [54], or adversarial learning [55]. Some supervision methods use single-step supervision based on human annotations of salient image regions [41], [42], [54], while others account for integrating attention across the visual reasoning procedure [13]. However, focusing on local alignments limits modeling relationships between steps in structured problem solving. Without propagating attention, these methods fail to capture complex interdependencies in multi-step procedures. Differently, in this work, we present a new metric that quantitatively measures the contributions of attention for constructing the task solution, and leverage it to progressively supervise both the intra- and inter-step attention. It provides an integral view of problem-solving procedures, resulting in enhanced performance in formulating a structured representation of the solutions.

III. Problem Statement

The vision-language problem solving task involves comprehending general vision-language problems and generating structured instructions to address them, incorporating both visual and textual information [23]. Previous research has explored instructional images [24], [56] or videos [26], [28], [29], [35], [57], but these were limited to predicting sequential instructions for specific task categories. In contrast, our work considers a wide range of problems and their corresponding solution structures. The fundamental goals of our proposed approach are twofold: (1) understanding the input problem description and (2) constructing a solution graph consisting of essential problem-solving steps, each associated with relevant images and captions.

As shown in Fig. 1, the input of our proposed approach consists of a problem description $\boldsymbol{g}$, such as “how to decorate the tables for a vintage-themed wedding,” and a pool of images $\{I_1, I_2, \cdots, I_N\}$ or captions $\{C_1, C_2, \cdots, C_N\}$. These images and captions serve as candidate steps or actions that could be relevant or irrelevant to solving the given problem. The main challenge in the vision-language problem solving task is to identify the essential steps and their correct order to construct a coherent and effective solution for the problem at hand.

To tackle this challenge, our proposed approach involves creating a solution graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ that encapsulates the problem-solving process. The graph nodes in $\mathcal{V}$ represent essential steps, including the start node (node 0), the end node (node $N+1$), and the nodes corresponding to the candidate steps (nodes $1, \cdots, N$) with their associated image or caption capturing the possible actions that can be taken to solve the problem. The edges in $\mathcal{E}$ represent the dependencies or chronological order between the steps. For instance, a directed edge between the nodes corresponding to “prepare vintage centerpieces for tables” and “add more flowers to the main table centerpiece” indicates that the latter step should happen after the former.

By constructing such a directed graph, our approach can effectively model the logical flow of the problem-solving procedure, enabling a structured and coherent representation of the solution. The directed graph representation also allows for the existence of multiple paths from the start node to the end node, corresponding to different ways of solving the problem. This flexibility in the graph structure is particularly beneficial for handling vision-language problems with multiple viable solutions or alternative sequences of steps.

IV. Method

Our proposed Solution Graph Attention Network (SGAN) addresses the vision-language problem solving task by leveraging both intra-step and inter-step attention mechanisms to iteratively refine the solution graph. The key technical components of our method are (1) a novel graph neural network approach that progressively predicts solutions with diverse structures, (2) an integrated attention mechanism combining intra-step attention and inter-step attention for a comprehensive understanding of the problem-solving procedure, and (3) new attention metrics and learning objectives to jointly supervise the attention throughout the solution graph by leveraging information propagation. Together, these components empower SGAN to effectively capture dependencies within individual steps and the relationships between them, providing a powerful ability to handle complex vision-language problems and generate coherent solutions.

A. Solution Graph Attention Network

In problem-solving scenarios, dependencies between steps can be complex and may not be readily apparent. To address this challenge and predict the solution graph $\mathcal{G}$, SGAN progressively learns integrated attention using a graph attention network, enabling a better understanding of the problem-solving procedure.

As depicted in Fig.2, the input features representing the candidates, denoted as $\boldsymbol{v} = \{\boldsymbol{v}_i | i = 1, \cdots, N\}$, are obtained with a pre-trained image encoder (*e.g.*, ResNeXT-101 [58], ViT [59]) for image candidates or a language embedding network (*e.g.*, BERT [48]) for caption candidates. The language embedding $\boldsymbol{g}$ represents the description of the input problem [23], [48]. SGAN is designed with a stack of L graph attention layers, allowing the step-by-step refinement of the solution graph. Specifically, the network iteratively updates the node representations $\boldsymbol{h}^{(\ell)} = \{\boldsymbol{h}_i^{(\ell)} | i = 0, \cdots, N+1\}$, where $\ell = 1, \cdots, L$ indicates the ℓ-th layer. It consists of the updated features of the graph nodes start ($i = 0$), end ($i = N+1$), and each candidate step ($i = 1, \cdots, N$). The node representations of the previous layer $\boldsymbol{h}^{(\ell-1)}$ are passed to the current layer as the input, while the first layer input is initialized as $\boldsymbol{h}^{(0)} = \{\boldsymbol{g}, \bar{\boldsymbol{v}}_1, \cdots, \bar{\boldsymbol{v}}_N, \boldsymbol{W}_e \boldsymbol{g}\}$, where $\bar{\boldsymbol{v}}_i$ is the average of $\boldsymbol{v}_{i,k}$ across all $k = 1, \cdots, K$ image patches or word tokens, and $\boldsymbol{W}_e$ represents the learnable parameters to transform the language embedding $\boldsymbol{g}$ as the end

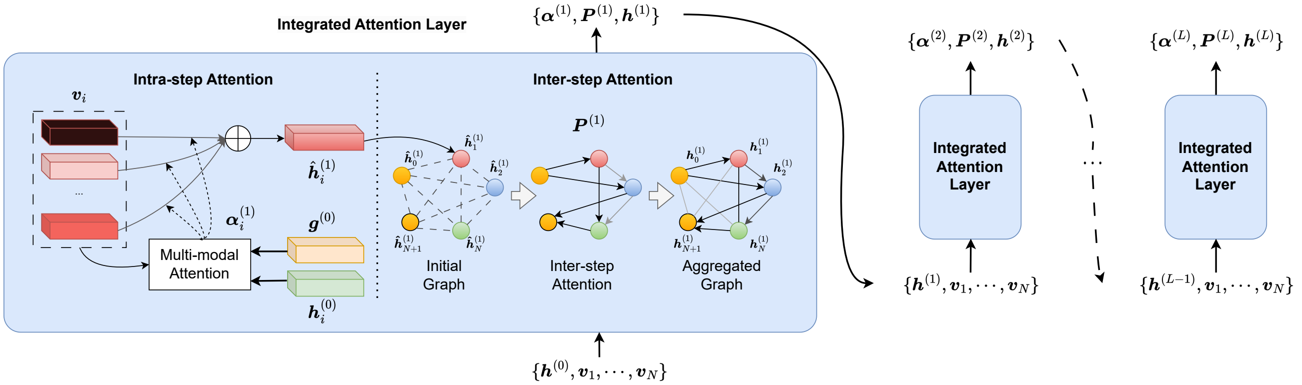
Integrated Attention Layer
$\{\boldsymbol{\alpha}^{(1)}, \boldsymbol{P}^{(1)}, \boldsymbol{h}^{(1)}\}$
Intra-step Attention
$\boldsymbol{v}_i$
...
$\hat{\boldsymbol{h}}_i^{(1)}$
$\boldsymbol{\alpha}_i^{(1)}$
$\boldsymbol{g}^{(0)}$
Multi-modal Attention
$\boldsymbol{h}_i^{(0)}$
Inter-step Attention
$\boldsymbol{P}^{(1)}$
$\hat{\boldsymbol{h}}_0^{(1)}$
$\hat{\boldsymbol{h}}_1^{(1)}$
$\hat{\boldsymbol{h}}_2^{(1)}$
$\hat{\boldsymbol{h}}_{N+1}^{(1)}$
$\hat{\boldsymbol{h}}_N^{(1)}$
Initial Graph
Inter-step Attention
$\boldsymbol{h}_0^{(1)}$
$\boldsymbol{h}_1^{(1)}$
$\boldsymbol{h}_2^{(1)}$
$\boldsymbol{h}_{N+1}^{(1)}$
$\boldsymbol{h}_N^{(1)}$
Aggregated Graph
$\{\boldsymbol{h}^{(0)}, \boldsymbol{v}_1, \cdots, \boldsymbol{v}_N\}$
$\{\boldsymbol{\alpha}^{(2)}, \boldsymbol{P}^{(2)}, \boldsymbol{h}^{(2)}\}$
Integrated Attention Layer
$\{\boldsymbol{h}^{(1)}, \boldsymbol{v}_1, \cdots, \boldsymbol{v}_N\}$
...
$\{\boldsymbol{\alpha}^{(L)}, \boldsymbol{P}^{(L)}, \boldsymbol{h}^{(L)}\}$
Integrated Attention Layer
$\{\boldsymbol{h}^{(L-1)}, \boldsymbol{v}_1, \cdots, \boldsymbol{v}_N\}$

Fig. 2. Overview of the proposed SGAN architecture. The input consists of node representations $\boldsymbol{h}^{(0)}$, features for images/caption candidates $\boldsymbol{v}_1, \cdots, \boldsymbol{v}_N$. The network leverages an integrated attention mechanism that progressively processes the input features and predicts the output intra-step attention $\boldsymbol{\alpha}^{(L)}$ for capturing salient information from the input images or captions, the inter-step attention $\boldsymbol{P}^{(L)}$ characterizing the probabilities of dependencies across different steps, and the final updated node representations $\boldsymbol{h}^{(L)}$.

node representation. Each layer also outputs the corresponding intra-step attention $\boldsymbol{\alpha}^{(\ell)}$ and the inter-step attention $\boldsymbol{P}^{(\ell)}$ (see Section IV-B for details).

To convert the final-layer output $\boldsymbol{P}^{(L)}$ into the solution graph $\mathcal{G}$, we employ the following process. Initially, a heuristic threshold is applied to the dependency matrix $\boldsymbol{P}^{(L)}$ to preserve the most pertinent nodes (see Steps 1-3 in Algorithm 1). Next, these selected nodes are iteratively added into the graph (see Steps 4-5 in Algorithm 1), along with their associated edges featuring the highest values in $\boldsymbol{P}^{(L)}$. This iterative process ensures that the graph remains a directed acyclic graph without loops or isolated nodes. Finally, attention weights $\boldsymbol{\alpha}^{(L)}$ assigned to each step's images and captions offer insights into what demands attention for effectively solving the given problem.

