# An End-to-End Adaptable Prototypical Framework for Explainable Fine-Grained Visual Question Answering

**Anonymous ACL submission**

## Abstract

Visual Question Answering (VQA) systems, while advancing through vision transformers, remain largely black boxes in critical applications. Current prototype-based interpretability methods struggle with multimodal reasoning, rigid feature representations, and lack of fine-grained explanations. We present ProtoVQA, introducing adaptable prototypes for cross-modal tasks, spatially-constrained matching for geometric variations, and systematic evaluation of visual-linguistic alignment. Our model achieves competitive accuracy on Visual7W while providing comprehensive explainability through explicit visual evidence. Our code is available at https://anonymous.4open.science/r/ARR-Submission136.

## 1 Introduction

Visual Question Answering (VQA) is a key challenge in AI, requiring systems to understand and reason about both visual content and natural language queries (Zhu et al., 2016; Kafle and Kanan, 2017). Recent advances in vision transformers (Dosovitskiy et al., 2021; Touvron et al., 2021) have significantly improved performance by enhancing multimodal feature learning, leading to better accuracy on VQA benchmarks.

As VQA systems are applied in critical fields such as medical diagnosis (Wang et al., 2022; Donnelly et al., 2024; Yang et al., 2024) and autonomous driving (Ramos et al., 2017), model transparency is essential. Current state-of-the-art models operate as black boxes (Figure 1), making it difficult to interpret their reasoning or verify reliability. Traditional VQA interpretability approaches, primarily using attention visualization or post-hoc explanation methods, often fail to faithfully represent the model's decision-making process (Chen et al., 2019; Ma et al., 2023, 2024).

Prototype-based learning has emerged as a promising approach to improving interpretability

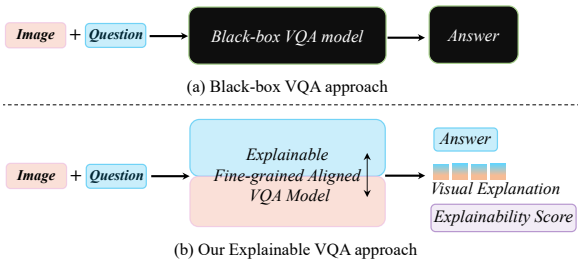



Figure 1: Comparison of different VQA approaches: (a) Traditional black-box approach provides only the answer without enough explanation. (b) Our approach provides comprehensive explainability through explicit visual explanation and quantifiable explainability score.

(Chen et al., 2019; Barnett et al., 2021; Donnelly et al., 2022; Ma et al., 2023). The latest work like ProtoViT (Ma et al., 2024) shows that Vision Transformers can enable flexible prototype learning while maintaining interpretability. However, existing prototype-based methods face challenges in multimodal reasoning and cross-modal interpretability, which are critical for VQA tasks. These challenges include: *(i)* Prototype-based approaches often focus on single modality (visual/textual) interpretability, struggling to bridge the visual-textual semantic gap; *(ii)* Rigid prototypical features fail to capture geometric variations and dynamic visual-question relationships; *(iii)* These methods lack the ability to provide fine-grained explanations at both the component level and system level, making it difficult to understand how individual parts contribute to the final decision. To address these issues, we propose ProtoVQA. Our contributions are:

- We introduce an end-to-end adaptable prototypical framework capable of seamlessly handling diverse visual-linguistic downstream tasks, including both visual question answering and grounding, through a shared prototype-based backbone with task-specific answer processing modules.

- We employ a spatially-constrained greedy matching strategy to model dynamic visual-question

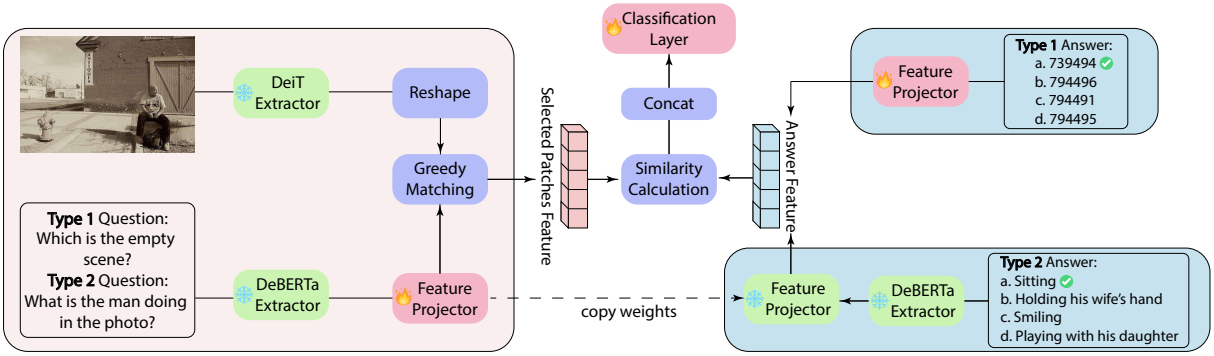



Figure 2: Overview of ProtoVQA. ProtoVQA extracts visual features through DeiT Extractor and encodes questions through DeBERTa Extractor. Image patch features undergo greedy matching with question-aware prototypes generated through a feature projector. For answering, the model processes either coordinate inputs (Type 1) through another projector, or textual answers (Type 2) through DeBERTa and a frozen feature projector sharing weights from the question branch. The matched patch features are concatenated with answer features for final classification.

relationships and geometric variations.

- Our model achieves comprehensive explainability through explicit visual evidence and systematic validation of visual-linguistic alignment.

## 2 ProtoVQA

We present ProtoVQA (Figure 2), a novel prototype-based approach to visual question answering that achieves interpretability through question-aware prototype learning and spatially-constrained patch matching. By explicitly mapping prototypes to discriminative image regions, ProtoVQA can provide transparent reasoning paths from questions to visual evidence.

### 2.1 Feature Extraction Module

The visual feature extraction leverages pre-trained DeiT (Touvron et al., 2021) as backbone to extract patch-level visual features. Let $I \in \mathbb{R}^{H\times W\times 3}$ denote the input image. The DeiT backbone processes $I$ to produce a feature map $F \in \mathbb{R}^{(N+1)\times D}$, where $N$ is the number of image patches and $D$ is the output feature dimension. To enhance local feature representation, we compute the difference between each patch feature and the global CLS token representation: $F_{local} = F[1:] - F[0]$, where $F[0]$ is the CLS token feature.

For textual input, ProtoVQA utilizes a pre-trained DeBERTa model (He et al., 2021). The question $Q$, represented as a token sequence $[q_1, q_2, ..., q_{l_q}]$, is encoded by DeBERTa, yielding embeddings $E_q \in \mathbb{R}^{l_q\times D_{\text{text}}}$, where $D_{\text{text}}$ is the DeBERTa hidden dimension. These embeddings are then projected into the shared visual-linguistic space $\mathbb{R}^D$ via a learnable feature projector $\mathcal{F}$. For answer processing, there are two pathways: For question answering tasks, answer candidates are encoded by DeBERTa and projected to $\mathbb{R}^D$ using the same feature projector $\mathcal{F}$ with frozen parameters whose weights are copied from the question encoding. This weight-sharing mechanism ensures consistent representation of question and answer candidates in the shared visual-linguistic space, while the frozen parameter design prevents potential overfitting. For visual grounding tasks, the coordinate inputs $P \in \mathbb{R}^4$ are directly projected to the same feature space through a separate feature projector.

### 2.2 Interpretable Prototypical Part Selection Module

This module constitutes the core novelty and interpretability mechanism of ProtoVQA. It introduces sub-patch prototypes and a greedy matching algorithm with spatial constraints to select salient image parts.

#### 2.2.1 Sub-patch Prototypes

From the projected question embeddings in the shared visual-linguistic space, we construct a structured set of prototypes by reshaping the first $m \times k$ tokens' representations into a three-dimensional tensor:

$$P = \text{Reshape}(\mathcal{F}(E_q[:m\times k])) \in \mathbb{R}^{m\times k\times D}. \quad (1)$$

This organization generates $m$ prototypes, each containing $k$ sub-patches that maintain the same feature dimensionality as the visual features $F$. The sub-patches within each prototype are designed to collectively capture different aspects of visual semantics, from object attributes to structural information. Through a learnable weighting mechanism, the model can adaptively adjust the importance of

each sub-patch based on the question type, enabling context-aware patch selection during the matching process.

#### 2.2.2 Greedy Matching with Spatial Constraints

The core matching mechanism employs a spatially-constrained greedy algorithm (Ma et al., 2024) to establish correspondences between sub-patch prototypes and image regions. For each prototype $P_i \in \mathbb{R}^{k \times D}$ from our prototype set $P$, the algorithm iteratively constructs a spatially coherent set of matched image patches through $k$ iterations.

At each iteration $t$, we first compute the similarity matrix $S^t \in \mathbb{R}^{N \times k}$ between local image features $F_{local}$ and prototype sub-patches $P_i$:

$$S^t_{n,j} = \frac{F_{local,n} \cdot P_{i,j}}{\|F_{local,n}\| \|P_{i,j}\|}, \tag{2}$$

where $n \in \{1, ..., N\}$ indexes image patches, and $j \in \{1, ..., k\}$ indexes sub-patches.