The proposed network is powerful for learning the dependencies between problem-solving steps. By using this iterative approach, the network can generate free-formed solutions with a better understanding of the problem-solving procedure. In the following, we will describe the specific design of our integrated attention mechanism to effectively capture the important contents and dependencies across problem-solving steps.

B. Integrated Attention Mechanism

Attention is a crucial component that drives advancements in natural language processing and computer vision, which enables models to selectively focus on the most relevant parts of the input data when performing different tasks. In the context of problem-solving, our integrated attention mechanism plays a critical role in identifying the key features and dependencies between the steps involved in a solution. It combines intra-step and inter-step attention to enable the network to capture both the fine-grained details of each step and the broader context in which they exist.

1) Intra-Step Attention: The intra-step attention focuses on capturing salient information from the input images or captions for understanding and completing each individual step. Specifically, in the ℓ-th layer, for the i-th candidate step, we define the intra-step attention weights as $\boldsymbol{\alpha}_i^{(\ell)}$, which

Algorithm 1 Graph post-processing method to obtain the final solution graph

```
INPUT: Predicted dependency matrix $\boldsymbol{P}^{(L)}$, retrieval threshold $\lambda_r$, and dependency threshold $\lambda_d$.
    1: Filter candidate steps using $\lambda_r$ over $\boldsymbol{P}^{(L)}_{0,1:N}$ to obtain the node set $S$, where $\boldsymbol{P}^{(L)}_{0,i} \geq \lambda_r$ for $i \in S$.
    2: Remove cycles between nodes $i$ and $j$ in $S$ by updating $\boldsymbol{P}^{(L)}_{i,j} = \max(0, \boldsymbol{P}^{(L)}_{i,j} - \boldsymbol{P}^{(L)}_{j,i})$.
    3: Initialize solution graph $\mathcal{G}$ with nodes $\mathcal{V} = \{0, N+1\}$, edges $\mathcal{E} = \{(0, N+1)\}$, and candidate edges $\mathcal{W} = \{(0, N+1)\}$ containing potential edges to add to the graph.
    For $u$ in $S$
        4: Find the best node $v_b$ with maximum
            $b = \max_{v \in \mathcal{S}, v \notin \mathcal{V}} \max_{(\bar{v}_1, \bar{v}_2) \in \mathcal{W}} \left( \sum_{\bar{v}_3 \in Pa(\bar{v}_1)} \boldsymbol{P}^{(L)}_{\bar{v}_3, v} + \boldsymbol{P}^{(L)}_{\bar{v}_1, v} + \boldsymbol{P}^{(L)}_{v, \bar{v}_2} + \sum_{\bar{v}_4 \in Ch(\bar{v}_2)} \boldsymbol{P}^{(L)}_{v, \bar{v}_4} \right),$
        where $Pa(\bar{v}_1)$ and $Ch(\bar{v}_2)$ represent the parent set of node $\bar{v}_1$ and child set of node $\bar{v}_2$ in the solution graph $\mathcal{G}$, respectively.
        5: If $b > \lambda_d$, Update the edge set $\mathcal{E}$ and node set $\mathcal{V}$ by adding node $v_b$ and candidate edges in $\mathcal{W}$ to ensure the graph remains a directed acyclic graph.
OUTPUT: The final solution graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$
```

is computed based on the input problem description $\boldsymbol{g}$, the candidate features $\boldsymbol{v}_i$, and the node features $\boldsymbol{h}_i^{(\ell-1)}$:

$$\boldsymbol{a}_{i,k}^{(\ell)} = \boldsymbol{w}_a^T \tanh(\boldsymbol{W}_v \boldsymbol{v}_{i,k} + \boldsymbol{W}_g \boldsymbol{h}_0^{(\ell-1)} + \boldsymbol{W}_h \boldsymbol{h}_i^{(\ell-1)}), \quad (1)$$

where $\boldsymbol{w}_a, \boldsymbol{W}_g, \boldsymbol{W}_v, \boldsymbol{W}_h$ are learnable parameters, and $k = 1, \cdots, K$ indicates the k-th element of the input candidate (*i.e.*, image patch or word token).

The attention weights $\boldsymbol{a}_i^{(\ell)}$ are normalized as $\boldsymbol{\alpha}_i^{(\ell)}$ with a masked softmax activation function

$$\boldsymbol{\alpha}_i^{(\ell)} = \text{softmax}(\boldsymbol{a}_i^{(\ell)}, \boldsymbol{m}_i), \quad (2)$$

where $\boldsymbol{m}_i$ is a binary vector and $\boldsymbol{m}_{i,k}$ indicates the k-th element (*i.e.*, image patch or word token) of the i-th candidate features is padded or not due to the variable length of the image or language inputs.

Finally, we apply the attention to the candidate features $\boldsymbol{v}$ to initialize the node representations for the ℓ-th layer:

$$\hat{\boldsymbol{h}}_i^{(\ell)} = \begin{cases} \boldsymbol{h}_i^{(\ell-1)} & \text{if } i = 0 \text{ or } N+1 \\ \sum_k \boldsymbol{\alpha}_{i,k}^{(\ell)} \boldsymbol{v}_{i,k}^{(\ell-1)} & \text{if } i = 1, \ldots, N. \end{cases} \quad (3)$$

2) Inter-Step Attention: The inter-step attention is responsible for capturing the chronological order between different problem-solving steps, providing a coherent and structured representation of the solution. By integrating inter-step attention into our model, we aim to enable more effective joint reasoning across multiple problem-solving steps. Specifically, we compute graph attention weights [44], [46] to estimate the existence of a dependency between each pair of steps based on the initial node features $\hat{\boldsymbol{h}}_i^{(\ell)}$ computed in Equation (3):

$$\boldsymbol{P}_{i,j}^{(\ell)} = \sigma\left(\boldsymbol{\gamma}^{(\ell)^T} \text{LeakyReLU}(\boldsymbol{W}_l^{(\ell)} \hat{h}_i^{(\ell)} + \boldsymbol{W}_r^{(\ell)} \hat{\boldsymbol{h}}_j^{(\ell)})\right), \quad (4)$$

where $\boldsymbol{\gamma}^{(\ell)}$, $\boldsymbol{W}_l^{(\ell)}$ and $\boldsymbol{W}_r^{(\ell)}$ are learnable parameters and $\sigma(\cdot)$ is the sigmoid function. This computation involves learning parameters that weigh the significance of each step's features in establishing a dependency with another step. The resulting weight matrix $\boldsymbol{P}^{(\ell)}$ explicitly represents the probabilities of dependency between steps in order to construct the final solution graph.

With these inter-step attention weights, we proceed to update the features of each node i by combining information from all graph nodes, which involves measuring how much weight is given to the connection between nodes i and j at the ℓ-th layer and then using these weights to update the features of node i:

$$\boldsymbol{h}_i^{(\ell)} = \text{ELU}\left(\sum_j \frac{\boldsymbol{P}_{i,j}^{(\ell)} \boldsymbol{W}_r^{(\ell)} \hat{\boldsymbol{h}}_j^{(\ell)}}{\sum_{j'} \boldsymbol{P}_{i,j'}^{(\ell)}}\right), \quad (5)$$

where ELU is the exponential linear unit function. This feature update allows the model to adaptively refine the representation of each node, incorporating insights from its connections in the solution graph.

By integrating both intra-step attention and inter-step attention mechanisms into SGAN's stack of attention layers, the model achieves a comprehensive understanding of the problem-solving procedure. The iterative refinement of the solution graph across these layers enables SGAN to progressively capture important features within individual steps and the relationships between the steps. This integration introduces a novel and powerful framework for SGAN to generate structured and coherent solutions for a wide range of vision-language problem solving tasks.

C. Learning Objectives

Our integrated attention mechanism progressively focuses on salient information in visual and textual inputs, capturing step dependencies for effective problem-solving. We propose novel learning objectives, supervising attention to identify important parts of the images and captions, and propagating information across steps for high-quality solution graphs.

1) Learning Intra-Step Attention: We present the attention learning loss to measure the prediction error of intra-step attention, based on the ground-truth multimodal attention annotations. These annotations are binary masks that indicate important image regions or word tokens in the captions. To measure the prediction error of the intra-step attention $\boldsymbol{\alpha}_i^{(\ell)}$, the intra-step attention loss is defined as

$$L_{att}^{(\ell)} = \sum_{i \in \mathcal{GT}} l_{att}(\boldsymbol{\alpha}_i^{(\ell)}, \boldsymbol{\alpha}_i'), \quad (6)$$

where $\mathcal{GT}$ is the set of ground-truth steps and l_{att} is a dissimilarity metric that measures the misalignment between the predicted $\boldsymbol{\alpha}_i^{(\ell)}$ and the softmax-normalized ground-truth attention $\boldsymbol{\alpha}_i'$ [23]. In our implementation, we define l_{att} as a cross-entropy loss:

$$L_{att}(\boldsymbol{\alpha}_i, \boldsymbol{\alpha}_i') = -\sum_k \boldsymbol{\alpha}_{i,k}' \log(\boldsymbol{\alpha}_{i,k}). \quad (7)$$

Similarly, other attention evaluation metrics like SIM [60], JSD [61], [62], and CC [60]) can also be used to measure the intra-step attention alignment.

2) Learning Inter-Step Attention: To gain deeper insights into the contributions of attention throughout the entire problem-solving process, we adopt an integrated approach that considers attention allocation across multiple problem-solving steps. Inspired by information retrieval techniques [45], [63], we introduce novel learning objectives that involve propagating the intra-step attention measurements along the edges of the predicted solution graph, quantifying the impact of attention in achieving successful solution prediction.