The algorithm then identifies the optimal patch-subpatch pair $(n^*, j^*)$ that maximizes the similarity score:

$$(n^*, j^*) = \underset{n,j}{\text{argmax}} \; S^t_{n,j} \cdot M^t_n \cdot A^t_n, \tag{3}$$

where $M^t \in \{0, 1\}^N$ is a binary mask indicating available patches at iteration $t$ (1 for available patches, 0 for unavailable), and $A^t \in \{0, 1\}^N$ is an adjacency mask enforcing spatial continuity with previously selected patches. After each selection, the masks are updated: $M^{t+1}$ marks the selected patch as unavailable by setting $M^{t+1}_{n^*} = 0$ to prevent repeated selection in subsequent iterations, and $A^{t+1}$ is updated to mark as valid only those patches within a spatial constraint radius $r$ from position $n^*$, ensuring spatial coherence in the matching process.

The final matching score for prototype $P_i$ is computed as a weighted combination of individual sub-patch similarities:

$$\text{score}(P_i) = \sum_{t=1}^{k} w_t \cdot S^t_{n^*_t, j^*_t}, \tag{4}$$

where $w_t$ are learnable slot weights that modulate the importance of each sub-patch match, and $(n^*_t, j^*_t)$ denotes the optimal pair selected at iteration $t$. This spatially-aware matching strategy ensures the selected patches form coherent visual regions while maintaining semantic relevance to the prototype.

### 2.3 Answer Processing

`ProtoVQA` supports two types of answer processing: **Type 1 (Visual Grounding)** for tasks requiring coordinate-based answers, where the input coordinates $P \in \mathbb{R}^4$ are projected directly into the feature space through a dedicated projector; and **Type 2 (Descriptive QA)** for tasks requiring textual answers, where candidates are encoded by DeBERTa and projected using a frozen feature projector that shares weights with the question branch, ensuring consistent representation while preventing overfitting. In both cases, the matched patch features are concatenated with the processed answer features and fed through a classification layer for final prediction.

### 2.4 Visual-Linguistic Alignment Evaluation

To systematically evaluate the alignment between visual and linguistic components, we propose the **Visual-Linguistic Alignment Score** (VLAS):

$$\text{VLAS} = \frac{\sum_{i=1}^{N} \mathcal{I}(M_i \cap G_i > \theta)}{N_{QA}}, \tag{5}$$

where $M_i$ represents the model-attended region for the $i$-th QA pair, $G_i$ denotes the corresponding ground truth region, $\mathcal{I}(\cdot)$ is an indicator function that returns 1 if the condition is satisfied and 0 otherwise, $\theta$ is the overlap threshold, and $N_{QA}$ is the total number of QA pairs. A correct match occurs when the model-attended regions correspond to the ground truth visual elements. This formulation directly measures the effectiveness of our prototype-based matching mechanism, providing a quantitative measure of the model's ability to ground its reasoning in appropriate visual evidence.

## 3 Experiments

### 3.1 Setup

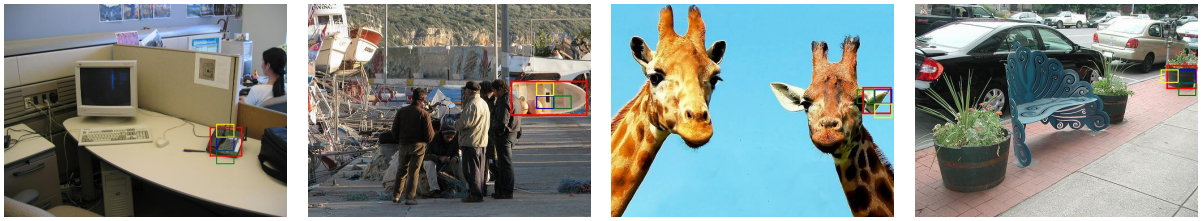
**Dataset** *(i)* Visual7W (Zhu et al., 2016): 327,939 QA pairs on 47,300 COCO images, with 1,311,756 human-generated multiple-choices and 561,459 object groundings from 36,579 categories. We follow standard training/test splits for our experiments.

**Baselines** Baseline details are in Appendix A.

**Configuration** The model was trained on an NVIDIA A800 GPU (80GB) for 200 epochs using Adam optimizer (lr=$1 \times 10^{-4}$, batch size=64). The vision transformer processed $224 \times 224$ images with $16 \times 16$ patches. The prototype learning used

| Method | Image Encoder | Text Encoder | Accuracy (%)↑ |
|---|---|---|---|
| SUPER (Han et al., 2023) | FasterRCNN | GRU | 64.07 |
| QOI_Attention (Gao et al., 2018) | FasterRCNN | GRU | 65.90 |
| SDF of VLT (Ding et al., 2022) | ViT-patch16 | BERT | 65.93 |
| STL (Wang et al., 2018) | ResNet200 | n-gram | 68.20 |
| CFR (Nguyen et al., 2022) | FasterRCNN | GRU | 71.90 |
| BriVL (Fei et al., 2022) | Custom image patch+CNN | RoBERTa | 72.06 |
| CTI (Do et al., 2019) | FasterRCNN | LSTM/GRU | 72.30 |
| Bi-CMA (Upadhyay and Tripathy, 2025) | ViT-patch16 | BERT | 70.53 |
| Bi-CMA (Upadhyay and Tripathy, 2025) | ViT-patch16 (finetune) | BERT | 73.07 |
| **ProtoVQA (Ours)** | ViT-patch16 | DeBERTa | 70.23 |

Table 1: Comparison with state-of-the-art VQA methods on Visual7W (Zhu et al., 2016) test set. Baselines include various architectures from CNN-RNN approaches to Transformer-based models.



(a) Which item can be used for communication?

(b) Which is framing a white sideways boat?

(c) Which ear is the left ear of the right giraffe?

(d) Which flower tub, with red flowers in it, is beside a parking meter?

Figure 3: Visualization of explanation results on Visual7W (Zhu et al., 2016) test set. The **red** bounding box indicates the ground truth answer box provided by the dataset. The **blue**, **green** and **yellow** bounding boxes show the **top-3 best-matched patches** projected back to the original image space. More visualization results on diverse visual question answering scenarios can be found in Appendix Section B.

$m = 10$ prototypes per class (each with $k = 3$ sub-patches) and a spatial constraint radius of $r = 3$. Other hyperparameters remained default.

### 3.2 Comparison with Baselines

As shown in Table 1, among the methods using ViT-patch16 as visual backbone, ProtoVQA (70.23%) achieves performance comparable to Bi-CMA (70.53% without fine-tuning, 73.07% with fine-tuning) and outperforms the SDF of VLT (65.93%). Notably, while maintaining competitive accuracy, our method enables explicit reasoning visualization (Figure 3), which is not available in other ViT-based approaches.

Figure 3 demonstrates ProtoVQA's comprehensive interpretability across diverse scenarios, with model-attended regions (**blue**, **green** and **yellow** boxes) consistently aligning with ground truth annotations (**red** boxes) for tasks ranging from object identification to spatial relationship understanding.

As shown in Table 2, ProtoVQA significantly outperforms baseline methods on VLAS (0.4103 vs 0.2466 on VLAS@1, 0.2466 vs 0.1123 on VLAS@3), representing a 66.4% and 119.6% improvement over Bi-CMA respectively, demonstrating superior visual-linguistic alignment capability.

| Method | VLAS@1↑ | VLAS@3↑ |
|---|---|---|
| SDF of VLT | 0.2013 | 0.0847 |
| Bi-CMA | 0.2466 | 0.1123 |
| **ProtoVQA (Ours)** | 0.4103 | 0.2466 |

Table 2: Visual-Linguistic Alignment Score (VLAS) comparison on Visual7W (Zhu et al., 2016) test set.

## 4 Conclusion

We present ProtoVQA, a novel framework for visual question answering that addresses the need for model transparency and cross-modal reasoning. ProtoVQA achieves comprehensive explainability by (i) introducing adaptable prototypes capable of seamlessly handling diverse visual-linguistic downstream tasks through a shared prototype-based backbone; (ii) employing a spatially-constrained greedy matching strategy to model dynamic visual-question relationships and geometric variations; and (iii) providing explicit visual evidence and systematic validation of visual-linguistic alignment. Our work provides a fundamental step towards VQA systems that achieve strong performance while maintaining comprehensive explainability.

## Limitations

While ProtoVQA demonstrates strong performance and interpretability in general visual question answering scenarios, current applications primarily focus on common object understanding and general scene comprehension. Future extensions could enhance its capabilities in specialized domains such as medical imaging question answering, where interpretable decision reasoning is particularly crucial for clinical decision support and patient care.

## Ethical Considerations

We examined the study describing the publicly available datasets used in this research and identified no ethical issues regarding the datasets.

## References


Alina Jade Barnett, Fides Regina Schwartz, Chaofan Tao, Chaofan Chen, Yinhao Ren, Joseph Y Lo, and Cynthia Rudin. 2021. A case-based interpretable deep learning model for classification of mass lesions in digital mammography. *Nature Machine Intelligence*.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. 2019. This Looks Like That: Deep Learning for Interpretable Image Recognition. In *Advances in Neural Information Processing Systems*.

Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. 2022. Vlt: Vision-language transformer and query generation for referring segmentation. *Transactions on Pattern Analysis and Machine Intelligence*.

Tuong Do, Thanh-Toan Do, Huy Tran, Erman Tjiputra, and Quang D Tran. 2019. Compact trilinear interaction for visual question answering. In *International Conference on Computer Vision*.

Jon Donnelly, Alina Jade Barnett, and Chaofan Chen. 2022. Deformable ProtoPNet: An Interpretable Image Classifier Using Deformable Prototypes. In *Conference on Computer Vision and Pattern Recognition*.

Jon Donnelly, Luke Moffett, Alina Jade Barnett, Hari Trivedi, Fides Schwartz, Joseph Lo, and Cynthia Rudin. 2024. AsymMirai: Interpretable Mammography-based Deep Learning Model for 1–5-year Breast Cancer Risk Prediction. *Radiology*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Nanyi Fei, Zhiwu Lu, Yizhao Gao, Guoxing Yang, Yuqi Huo, Jingyuan Wen, Haoyu Lu, Ruihua Song, Xin Gao, Tao Xiang, et al. 2022. Towards artificial general intelligence via a multimodal foundation model. *Nature Communications*.

Lianli Gao, Pengpeng Zeng, Jingkuan Song, Xianglong Liu, and Heng Tao Shen. 2018. Examine before you answer: Multi-task learning with adaptive-attentions for multiple-choice vqa. In *International Conference on Multimedia*.

Yudong Han, Jianhua Yin, Jianlong Wu, Yinwei Wei, and Liqiang Nie. 2023. Semantic-aware modular capsule routing for visual question answering. *Transactions on Image Processing*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Kushal Kafle and Christopher Kanan. 2017. An analysis of visual question answering algorithms. In *International Conference on Computer Vision*.

Chiyu Ma, Jon Donnelly, Wenjun Liu, Soroush Vosoughi, Cynthia Rudin, and Chaofan Chen. 2024. Interpretable image classification with adaptive prototype-based vision transformers. In *Advances in Neural Information Processing Systems*.

Chiyu Ma, Brandon Zhao, Chaofan Chen, and Cynthia Rudin. 2023. This Looks Like Those: Illuminating Prototypical Concepts Using Multiple Visualizations. In *Advances in Neural Information Processing Systems*.

Binh X Nguyen, Tuong Do, Huy Tran, Erman Tjiputra, Quang D Tran, and Anh Nguyen. 2022. Coarse-to-fine reasoning for visual question answering. In *Conference on Computer Vision and Pattern Recognition*.

Sebastian Ramos, Stefan Gehrig, Peter Pinggera, Uwe Franke, and Carsten Rother. 2017. Detecting Unexpected Obstacles for Self-Driving Cars: Fusing Deep Learning and Geometric Modeling. In *IEEE Intelligent Vehicles Symposium*.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. 2021. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*.

Sushmita Upadhyay and Sanjaya Shankar Tripathy. 2025. Bidirectional cascaded multimodal attention for multiple choice visual question answering. *Machine Vision and Applications*.

Chong Wang, Yuanhong Chen, Yuyuan Liu, Yu Tian, Fengbei Liu, Davis J McCarthy, Michael Elliott, Helen Frazer, and Gustavo Carneiro. 2022. Knowledge distillation to ensemble global and interpretable prototype-based mammogram classification models.

In *International Conference on Medical Image Computing and Computer-Assisted Intervention*.

Zhe Wang, Xiaoyi Liu, Limin Wang, Yu Qiao, Xiaohui Xie, and Charless Fowlkes. 2018. Structured triplet learning with pos-tag guided attention for visual question answering. In *Winter Conference on Applications of Computer Vision*.

Haoming Yang, Pramod KC, Panyu Chen, Hong Lei, Simon Sponberg, Vahid Tarokh, and Jeffrey Riffell. 2024. Neuron synchronization analyzed through spatial-temporal attention. *bioRxiv*.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Conference on Computer Vision and Pattern Recognition*.

## A Baselines

- **SUPER** (Han et al., 2023): Introduces a semantic-aware modular capsule routing framework for Visual Question Answering (VQA) to enhance adaptability to semantically complex inputs. It features five specialized modules and dynamic routers that refine vision-semantic representations, offering a novel approach to architecture learning and representation calibration for VQA tasks.

- **QOI_Attention** (Gao et al., 2018): Proposes a Multi-task Learning with Adaptive-attention (MTA) model for multiple-choice (MC) VQA. It mimics human reasoning by integrating answer options and adapting attention to visual features, achieving remarkable performance on MC VQA benchmarks.