Specifically, given the ground truth solution graph represented as an adjacency matrix $\boldsymbol{G}$ and the inter-step attention $\boldsymbol{P}^{(\ell)}$ predicted by the ℓ-th layer, we compute $\boldsymbol{F}^{out^{(\ell)}}$ and $\boldsymbol{F}^{in^{(\ell)}}$ that denote the probabilities of information propagation along the ground-truth edges from step i, and those to step j, from out-degree and in-degree perspectives, respectively:

$$\boldsymbol{F}_{i,j}^{out^{(\ell)}} = \frac{\sum_k \boldsymbol{G}_{i,j} \boldsymbol{P}_{i,k}^{(\ell)}}{\sum_k \boldsymbol{P}_{i,k}^{(\ell)}}, \quad j = 0, \cdots, N+1 \quad (8)$$

$$\boldsymbol{F}_{i,j}^{in^{(\ell)}} = \frac{\sum_k \boldsymbol{G}_{k,j} \boldsymbol{P}_{k,j}^{(\ell)}}{\sum_k \boldsymbol{P}_{k,j}^{(\ell)}}, \quad i = 0, \cdots, N+1 \quad (9)$$

Based on these propagation probabilities, we define two inter-step attention scores that quantify the information flow from both in-degree and out-degree perspectives at the ℓ-th layer, respectively:

$$S_{out}^{(\ell)} = \text{mean}[(\boldsymbol{F}^{out^{(\ell)}} \odot \boldsymbol{D}^{(\ell)})^T \boldsymbol{s}^{(\ell)}], \tag{10}$$

$$S_{in}^{(\ell)} = \text{mean}[(\boldsymbol{D}^{(\ell)} \odot \boldsymbol{F}^{in^{(\ell)}})^T \boldsymbol{s}^{(\ell)}], \tag{11}$$

where $\odot$ represents the Hadamard product, $\boldsymbol{s}^{(\ell)} = [1, \boldsymbol{s}_1^{(\ell)}, \cdots, \boldsymbol{s}_N^{(\ell)}, 0]^T$ denotes an intra-step attention similarity measure, and $\boldsymbol{D}^{(\ell)}$ is a distribution matrix measuring the probability distribution of attention weights from step i to step j:

$$\boldsymbol{D}_{i,j}^{(\ell)} = \frac{\boldsymbol{G}_{i,j} \boldsymbol{P}_{i,j}^{(\ell)}}{\sum_k \boldsymbol{G}_{i,k}}. \tag{12}$$

Specifically, the similarity $\boldsymbol{s}_i^{(\ell)}$ is defined as

$$\boldsymbol{s}_i^{(\ell)} = 1 - \frac{\text{JSD}(\boldsymbol{\alpha}_i^{(\ell)}, \boldsymbol{\alpha}'_i)}{\ln 2}, \tag{13}$$

where JSD is the Jensen–Shannon divergence [61], [62].

The above inter-step attention scores $S_{out}^{(\ell)}$ and $S_{in}^{(\ell)}$ comprehensively quantify the performance of inter-step attention prediction from the out-degree and in-degree perspectives, where higher scores indicate that attention can be more effectively allocated over the important steps and dependencies to build the solution graph, and the maximum score of 1 indicates the perfect alignment with the ground-truth solution graph.

3) Overall Objectives: Our final objective function is defined as a combination of the binary cross entropy loss L_{BCE} that evaluates the solution graph, the intra-step attention loss $L_{att}{}^{(\ell)}$, and the inter-step attention scores $S_{out}^{(\ell)}$ and $S_{in}^{(\ell)}$ across all graph attention layers:

$$L = L_{\text{BCE}} + \sum_{\ell=1}^{L} L_{att}^{(\ell)} - \sum_{\ell=1}^{L} (S_{out}^{(\ell)} + S_{in}^{(\ell)}), \tag{14}$$

where

$$L_{\text{BCE}} = -\sum_{\ell=1}^{L} \sum_{i,j} (\boldsymbol{G}_{i,j} \log \boldsymbol{P}_{i,j}^{(\ell)} + (1 - \boldsymbol{G}_{i,j}) \log(1 - \boldsymbol{P}_{i,j}^{(\ell)})), \tag{15}$$

is the binary cross-entropy loss.

With this objective function, our method jointly and progressively supervises both intra-step attention and inter-step attention. It enables an integrated optimization of the solution with respect to multimodal attention alignment within individual problem-solving steps, information propagation for between-step connections, and the final solution graph. With the ability to traverse the graph and selectively aggregate information, our method achieves significant improvement in formulating solutions to various problems.

V. Experiments

In this section, we present comprehensive experiments to demonstrate the advantages of our proposed method and assess the contributions of its major components. The experimental results underscore the significance of progressive attention learning and the effectiveness of the proposed objectives, shedding light on the intricacies of complex problem-solving processes. These findings hold promise in substantially advancing the domain of vision-language problem solving and paving the way for more sophisticated intelligent systems.

A. Experimental Setup

In this subsection, we provide a thorough description of our experiments and implementation details. We introduce the dataset used for our multimodal problem-solving task, the compared state-of-the-art models and baselines, the evaluation methods, and the implementation details of our proposed SGAN method.

1) Dataset: Our experimental evaluation is conducted on the VisualHow dataset [23], which comprises 20,028 real-life problems categorized hierarchically into 18 main categories and 317 subcategories. The number of problems in each category ranges from 405 to 2,952, providing a diverse set of problem-solving scenarios. Unlike previous datasets [24]–[26], [28], [29], [35] that focus solely on sequential procedures, the VisualHow dataset includes a solution graph for each problem, representing the structured dependencies between individual steps. Importantly, a substantial portion of the graphs exhibit non-sequential characteristics, featuring more complex inter-step dependencies. Each solution graph consists of 3 to 10 steps, each described with images and captions. The images encompass a variety of formats, including realistic photos, cartoons, drawings, handwriting, charts, among others. The captions have a vocabulary of 30,000 tokens, ensuring rich and informative descriptions. To facilitate attention learning and evaluation, fine-grained attention annotations are provided for both images and captions.

2) Models: To evaluate the effectiveness of our method in handling vision-language problem-solving tasks, we compare it with state-of-the-art approaches on the VisualHow dataset [23]. We treat these methods as multi-task models, addressing both the retrieval of the multimodal instructions and the prediction of step dependencies. The compared methods, including SEQ GPO [64], SEQ GAP [23], and SEQ ATT [23], aim to predict individual problem-solving steps and their dependencies using various sequential processes. Specifically, SEQ GPO employs a generalized pooling operator to align visual and language features and jointly aggregates them during feature aggregation. Similarly, SEQ GAP adopts a global average pooling method to process features from different image regions and word tokens independently, without considering their importance. Finally, SEQ ATT utilizes an attention mechanism to highlight important semantics in each modality and then aggregates them based on learned weights, supervised with ground-truth attention annotations from VisualHow [23].

To further investigate the role and significance of the integrated attention mechanism, we conduct a comprehensive ablation study using three variants of our proposed model: SGAN-Base, SGAN-Intra, and SGAN-Inter. SGAN-Base is a basic model that uses the same architecture as SGAN but doesn't rely on any extra attention supervision from outside sources. This helps us understand how well the model performs when it learns attention on its own from the solution graph. For SGAN-Intra and SGAN-Inter, we supervise the model with the intra-step attention loss and inter-step attention loss terms, respectively. By comparing the performance of these three variants with our full SGAN model, which incorporates intra-step and inter-step attention supervision, we can analyze the specific contributions of each attention component.

3) Evaluation: To ensure a fair comparison with other methods, we adhere to the official training and validation splits provided by the dataset. We construct candidate pools by sampling images and captions from the corresponding subsets. These candidate pools include positive samples corresponding to the given problem and negative samples from other problems randomly sampled from the dataset. Note that the candidate pools contain only training data during the training phase, and only validation data during the validation phase. Different from the previous study [23] that samples unrelated steps from different problems, in this paper, to obtain negative step dependencies, we sample negative problems first, and include all steps and their dependencies in the negative problems. This approach serves as a suitable test bed for robustly evaluating and justifying the model's performance. Following the VisualHow [23] study and our proposed attention evaluation methods, we evaluate model performances with four categories of metrics:

Retrieval of Steps. To evaluate the performance models in retrieving the correct ground-truth steps, we rank the candidate steps based on their predicted relevance to the input problem (*i.e.*, $\boldsymbol{P}_{0,i}^{(L)},\ i = 1, 2, \cdots, N$). We employ the mean reciprocal rank (MRR) [17], [18], [23], Recall@K [17], [18], [23], [64]–[66], and recall sum (RSUM) [23], [64]–[66] metrics. The MRR computes the reciprocal rank of a correct step, which is defined as 1 divided by its position in the ranked list. Recall@K measures the presence of the correct step in the top-K ranked steps. The RSUM is defined as the sum of recall metrics at different values of K (*e.g.*, $K = \{1, 5, 10\}$). The combination of these metrics provides a comprehensive summary of the model's overall performance in image and caption retrieval.

Step Dependency Prediction. The prediction of dependencies between steps is evaluated using the area under the ROC curve (AUC) [23], [67], the area under the precision-recall curve (AUPR) [67], and the intersection over union (IoU) [23], [27], [68]. The AUC represents the overall performance of the model in distinguishing positive (correctly predicted edges) from negative (incorrectly predicted edges) dependencies between steps. The AUPR is a useful performance metric for imbalanced data in a setting with a bigger focus on positive examples, which is the case for our experiments. To measure IoU, we apply a threshold (e.g., 0.25, 0.5, 0.75) [23] to the model output $\boldsymbol{P}^{(L)}$ to determine the graph edges and count the edges for the intersection and union between the predicted graph and the ground truth. These metrics enable a comprehensive evaluation of the model's performance in predicting the structure of solutions.

Intra-Step Attention. To evaluate the intra-step attention, the output $\boldsymbol{\alpha}^{(L)}$ is first normalized and converted into an attention map, where each value indicates the attention probability of an image patch or word token. The ground-truth attention maps are computed similarly as the annotations. Three attention metrics are used to compute the attention maps: the linear Correlation Coefficient (CC) [60], [69] scores are computed as Pearson's linear correlation between the attention maps; the similarity of histogram intersection (SIM) [60] computes the sum of the minimum values at every location; Kullback-Leibler divergence (KL) [60] measures the difference between two distributions based on information theory.

Inter-Step Attention. The inter-step attention is evaluated based on the final-layer outputs $\boldsymbol{\alpha}^{(L)}$ and $\boldsymbol{P}^{(L)}$ simultaneously by three metrics that measure out-degree $S_{out}^{(L)}$ (see Equation (10)), in-degree $S_{in}^{(L)}$ (see Equation (11)) attention scores, and an overall attention score $S_{all}^{(L)}$ computed as

$$S_{all}^{(L)} = mean[(\boldsymbol{F}^{out(L)} \odot \boldsymbol{D}^{(L)} \odot \boldsymbol{F}^{in(L)})^T \boldsymbol{s}^{(L)}]. \tag{16}$$

4) Implementation Details: To extract discriminative visual-linguistic features, we adopt state-of-the-art pre-trained models. For the visual features, we use ResNeXT-101 [58] (32×8d) trained on Instagram images (WSL) [70], with image size 256×256. Regarding the language features, we use a pre-trained BERT model [48] optimized on a massive corpus of text. We use these models to extract features from the candidate image and caption pools, which are then used as inputs to our SGAN model. We train our model using the Adam [71] optimizer with learning rate 2×10^{-4}, weight decay 10^{-4} and batch size 16. A cosine annealing scheduler schedules the learning rate. We set $L = 3$ as the total number of network layers. To address the imbalance between the positive and negative samples from the solution graph, we train the model with the loss related to the retrieval task for 5 epochs and then train the model with the loss related to the whole solution graph for the remaining 20 epochs. A hard negative mining strategy [72], [73] is also used. The post-processing method to obtain the final solution graph is implemented following Algorithm 1, where we set dependency threshold $\lambda_d = 0.8$ and retrieval threshold $\lambda_r = 0.45$.

B. Quantitative Results

1) Comparison with the State-of-the-Art: Our approach demonstrates superior performance across all metrics for generalizing solutions to vision-language problems, as shown in Table I. Overall, it outperforms the state-of-the-art SEQ GPO, SEQ GAP, and SEQ ATT methods [23] across all evaluation metrics. In terms of retrieving multimodal instructions for individual problem-solving steps, it achieves an impressive improvement of 11.1% and 12.5% in MRR scores for images and captions, respectively, as well as an improvement of 10.9% in RSUM scores which aggregate the Recall@K scores over both modalities. Further, in terms of predicting the step dependencies, our method exhibits strong capability in capturing the diverse structures of solutions, which has been a challenge for existing methods. It shows 81.0% and 45.3% improvements in the average IoU scores (*i.e.*, 0.25, 0.5, and 0.75) for images and captions, respectively. These observations not only demonstrate the advantages of our approach in solving complex vision-language problems but also highlight the significance of progressively constructing task solutions.

TABLE I
SOLUTION GRAPH PREDICTION RESULTS FROM RETRIEVAL AND DEPENDENCY ASPECTS. IN EACH PANEL, THE FIRST ROW (I) INDICATES THE IMAGE MODALITY AND THE SECOND ROW (C) INDICATES THE CAPTION MODALITY. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Method	Mode	Retrieval ↑					Dependency ↑				
		MRR	R@1	R@5	R@10	RSUM	AUC	AUPR	IoU@0.25	IoU@0.5	IoU@0.75
SEQ GPO [64]	I	0.4529	31.03	61.89	79.99	386.28	0.713	0.768	0.455	0.262	0.102
	C	0.6066	47.25	77.27	88.85		0.816	0.847	0.544	0.443	0.269
SEQ GAP [23]	I	0.5311	39.28	69.55	85.30	404.72	0.685	0.752	0.444	0.250	0.080
	C	0.5935	45.55	76.75	88.29		0.817	0.850	0.557	0.430	0.258
SEQ ATT [23]	I	0.5069	37.10	66.48	82.82	410.77	0.698	0.760	0.422	0.274	0.106
	C	0.6509	51.94	82.02	91.41		0.815	0.849	0.520	0.432	0.263
SGAN-Base	I	0.5524	41.99	71.42	86.95	433.55	0.721	0.777	0.487	0.307	0.152
	C	0.6820	56.66	83.10	93.43		0.799	0.834	0.553	0.481	0.321
SGAN-Intra	I	0.5833	45.03	74.89	**88.93**	451.82	0.700	0.765	0.488	0.316	0.189
	C	0.7247	61.61	86.52	94.84		0.780	0.825	0.578	0.428	0.324
SGAN-Inter	I	0.5703	43.72	73.45	88.26	442.26	**0.801**	0.816	0.577	0.504	0.385
	C	0.6954	57.98	84.76	94.09		0.861	0.863	0.653	0.616	0.538
SGAN	I	**0.5898**	**45.97**	**75.61**	88.89	**455.56**	0.800	**0.817**	**0.580**	**0.508**	**0.394**
	C	**0.7324**	**62.77**	**86.95**	**95.37**		**0.862**	**0.864**	**0.659**	**0.620**	**0.547**

2) Comparison with Baseline Models: Table I also compares our proposed SGAN model with different baselines, including the SGAN-Base model that is learned without supervision from attention annotations, the SGAN-Intra model supervised with the intra-step attention loss, and the SGAN-Inter model supervised with the inter-step attention loss. The comparison shows that even without any external supervision, the SGAN-Base can still effectively learn the integrated attention from the ground-truth solution graph, and achieve promising results. Its MRR, RSUM, and IoU scores are all significantly better than those of the SEQ ATT method (*e.g.*, RSUM is improved from 410.77 to 433.55), demonstrating the effectiveness of the proposed network design. Notably, the introduction of either intra-step or inter-step attention supervision leads to substantial improvements. In particular, compared with SGAN-Base, SGAN-Intra achieves an improvement of 5.6% and 6.3% in MRR scores for images and captions, respectively. Its RSUM score is improved from 433.55 to 451.82, outperforming the SGAN-Base by 4.2%. These improvements suggest that the supervision of intra-step attention can benefit the localization of important information in both modalities. Furthermore, SGAN-Inter's performance highlights its practical significance in predicting step dependencies. With inter-step attention supervision, it achieves an impressive average improvement of 37.7% across AUC, AUPR, and IoU scores. This suggests the models' applicability in real-world scenarios where detailed annotations may be limited. Overall, incorporating both types of attention supervision achieves the best results, demonstrating the effectiveness of the integral design of our method in modeling attention for vision-language problem solving.

C. Qualitative Results

To further understand the proposed integrated attention mechanism and how it contributes to the prediction of problem-solving procedures, we conduct a qualitative comparison of the predicted solution graph and their intra-step attention maps. The qualitative examples are shown in Fig. 3, where the proposed SGAN method is compared with the state-of-the-art SEQ ATT [23] method and the ground truth. For a clearer illustration, we present the optimal predicted solution graph obtained from the image or caption candidate pool. The results consist of (1) the final solution graph obtained with Algorithm 1 showing the procedure flows across all steps, and (2) the intra-step attention maps for each problem-solving step overlaid on the images (*i.e.*, hot areas) and the captions (*i.e.*, bold text).

Despite leveraging explicit intra-step attention supervision based on fine-grained annotations, SEQ ATT sometimes fails to adequately attend to crucial objects relevant to problem-solving. As shown in Fig. 3, SEQ ATT allocates insufficient attention on the *conditioner* (see Fig. 3A, step 1), the *sugar* and *cocoa powder* (see Fig. 3B, step 1), the *structured meal plan* (see Fig. 3C, step 1), and the *rinse* action (see Fig. 3D, step 3). On the contrary, our proposed SGAN exhibits promising performance by attending to essential information within various steps. The comparison of intra-step attention between SEQ ATT and SGAN shows that progressively refining attention is effective in terms of learning accurate attention distribution in images and captions.

Furthermore, the inter-step attention mechanism is also shown to be effective in predicting the solution graph correctly. Because SEQ ATT sequentially predicts the dependencies one step at a time, it results in suboptimal solutions (see Fig. 3A-D). Differently, the integration of intra-step attention and inter-step attention in SGAN allows it to better understand the importance of key objects (*e.g.*, *conditioner*, *sugar*, *cocoa powder*, *structured meal plan*, *rinse*, *etc.*) across multiple steps. In addition, the progressive learning of the integrated attention mechanism allows SGAN to improve the solution graph by interactively refining it. Therefore, with a holistic view of the problem-solving procedure and interactive refinement, SGAN