- **SDF of VLT** (Ding et al., 2022): Presents a Vision-Language Transformer (VLT) framework for referring segmentation, introducing a Query Generation Module to dynamically produce input-specific queries. It improves handling diverse language expressions with a Query Balance Module and masked contrastive learning, setting new benchmarks on five datasets.

- **STL** (Wang et al., 2018): Proposes a VQA model focused on the multiple-choice task, incorporating part-of-speech (POS) tag-guided attention, convolutional n-grams, and triplet attention interactions between the image, question, and candidate answer. The model also employs structured learning for triplets based on image-question pairs.

- **CFR** (Nguyen et al., 2022): Introduces a reasoning framework for Visual Question Answering (VQA) that bridges the semantic gap between image and question by jointly learning features and predicates in a coarse-to-fine manner. The model achieves superior VQA accuracy and provides an explainable decision-making process.

- **BriVL** (Fei et al., 2022): Develops a foundation model pre-trained on multimodal data for artificial general intelligence (AGI), focusing on self-supervised learning with weak semantic correlation data. The model demonstrates strong imagination ability, achieving promising results across various downstream tasks including VQA.

- **Bi-CMA** (Upadhyay and Tripathy, 2025): Proposes a Bidirectional Cascaded Multimodal Attention network for VQA, utilizing bidirectional attention and sparsity to enhance feature integration between image and text. The model performs competitively on multiple-choice VQA tasks, providing insightful attention maps that reveal the model's decision-making focus.

## B Additional Visualization Results

In addition to results shown in Figure 3 in Section 3, we provide 10 representative samples from Visual7W test set. These examples demonstrate ProtoVQA's ability to handle diverse visual question answering scenarios, including human/animal anatomy (Figures 4, 5), object identification (Figures 6, 7, 8), object interactions (Figures 9, 10, 11), and spatial relationships (Figures 12, 13). In each case, the model provides clear visual explanations by highlighting relevant patches that directly correspond to the questions being asked.

Figure 4: Question: Which is the players arms?