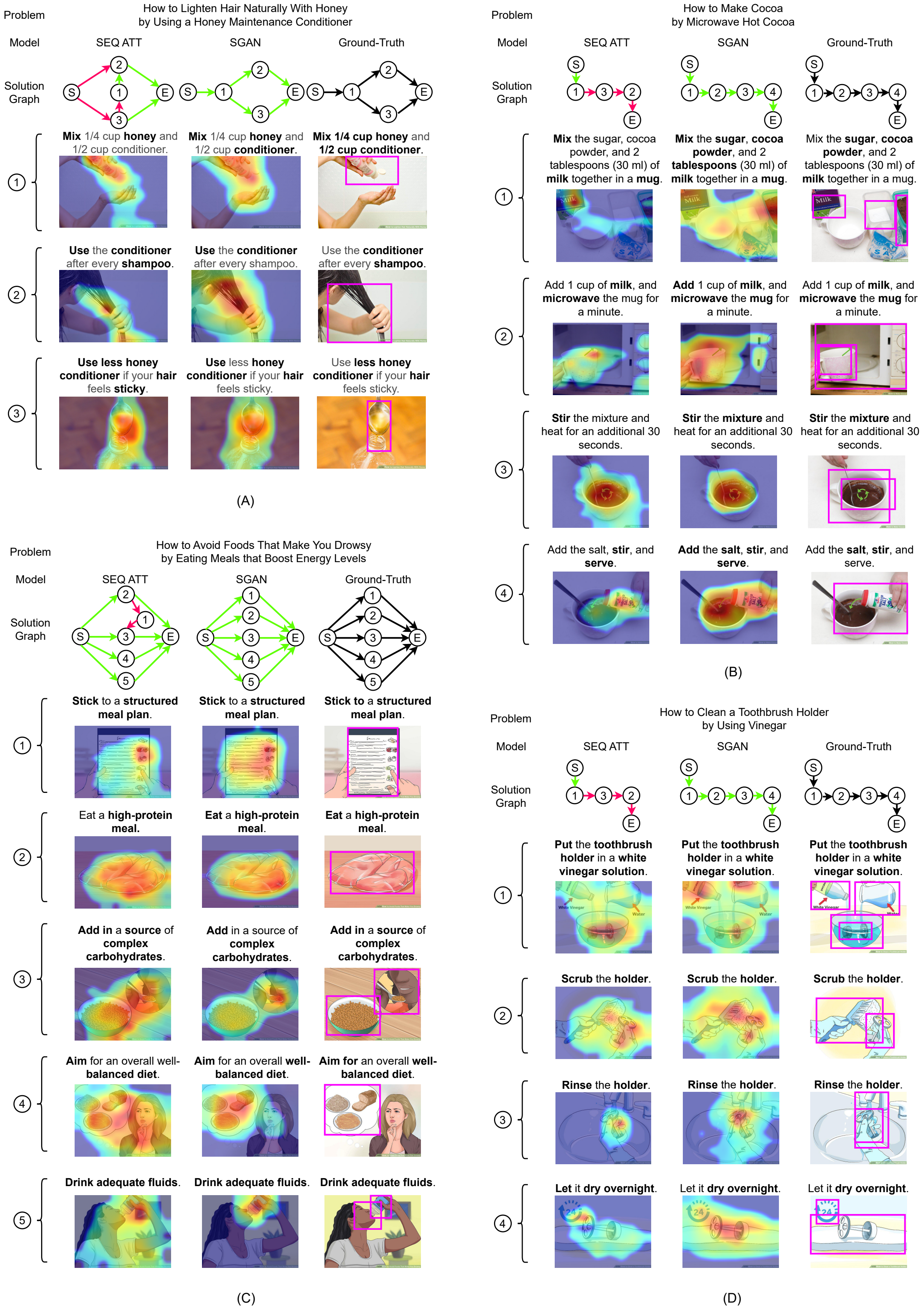
Problem
How to Lighten Hair Naturally With Honey by Using a Honey Maintenance Conditioner
Model
SEQ ATT
SGAN
Ground-Truth
Solution Graph
S
1
2
3
E
Mix 1/4 cup honey and 1/2 cup conditioner.
Use the conditioner after every shampoo.
Use less honey conditioner if your hair feels sticky.
(A)
Problem
How to Make Cocoa by Microwave Hot Cocoa
Model
SEQ ATT
SGAN
Ground-Truth
Solution Graph
S
1
2
3
4
E
Mix the sugar, cocoa powder, and 2 tablespoons (30 ml) of milk together in a mug.
Add 1 cup of milk, and microwave the mug for a minute.
Stir the mixture and heat for an additional 30 seconds.
Add the salt, stir, and serve.
Milk
(B)
Problem
How to Avoid Foods That Make You Drowsy by Eating Meals that Boost Energy Levels
Model
SEQ ATT
SGAN
Ground-Truth
Solution Graph
S
1
2
3
4
5
E
Stick to a structured meal plan.
Eat a high-protein meal.
Add in a source of complex carbohydrates.
Aim for an overall well-balanced diet.
Drink adequate fluids.
(C)
Problem
How to Clean a Toothbrush Holder by Using Vinegar
Model
SEQ ATT
SGAN
Ground-Truth
Solution Graph
S
1
2
3
4
E
Put the toothbrush holder in a white vinegar solution.
White Vinegar
Water
Scrub the holder.
Rinse the holder.
Let it dry overnight.
24
(D)

Fig. 3. Qualitative comparison of the predicted solution graphs and intra-step attention maps. The green edges in the graphs indicate correct predictions, while the red ones indicate wrong predictions.

TABLE II
SOLUTION GRAPH PREDICTION RESULTS FOR SEQUENTIAL AND NON-SEQUENTIAL SOLUTIONS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Method	Mode	Sequence		Non-Sequence	
		MRR ↑	IoU@0.5 ↑	MRR ↑	IoU@0.5 ↑
SEQ GPO [64]	I C	0.4435 0.5707	0.190 0.326	0.4640 0.6487	0.337 0.564
SEQ GAP [23]	I C	0.5338 0.5584	0.181 0.302	0.5280 0.6346	0.321 0.563
SEQ ATT [23]	I C	0.4968 0.6130	0.187 0.286	0.5187 0.6952	0.363 0.584
SGAN	I C	**0.5823** **0.7087**	**0.404** **0.537**	**0.5986** **0.7602**	**0.616** **0.706**

manages to predict the dependencies more accurately.

D. Performance Analyses

We further present extensive analyses to understand the roles and contributions of different components in our proposed approach. Through these in-depth analyses, we aim to gain a deeper understanding of the key factors that contribute to the success of our approach in solving complex vision-language problems.

1) Sequential and Non-Sequential Solutions: Unlike previous datasets that focus on sequential solutions, VisualHow is a unique dataset that contains a variety of complex problem-solving tasks. To demonstrate the effectiveness of our proposed method on different types of solution structures, we present the model's performance on both sequential and non-sequential solutions separately. In Table II, we evaluate the performance of our method in both sequential and non-sequential problem-solving scenarios. The results show that our SGAN method outperforms the state-of-the-art methods in both scenarios, achieving the highest MRR and IoU@0.5 scores. This demonstrates that SGAN excels in capturing the structure and dependencies of solution steps, regardless of whether the problem-solving process is sequential or not, making it a versatile approach for a wide range of real-world applications that involve complex structures and diverse multimodal instructions.

2) Intra-Step Attention: The results presented in Table III provide insights into the performance of our intra-step attention mechanism. The attention output $\boldsymbol{\alpha}^{(L)}$ is evaluated using three metrics: CC, KLD, and SIM to quantify the quality of intra-step attention learning and help assess the effectiveness of this method in focusing on salient information. The state-of-the-art method SEQ ATT [23], which also learns intra-step attention following a sequential approach, achieves moderate results for both image and caption modalities. However, our proposed SGAN with intra-step attention (SGAN-Intra) outperforms SEQ ATT consistently across almost all the metrics (5/6) for both modalities. This demonstrates that the progressive refinement of the solution graph with intra-step attention enables the model to focus on relevant information within each step, leading to improved attention quality. On the other hand, the impact of inter-step attention (SGAN-Inter) alone is not as significant on these evaluation metrics. However, integrating the two attention mechanisms is able to further improve the model's ability in finding important information in the images and captions. This highlights the importance of combining both attention mechanisms to achieve a comprehensive understanding of the problem-solving procedure.

TABLE III
INTRA-STEP ATTENTION EVALUATION RESULTS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Method	Mode	Intra-Step Attention		
		CC ↑	SIM ↑	KLD ↓
SEQ ATT	I C	0.600 0.764	0.571 0.623	0.705 0.616
SGAN-Base	I C	0.133 0.307	0.401 0.337	1.540 1.915
SGAN-Intra	I C	**0.611** 0.768	0.586 0.656	0.668 0.648
SGAN-Inter	I C	0.456 0.726	0.523 0.625	0.880 0.740
SGAN	I C	0.608 **0.772**	**0.588** **0.657**	**0.666** **0.605**

TABLE IV
INTER-STEP ATTENTION EVALUATION RESULTS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Method	Mode	Inter-Step Attention		
		$S_{in}^{(L)}$ ↑	$S_{out}^{(L)}$ ↑	$S_{all}^{(L)}$ ↑
SEQ ATT	I C	0.0696 0.1313	0.0551 0.0998	0.0102 0.0244
SGAN-Base	I C	0.0766 0.1102	0.0627 0.0934	0.0209 0.0309
SGAN-Intra	I C	0.0878 0.1354	0.0654 0.0971	0.0125 0.0217
SGAN-Inter	I C	0.2017 0.3081	0.2217 0.3158	0.1366 0.2288
SGAN	I C	**0.2220** **0.3214**	**0.2360** **0.3274**	**0.1493** **0.2390**

3) Inter-Step Attention: Understanding how attention is aligned across multiple steps in complex problem-solving is crucial for developing effective learning models. Here, we provide a detailed analysis of our method by examining the attention alignment between steps. Table IV presents the results of the inter-step attention evaluation, which sheds light on the model's ability to capture dependencies between problem-solving steps. The metrics used to evaluate inter-step attention include $S_{in}^{(L)}$, $S_{out}^{(L)}$, and $S_{all}^{(L)}$, which quantify the quality of attention propagation within the solution graph. The state-of-the-art method SEQ ATT [23] exhibits limited performance in capturing inter-step dependencies, as evidenced by the relatively low values of all metrics for both image and caption modalities. This is because the sequential design of SEQ ATT cannot effectively propagate attention to other steps across multiple steps, resulting in suboptimal predictions. However, the most significant improvement is observed with the addition of inter-step attention in the SGAN-Inter model. The values

TABLE V
PEARSON'S r BETWEEN ATTENTION EVALUATION SCORE AND OUR PROPOSED SGAN MODEL'S PERFORMANCE. BOLD NUMBERS INDICATE SIGNIFICANT POSITIVE CORRELATIONS ($p < 0.05$).

Attention Type	Mode	Sequence		Non-Sequence	
		MRR	IoU@0.5	MRR	IoU@0.5
Intra-step	I	-0.068	**0.212**	-0.001	0.001
	C	-0.091	**0.236**	-0.048	-0.058
Inter-step	I	**0.677**	**0.417**	**0.625**	**0.738**
	C	**0.629**	**0.435**	**0.607**	**0.732**

of $S_{in}^{(L)}$, $S_{out}^{(L)}$, and $S_{all}^{(L)}$ for SGAN-Inter are notably higher than those of SEQ ATT, SGAN-Base, and SGAN-Intra. The full SGAN model, which combines both intra-step and inter-step attention mechanisms, achieves the best results among all methods and modalities across all metrics. These observations indicate that the inter-step attention mechanism effectively captures the dependencies between problem-solving steps, allowing the attended information to effectively propagate across multiple steps, leading to improved reasoning about the chronological order of various solution steps.