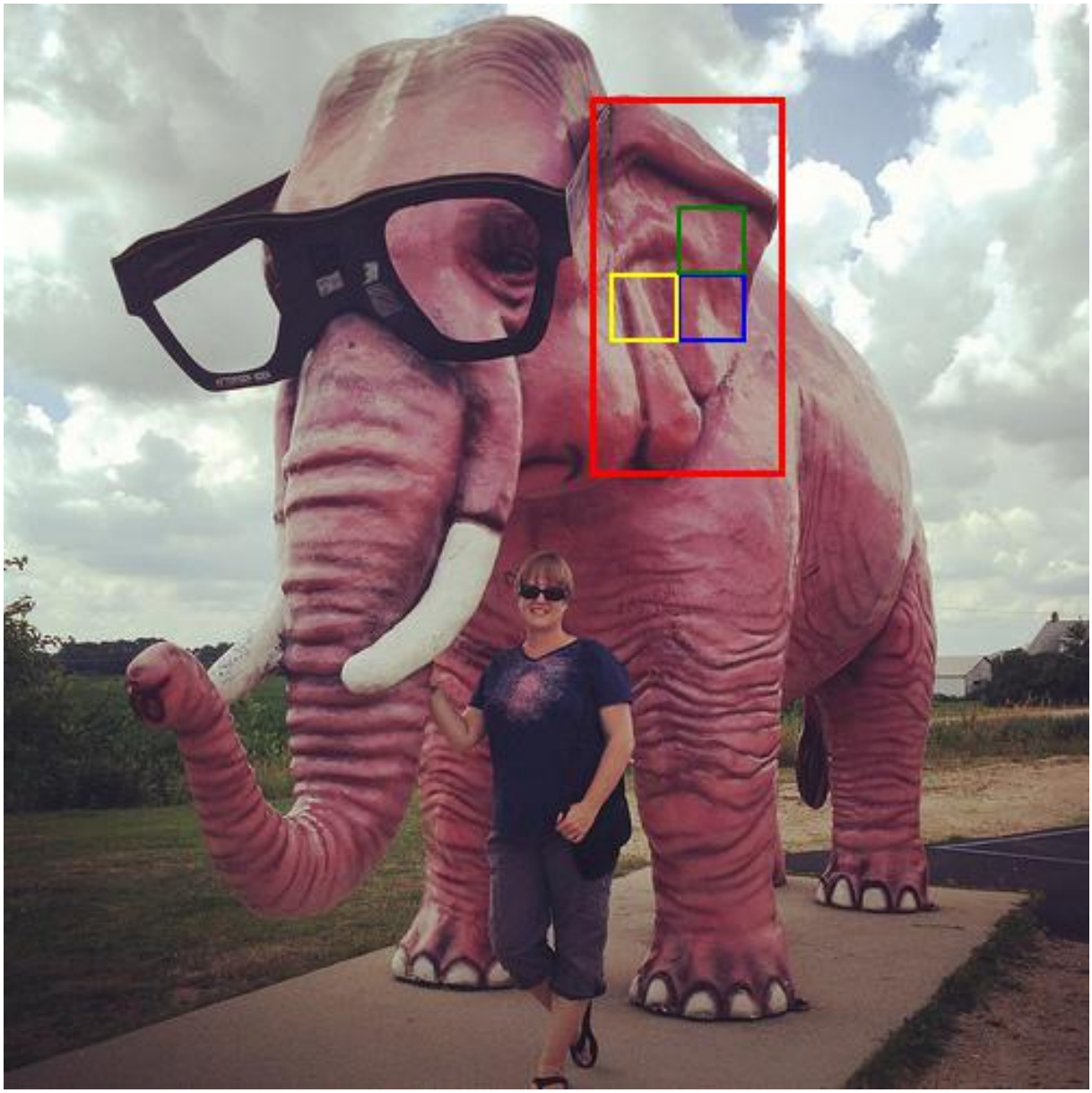


Figure 5: Question: Which part helps the elephant hear?

Figure 6: Question: Which plant is hanging in the room?