4) Correlation Between Attention Performance and Task Performance: To further investigate how the intra-step and inter-step attention contribute to the model performance in tackling vision-language problems, we compute the Pearson's r between the attention evaluation scores CC, $S_{all}^{(L)}$ and task evaluation scores MRR and IoU@0.5. Table V shows the Pearson's correlation coefficient (r) between the attention evaluation scores and the performance of our proposed SGAN model on predicting sequential solutions and non-sequential ones. For the intra-step attention evaluation, we observe a significant positive correlation between attention performance and model's ability to predict the dependencies in sequential solutions. The correlation coefficients for IoU@0.5 are 0.212 and 0.238 for the image and caption modalities, respectively. On non-sequential problems, the correlation coefficients are close to zero, indicating a weak correlation between intra-step attention performance and model performance. The weak correlations suggest that in the final SGAN model, the quality of attention within individual steps has limited impacts on the model's performance. In contrast, the inter-step attention evaluation shows strong positive correlations between attention performance and model performance on both sequential and non-sequential solutions. In particular, for non-sequential ones, attention performance is highly correlation with the IoU@0.5, with values of 0.738 and 0.732 for the image and caption modalities, respectively. The strong positive correlations suggest that the quality of inter-step attention is closely related to the model's ability to capture dependencies between problem-solving steps and predict coherent and structured solutions. These results indicate that the inter-step attention mechanism plays a crucial role in improving the model's performance on both sequential and non-sequential problems.

5) Number of Attention Layers: Progressively refining attention is a fundamental component of our proposed SGAN architecture, which enables the network to iteratively focus on key information within from visual and textual inputs and discover the dependencies between the steps. To verify the effect of the number of integrated attention layers, we conduct experiments with four variants of our models. As shown in Table VI, for the retrieval of the most relevant images and captions, increasing the number of attention layers consistently improves the model's performance. We observe that with three attention layers, the SGAN model achieves the highest MRR, Recall@K, and RSUM scores for both the image and caption modalities. However, adding more layers does not lead to further improvements in the retrieval performance. Similar trends are observed for the evaluation of step dependencies, with AUC, AUPR, and IoU scores. Overall, this ablation study demonstrates that a three-layer SGAN model results in the right balance between capturing relevant information within individual steps and modeling the dependencies between steps. This configuration achieves the best performance for both retrieval and dependency aspects, indicating its effectiveness in tackling complex multimodal problem-solving tasks.

6) Progressive Attention Refinement Across Layers: Gradually refining attention constitutes an important element within our proposed SGAN method, empowering the model to progressively concentrate on key information across visual and textual inputs, unraveling inter-step dependencies. To illustrate the effectiveness of progressively refining attention in our proposed SGAN, we compare the outputs of different layers, including the intra-step attention $\boldsymbol{\alpha}^{(\ell)}$ and the inter-step attention $\boldsymbol{P}^{(\ell)}$ ($\ell = 1, 2, 3$). As shown in Table VII, we find that the attention alignments (Intra-Step Attention and Inter-Step Attention) exhibit a progressive enhancement as the layers delve deeper. This suggests that, with each subsequent layer, the model refines its ability to focus on relevant information, capturing more detailed relationships. This refinement in attention aligns with an observed improvement in prediction performance metrics, including MRR and IoU@0.5, suggesting the significance of this progressive attention mechanism in the success of problem solving.

7) Proportion of Attention Annotations: In Table I, we have demonstrated that SGAN-Base can self-learn attention from the solution graph, which has performed better than the SEQ ATT [23] model that requires additional attention annotations, while learning from annotations with the proposed objectives can further improve the model's performance. To study the impact of the annotations on model performance, we use different proportions of annotations in training, ranging from 0% to 100%, and evaluate the model's performance using various metrics. Table VIII presents the results of our ablation study on the proportion of fine-grained data annotations used in training the SGAN model. For both the retrieval and the dependency evaluations, we observe that all evaluation scores increase steadily with a higher proportion of fine-grained annotations. This indicates that providing more detailed annotations enhances the model's ability to accurately retrieve multimodal instructions for individual problem-solving steps, as well as to better predict the structured dependencies between steps.

8) Using Pre-Trained Grounding As Attention Annotations: Although providing more attention annotations can improve model performance, the practicality of obtaining such annota-

TABLE VI
ABLATION STUDY OF THE NUMBER OF INTEGRATED ATTENTION LAYERS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Number	Mode	Retrieval ↑					Dependency ↑				
		MRR	R@1	R@5	R@10	RSUM	AUC	AUPR	IoU@0.25	IoU@0.5	IoU@0.75
1	I	0.5649	42.90	73.11	87.52	432.90	0.781	0.795	0.501	0.427	0.308
	C	0.6624	53.62	82.62	93.13		0.845	0.847	0.582	0.515	0.413
2	I	0.5846	45.00	75.35	88.46	451.77	0.796	0.809	0.553	0.482	0.369
	C	0.7224	61.26	86.83	94.86		0.855	0.854	0.629	0.583	0.504
3	I	**0.5898**	**45.97**	**75.61**	**88.89**	**455.56**	**0.800**	**0.817**	**0.580**	**0.508**	**0.394**
	C	**0.7324**	**62.77**	**86.95**	**95.37**		**0.862**	**0.864**	**0.659**	**0.620**	**0.547**
4	I	0.5815	44.98	74.38	88.08	448.05	0.799	0.811	**0.580**	0.491	0.363
	C	0.7169	60.89	85.36	94.36		0.861	0.858	0.655	0.618	0.532

TABLE VII
EVALUATIONS ON INTRA-STEP ATTENTION, INTER-STEP ATTENTION, AND SOLUTION GRAPH PREDICTION RESULTS ACROSS LAYERS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Layer	Mode	Intra-Step Attention			Inter-Step Attention			Retrieval	Dependency
		CC ↑	SIM ↑	KLD ↓	$S_{in}^{(L)}$ ↑	$S_{out}^{(L)}$ ↑	$S_{all}^{(L)}$ ↑	MRR ↑	IoU@0.5 ↑
1	I	0.607	0.587	0.669	0.1994	0.2039	0.1366	0.5867	0.450
	C	0.764	0.656	0.608	0.2981	0.2993	0.2275	0.7324	0.569
2	I	0.607	0.588	0.667	0.2201	0.2297	**0.1512**	0.5862	0.488
	C	0.765	0.656	0.606	0.3191	0.3210	**0.2421**	**0.7327**	0.601
3	I	**0.608**	**0.588**	**0.666**	**0.2220**	**0.2360**	0.1493	**0.5898**	**0.508**
	C	**0.772**	**0.657**	**0.605**	**0.3214**	**0.3274**	0.2390	0.7324	**0.620**

TABLE VIII
ABLATION STUDY OF THE PROPORTION OF FINE-GRAINED DATA ANNOTATIONS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Proportion	Mode	Retrieval ↑					Dependency ↑				
		MRR	R@1	R@5	R@10	RSUM	AUC	AUPR	IoU@0.25	IoU@0.5	IoU@0.75
0%	I	0.5524	41.99	71.42	86.95	433.55	0.721	0.777	0.487	0.307	0.152
	C	0.6820	56.66	83.10	93.43		0.799	0.834	0.553	0.481	0.321
20%	I	0.5551	42.12	71.84	87.00	434.13	0.774	0.803	0.546	0.405	0.286
	C	0.6829	56.65	83.50	93.03		0.834	0.849	0.612	0.453	0.383
40%	I	0.5875	45.48	75.35	**89.35**	453.59	0.786	0.810	0.548	0.423	0.299
	C	0.7272	61.95	86.58	94.87		0.846	0.857	0.635	0.565	0.443
60%	I	0.5890	45.94	75.14	89.12	454.29	0.793	0.814	0.560	0.457	0.335
	C	0.7286	62.26	86.92	94.92		0.853	0.859	0.643	0.593	0.488
80%	I	0.5863	45.52	74.78	89.02	454.21	0.796	0.815	0.570	0.484	0.367
	C	**0.7334**	**63.16**	86.61	95.12		0.859	0.863	0.658	0.613	0.527
100%	I	**0.5898**	**45.97**	**75.61**	88.89	**455.56**	**0.800**	**0.817**	**0.580**	**0.508**	**0.394**
	C	0.7324	62.77	**86.95**	**95.37**		**0.862**	**0.864**	**0.659**	**0.620**	**0.547**

tions may raise scalability concerns. To address this, instead of leveraging human annotations, we generate ground-truth attention annotations using a pre-trained GLIP [74] model, which exhibits strong zero-shot and few-shot transferability to diverse object-level recognition tasks. As shown in Table IX, the GLIP-generated annotations demonstrate comparable performance as the human annotations from the VisualHow dataset. This consistency suggests that large pre-trained vision-language models can provide sufficient attention annotations for modeling intra-step attention across various problems, offering a viable approach to scalability.

9) Multimodal Procedure Planning Models: Table X compares the performance of our method with state-of-the-art multimodal procedure planning models, including Text-Image Prompting (TIP) [75] and Skip-Plan [39]. TIP generates a sequence of step captions using the text-davinci-003 model [76], and subsequently converting these captions into images using Stable Diffusion [77]. Skip-Plan learns to predict solutions by breaking down a long chain of steps into several reliable sub-chains, addressing error accumulation in long sequence predictions. Since these sequential methods cannot handle complex graph structures, we only compare them with our method by evaluating them through image and caption retrieval. As shown in Table X, there is a notable discrepancy in the retrieval capabilities of the TIP model between image and caption retrieval tasks, indicating a greater proficiency in processing

TABLE IX
SOLUTION GRAPH PREDICTION RESULTS WITH DIFFERENT SOURCES OF ATTENTION ANNOTATIONS. IN EACH PANEL, THE FIRST ROW (I) INDICATES THE IMAGE MODALITY AND THE SECOND ROW (C) INDICATES THE CAPTION MODALITY. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Source	Mode	Retrieval ↑					Dependency ↑				
		MRR	R@1	R@5	R@10	RSUM	AUC	AUPR	IoU@0.25	IoU@0.5	IoU@0.75
GLIP [74]	I	**0.5907**	**46.06**	75.39	**89.28**	453.21	0.797	0.812	0.572	0.507	**0.395**
	C	0.7238	61.52	86.32	94.65		0.852	0.856	0.648	0.495	0.451
VisualHow [23]	I	0.5898	45.97	**75.61**	88.89	**455.56**	**0.800**	**0.817**	**0.580**	**0.508**	0.394
	C	**0.7324**	**62.77**	**86.95**	**95.37**		**0.862**	**0.864**	**0.659**	**0.620**	**0.547**

TABLE X
COMPARISON OF MULTIMODAL PROCEDURE PLANNING MODELS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Model	Mode	Retrieval ↑				
		MRR	R@1	R@5	R@10	RSUM
TIP [75]	I	0.4046	29.09	50.72	64.06	377.67
	C	0.7124	62.08	82.40	89.32	
Skip-Plan [39]	I	0.4819	33.72	64.21	82.25	384.08
	C	0.5571	40.92	74.24	88.76	
SGAN	I	**0.5898**	**45.97**	**75.61**	**88.89**	**455.56**
	C	**0.7324**	**62.77**	**86.95**	**95.37**	

and extracting information from textual data compared to visual inputs. On the other hand, the Skip-Plan model exhibits an improved retrieval performance, a result of its end-to-end training on the VisualHow [23] dataset. However, these state-of-the-art procedure planning methods still underperform our SGAN model, because of their sequential nature. The graph-based model architecture and the novel attention mechanisms allow SGAN to capitalize on the extensive in-domain problem-solving knowledge embedded in the VisualHow dataset [23], achieving a significant performance improvement. This improvement solidifies SGAN's status as a promising solution for effectively addressing multimodal complexities in problem solving.

10) Similarity Functions used in Attention Learning: In this ablation study, we investigate the impact of adopting different attention evaluation metrics on attention learning. We consider three widely used similarity functions: SIM [60], JSD [61], [62], and CC [60], which are applied to supervise the inter-step attention mechanism in our proposed SGAN model. The results in Table XI demonstrate that our attention supervision method is robust against the choice of similarity function, as all three metrics produce similar performance. This consistency in performance indicates that our method effectively captures the attention alignment from different perspectives, leading to comparable results regardless of the selected similarity function. Based on these findings, we adopt JSD similarity in the measurement of inter-step attention. Overall, these results affirm the effectiveness of our approach in measuring attention alignment from multiple angles. This versatility is crucial for the success of our SGAN model in solving complex multimodal problem-solving tasks, as it allows the model to capture fine-grained dependencies between individual solution steps, leading to more accurate and coherent predictions.

11) Graph Post-Processing Thresholds: Finally, we investigate the impacts of the thresholds (*i.e.*, dependency threshold λ_d and retrieval threshold λ_r) on the predicted solution graph. It is noteworthy that following the VisualHow [23] benchmark, quantitative results presented in this paper, including the evaluation of retrieval, dependency, intra-step attention, and inter-step attention, are based on the probabilistic output $\boldsymbol{P}^{(L)}$. The dependency threshold λ_d and retrieval threshold λ_r are only used to binarize the soft probabilities into the final deterministic solution graph. In Table XII, we show various threshold combinations and their corresponding precision, recall, and $F1$ scores computed with the final solution graph. These scores are derived from comparing the ground-truth solution graph with binarized solution graphs after post-processing. The analysis reveals that the final solution graphs are not significantly affected by the choice of the dependency threshold λ_d ($0.2 \leq \lambda_d \leq 1.1$). The retrieval threshold λ_r acts as a balancing factor between precision and recall, and the final solution graphs are not sensitive to the choice of it ($0.05 \leq \lambda_r \leq 0.65$). Based on this observation, we empirically choose $\lambda_d = 0.8$ and $\lambda_r = 0.45$ for our experiment.

VI. CONCLUSION

In this paper, we focus on addressing existing gaps in understanding and providing effective step-by-step instructions for problem-solving in vision-and-language applications. Our contribution is a novel Solution Graph Attention Network (SGAN) approach that takes into account both intra-step and inter-step attention mechanisms, enabling a progressive construction of solutions by refining the dependencies between relevant problem-solving steps. The flexibility of our method allows for the formulation of solutions with various structures, accommodating both sequential and non-sequential patterns. In order to enhance the accuracy of attention in the problem-solving process, we have introduced quantitative metrics to study the role of attention in task accomplishment. These metrics serve as valuable tools for attention supervision, providing insights into how attention mechanisms can be leveraged effectively.

Our experimental results showcase the advantages of our proposed method in tackling a wide range of vision-language problems. By employing our model, we achieved significant improvements in formulating solutions with complex graph structures. Moreover, our findings shed light on the crucial components that contribute to successful problem-solving, thus offering valuable insights for future research and applications.

TABLE XI
ABLATION STUDY OF SIMILARITY FUNCTIONS USED IN THE PROPOSED EVALUATION METRICS.

Similarity	Mode	Retrieval ↑					Dependency ↑				
		MRR	R@1	R@5	R@10	RSUM	AUC	AUPR	IoU@0.25	IoU@0.5	IoU@0.75
SIM [60]	I	0.5896	45.81	75.74	89.03	455.44	0.796	0.815	0.576	0.485	0.377
	C	0.7310	62.48	87.22	95.17		0.861	0.865	0.657	0.615	0.531
JSD [61], [62]	I	0.5898	45.97	75.61	88.89	455.56	0.800	0.817	0.580	0.508	0.394
	C	0.7324	62.77	86.95	95.37		0.862	0.864	0.659	0.620	0.547
CC [60]	I	0.5921	46.24	75.52	88.89	456.62	0.798	0.816	0.579	0.507	0.389
	C	0.7370	63.47	87.14	95.37		0.862	0.863	0.659	0.621	0.553

TABLE XII
ABLATION STUDY ON DIFFERENT COMBINATIONS OF DEPENDENCY THRESHOLD λ_d AND RETRIEVAL THRESHOLD λ_r. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

λ_d	λ_r	Precision	Recall	$F1$
0.2	0.45	0.477	0.605	0.499
0.5		0.477	0.605	0.499
0.8		0.482	0.601	**0.500**
1.1		0.490	0.586	0.498
1.4		0.495	0.529	0.473
0.8	0.05	0.444	0.650	0.493
	0.25	0.462	0.629	0.499
	0.45	0.482	0.601	**0.500**
	0.65	0.505	0.557	0.494
	0.85	0.525	0.473	0.463

We believe that the insights gained from our work will have a profound impact on solving intricate visual problems and providing effective guidance for various daily-life activities. Our method not only advances the state-of-the-art in vision-language problem solving, but also lays the groundwork for the development of more powerful and flexible attention mechanisms. With the hope that our work will inspire further advancements in this field, we envision that our proposed GNN-based model and attention supervision techniques will continue to drive progress in solving problems more effectively and efficiently.

While our proposed method shows promising results in tackling vision-language problem-solving tasks, it also has several limitations and opens up interesting avenues for future research. One limitation is that our method relies on annotated data for training and supervision. We have explored GLIP-generated annotations to reduce the data dependency and improve the generalization capabilities of our model, which has shown promising results. Another challenge we face in this work is that the dependencies between steps may not always be clear-cut. There can be cases where multiple possible dependencies exist, leading to ambiguity in constructing the solution graph. Developing methods to handle such ambiguity and effectively capture uncertain dependencies is an important direction for future research.

ACKNOWLEDGEMENTS

This work is supported by NSF Grant 2143197.

REFERENCES

[1] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "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 (CVPR)*, 2018.

[2] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and C. L. Zitnick, "Microsoft COCO captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.

[3] M. Cornia, M. Stefanini, L. Baraldi, and R. Cucchiara, "M^2: Meshed-memory transformer for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[4] S. Chen and Q. Zhao, "Boosted attention: Leveraging human attention for image captioning," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[5] X. Chen, M. Jiang, and Q. Zhao, "Leveraging human attention in novel object captioning," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

[6] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni, and R. Cucchiara, "From show to tell: A survey on deep learning-based image captioning," *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2023.

[7] T.-H. K. Huang, F. Ferraro, N. Mostafazadeh, I. Misra, A. Agrawal, J. Devlin, R. Girshick, X. He, P. Kohli, D. Batra, C. L. Zitnick, D. Parikh, L. Vanderwende, M. Galley, and M. Mitchell, "Visual storytelling," in *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, 2016.

[8] T. Li, H. Wang, B. He, and C. W. Chen, "Knowledge-enriched attention network with group-wise semantic for visual storytelling," *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2023.

[9] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Batra, and D. Parikh, "VQA: Visual question answering," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015.

[10] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA matter: Elevating the role of image understanding in visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[11] A. Agrawal, D. Batra, D. Parikh, and A. Kembhavi, "Don't just assume; look and answer: Overcoming priors for visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[12] D. A. Hudson and C. D. Manning, "GQA: A new dataset for real-world visual reasoning and compositional question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[13] S. Chen, M. Jiang, J. Yang, and Q. Zhao, "AiR: Attention with reasoning capability," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[14] ——, "Attention in reasoning: Dataset, analysis, and modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2021.

[15] L. Chen, Y. Zheng, Y. Niu, H. Zhang, and J. Xiao, "Counterfactual samples synthesizing and training for robust visual question answering," *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2023.

[16] Y. Zhou, R. Ji, X. Sun, J. Su, D. Meng, Y. Gao, and C. Shen, "Plenty is plague: Fine-grained learning for visual question answering,"

IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI), 2022.

[17] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. F. Moura, D. Parikh, and D. Batra, "Visual dialog," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[18] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, S. Lee, J. M. F. Moura, D. Parikh, and D. Batra, "Visual dialog," *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2019.

[19] D. Guo, H. Wang, and M. Wang, "Context-aware graph inference with knowledge distillation for visual dialog," *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2022.

[20] S. Gehrmann, E. Clark, and T. Sellam, "Repairing the cracked foundation: A survey of obstacles in evaluation practices for generated text," *arXiv preprint arXiv:2202.06935*, 2022.

[21] S. Oraby, V. Harrison, A. Ebrahimi, and M. Walker, "Curate and generate: A corpus and method for joint control of semantics and style in neural NLG," in *Annual Conference of the Association for Computational Linguistics (ACL)*, 2019.

[22] S. Welleck, I. Kulikov, S. Roller, E. Dinan, K. Cho, and J. Weston, "Neural text generation with unlikelihood training," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[23] J. Yang, X. Chen, M. Jiang, S. Chen, L. Wang, and Q. Zhao, "VisualHow: Multimodal problem solving," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[24] K. R. Chandu, R.-P. Dong, and A. Black, "Reading between the lines: Exploring infilling in visual narratives," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[25] Y. Tang, D. Ding, Y. Rao, Y. Zheng, D. Zhang, L. Zhao, J. Lu, and J. Zhou, "COIN: A large-scale dataset for comprehensive instructional video analysis," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[26] Y. Tang, J. Lu, and J. Zhou, "Comprehensive instructional video analysis: The COIN dataset and performance evaluation," *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2020.

[27] C.-Y. Chang, D.-A. Huang, D. Xu, E. Adeli, L. Fei-Fei, and J. C. Niebles, "Procedure planning in instructional videos," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[28] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[29] L. Zhou, C. Xu, and J. J. Corso, "Towards automatic learning of procedures from web instructional videos," in *Association for the Advancement of Artificial Intelligence (AAAI)*, 2018.

[30] A. Miech, D. Zhukov, and J.-B. Alayrac, "The language of actions: Recovering the syntax and semantics of goal-directed human activities," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[31] H. Zhou, R. Martín-Martín, M. Kapadia, S. Savarese, and J. C. Niebles, "Procedure-aware pretraining for instructional video understanding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[32] H. Wang, Y. Wu, S. Guo, and L. Wang, "PDPP: Projected diffusion for procedure planning in instructional videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[33] A. Zala, J. Cho, S. Kottur, X. Chen, B. Oğuz, Y. Mehdad, and M. Bansal, "Hierarchical video-moment retrieval and step-captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[34] F. Sener, R. Saraf, and A. Yao, "Transferring knowledge from text to video: Zero-shot anticipation for procedural actions," *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2023.

[35] D. Zhukov, J.-B. Alayrac, R. G. Cinbis, D. Fouhey, I. Laptev, and J. Sivic, "Cross-task weakly supervised learning from instructional videos," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[36] N. Dvornik, I. Hadji, H. Pham, D. Bhatt, B. Martinez, A. Fazly, and A. D. Jepson, "Graph2Vid: Flow graph to video grounding for weakly-supervised multi-step localization," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[37] J. Sun, D.-A. Huang, B. Lu, Y.-H. Liu, B. Zhou, and A. Garg, "PlaTe: Visually-grounded planning with transformers in procedural tasks," *IEEE Robotics and Automation Letters*, 2021.

[38] H. Zhao, I. Hadji, N. Dvornik, K. G. Derpanis, R. P. Wildes, and A. D. Jepson, "P^{3}IV: Probabilistic procedure planning from instructional videos with weak supervision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[39] Z. Li, W. Geng, M. Li, L. Chen, Y. Tang, J. Lu, and J. Zhou, "Skip-Plan: Procedure planning in instructional videos via condensed action space learning," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[40] A. Das, H. Agrawal, C. L. Zitnick, D. Parikh, and D. Batra, "Human attention in visual question answering: Do humans and deep networks look at the same regions?" in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016.

[41] R. R. Selvaraju, S. Lee, Y. Shen, H. Jin, S. Ghosh, L. Heck, D. Batra, and D. Parikh, "Taking a HINT: Leveraging explanations to make vision and language models more grounded." in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[42] J. Wu and R. Mooney, "Self-critical reasoning for robust visual question answering," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.

[43] H. Tan and M. Bansal, "LXMERT: Learning cross-modality encoder representations from transformers," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[44] S. Brody, U. Alon, and E. Yahav, "How attentive are graph attention networks?" in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

[45] P. A. Gagniuc, *Markov Chains: From Theory to Implementation and Experimentation*. John Wiley & Sons, 2017.

[46] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018.

[47] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," *International Conference on Machine Learning (ICML)*, 2015.

[48] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *North American Chapter of the Association for Computational Linguistics (NAACL)*, 2019.

[49] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2017.

[50] A. Arnab, C. Sun, and C. Schmid, "Unified graph structured models for video understanding," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021.

[51] F. Chen, X. Chen, F. Meng, P. Li, and J. Zhou, "GoG: Relation-aware graph-over-graph network for visual dialog," in *Findings of Annual Conference of the Association for Computational Linguistics (Findings of ACL)*, 2021.

[52] L. Li, Z. Gan, Y. Cheng, and J. Liu, "Relation-aware graph attention network for visual question answering," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[53] L. Peng, S. Yang, Y. Bin, and G. Wang, "Progressive graph attention network for video question answering," in *Proceedings of the International Conference on Multimedia (MM)*, 2021.

[54] H. R. Tavakoli, R. Shetty, A. Borji, and J. Laaksonen, "Paying attention to descriptions generated by image captioning models," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[55] B. N. Patro, Anupriy, and V. P. Namboodiri, "Explanation vs attention: A two-player game to obtain attention for VQA," in *Association for the Advancement of Artificial Intelligence (AAAI)*, 2020.

[56] S. Yagcioglu, A. Erdem, E. Erdem, and N. Ikizler-Cinbis, "RecipeQA: A challenge dataset for multimodal comprehension of cooking recipes," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

[57] R. Zellers, X. Lu, J. Hessel, Y. Yu, J. S. Park, J. Cao, A. Farhadi, and Y. Choi, "MERLOT: Multimodal neural script knowledge models," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2021.

[58] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[59] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

[60] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2019.

[61] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2014.

[62] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Transactions on Information Theory (IEEE TIT)*, 1991.

[63] L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web," in *The Web Conference*, 1999.

[64] J. Chen, H. Hu, H. Wu, Y. Jiang, and C. Wang, "Learning the best pooling strategy for visual semantic embedding," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[65] H. Chen, G. Ding, X. Liu, Z. Lin, J. Liu, and J. Han, "IMRAM: Iterative matching with recurrent attention memory for cross-modal image-text retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[66] H. Wu, J. Mao, Y. Zhang, Y. Jiang, L. Li, W. Sun, and W.-Y. Ma, "Unified visual-semantic embeddings: Bridging vision and language with structured meaning representations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[67] Y. Yang, R. N. Lichtenwalter, and N. V. Chawla, "Evaluating link prediction methods," *Knowledge and Information Systems*, 2015.

[68] D. S. Pilco and A. R. Rivera, "Graph learning network: A structure learning algorithm," in *Proceedings of the International Conference on Machine Learning Workshop (ICMLW)*, 2019.

[69] A. Toet, "Computational versus psychophysical bottom-up image saliency: A comparative evaluation study," *IEEE Transactions on Pattern Analysis and Machine Intelligence (IEEE TPAMI)*, 2011.

[70] D. Mahajan, R. Girshick, V. Ramanathan, K. He, M. Paluri, Y. Li, A. Bharambe, and L. van der Maaten, "Exploring the limits of weakly supervised pretraining," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[71] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015.

[72] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[73] A. Shrivastava, A. Gupta, and R. Girshick, "Training region-based object detectors with online hard example mining," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[74] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J.-N. Hwang, K.-W. Chang, and J. Gao, "rounded language-image pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[75] Y. Lu, P. Lu, Z. Chen, W. Zhu, X. E. Wang, and W. Y. Wang, "Multimodal procedural planning via dual text-image prompting," *arXiv preprint arXiv:2305.01795*, 2023.

[76] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

[77] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Xianyu Chen received his B.E. and M.E. degrees from the School of Information Science and Technology and the School of Electronics and information technology, Sun Yat-sen University, Guangzhou, China, in 2015 and 2018, respectively. He is currently a Ph.D. student at the Department of Computer Science, University of Minnesota, USA. His research interests include computer vision, pattern recognition, and machine learning.

Jinhui Yang received BA degree in computer science and statistics from Carleton College in 2019. He is currently a Ph.D. student at the Department of Computer Science, University of Minnesota. His current research interests include computer vision, interpretable machine learning, and deep neural networks.

Shi Chen received the B.E. degree from the School of Computer Science, Wuhan University, Wuhan, China, in 2015. He received the M.S. and Ph.D. degree from Department of Computer Science, University of Minnesota, in 2017 and 2023, respectively. His research interests broadly include computer vision, vision and language, human vision, and machine learning.

Louis Wang is currently a Ph.D. student in the Computer Science program with a research focus in computer vision and machine learning, as well as a broader interest in transfer learning. He obtained His B.Sc. degree in Computer Science and Mathematics from the University of Minnesota in 2019.

Ming Jiang received the B.E. and M.E. degrees from Zhejiang University, Hangzhou, China and the Ph.D. degree in electrical and computer engineering from the National University of Singapore. He is currently a researcher at the Department of Computer Science and Engineering, University of Minnesota. His research interests include computer vision, cognitive vision, machine learning, psychophysics, neuroscience, and brain-machine interface.

Qi Zhao (Senior Member, IEEE) received Ph.D. degree in computer engineering from the University of California, Santa Cruz in 2009. She is currently an associate professor in the Department of Computer Science and Engineering at the University of Minnesota, Twin Cities. She was a postdoctoral researcher in the Computation & Neural Systems, and the Division of Biology at the California Institute of Technology from 2009 to 2011. She has published more than 100 journal and conference papers in computer vision, machine learning, and cognitive neuroscience venues, and edited a book with Springer, titled Computational and Cognitive Neuroscience of Vision, that provides a systematic and comprehensive overview of vision from various perspectives. She serves as an associate editor at the IEEE Transactions on Neural Networks and Learning Systems (TNNLS), IEEE Transactions on Multimedia (TMM), and IEEE Transactions on Cognitive and Developmental Systems (TCDS), as a program chair at IEEE Winter Conference on Applications of Computer Vision (WACV), and as an organizer and/or area chair at IEEE Conference on Computer Vision and Pattern Recognition (CVPR) and other major venues in computer vision and AI regularly. Her main research interests include computer vision, machine learning, cognitive neuroscience, and healthcare.