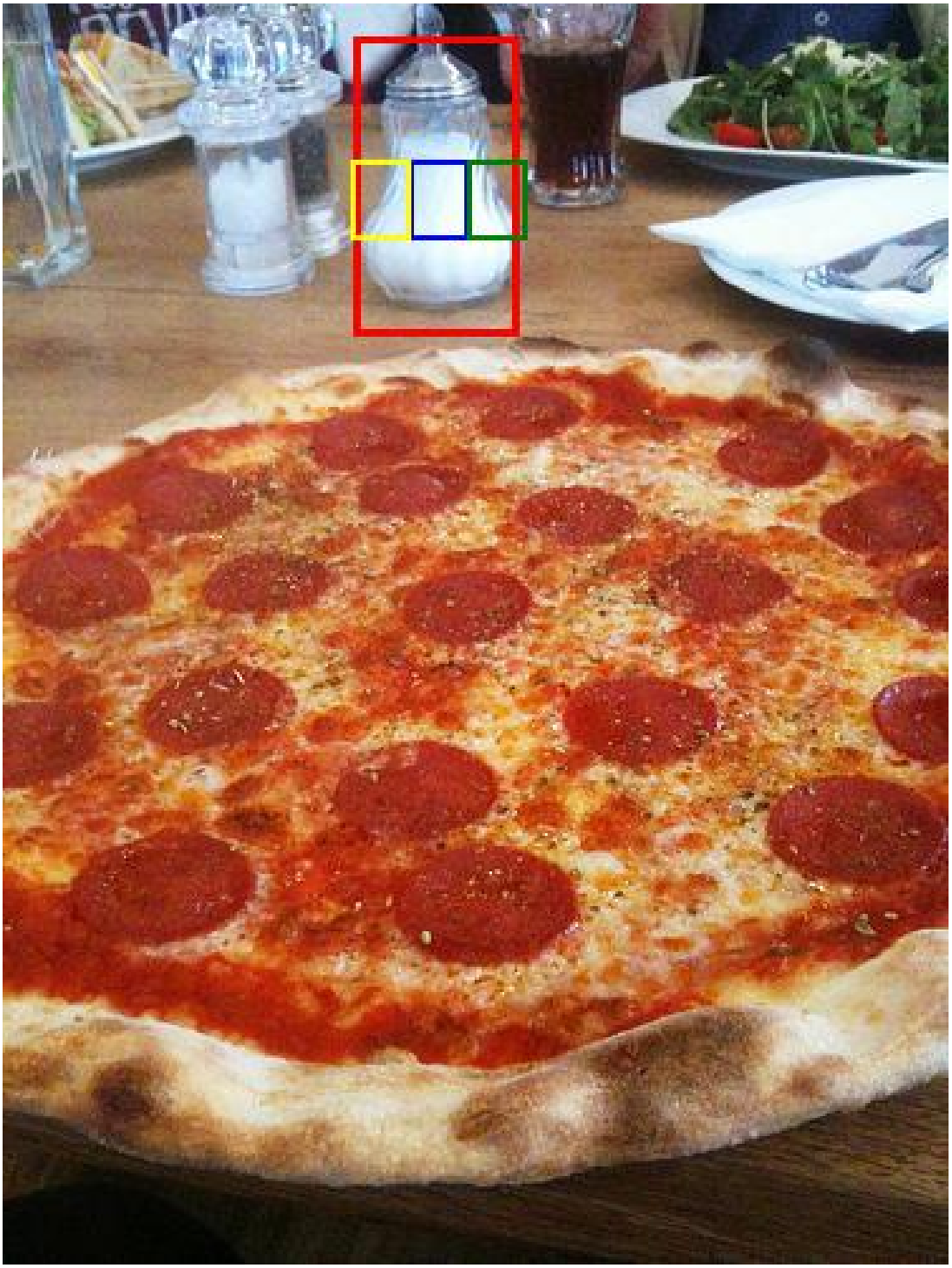


Figure 7: Question: Which is the glass containing?

Figure 8: Question: Which object is a large beige cylinder next to the dirt?


Figure 9: Question: Which object is she wearing on her face?

Figure 10: Question: Which object is being flown?

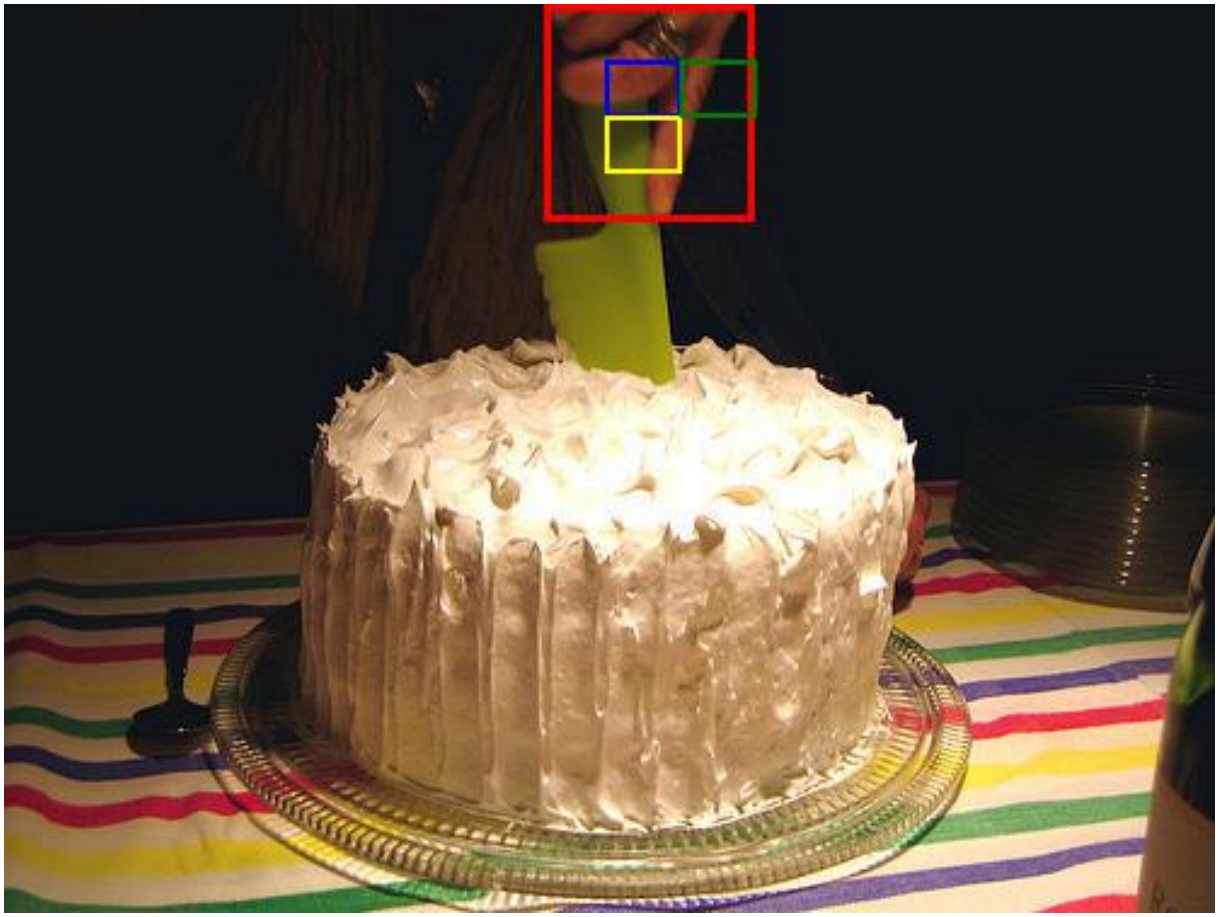


Figure 11: Question: Which hand is holding a knife?

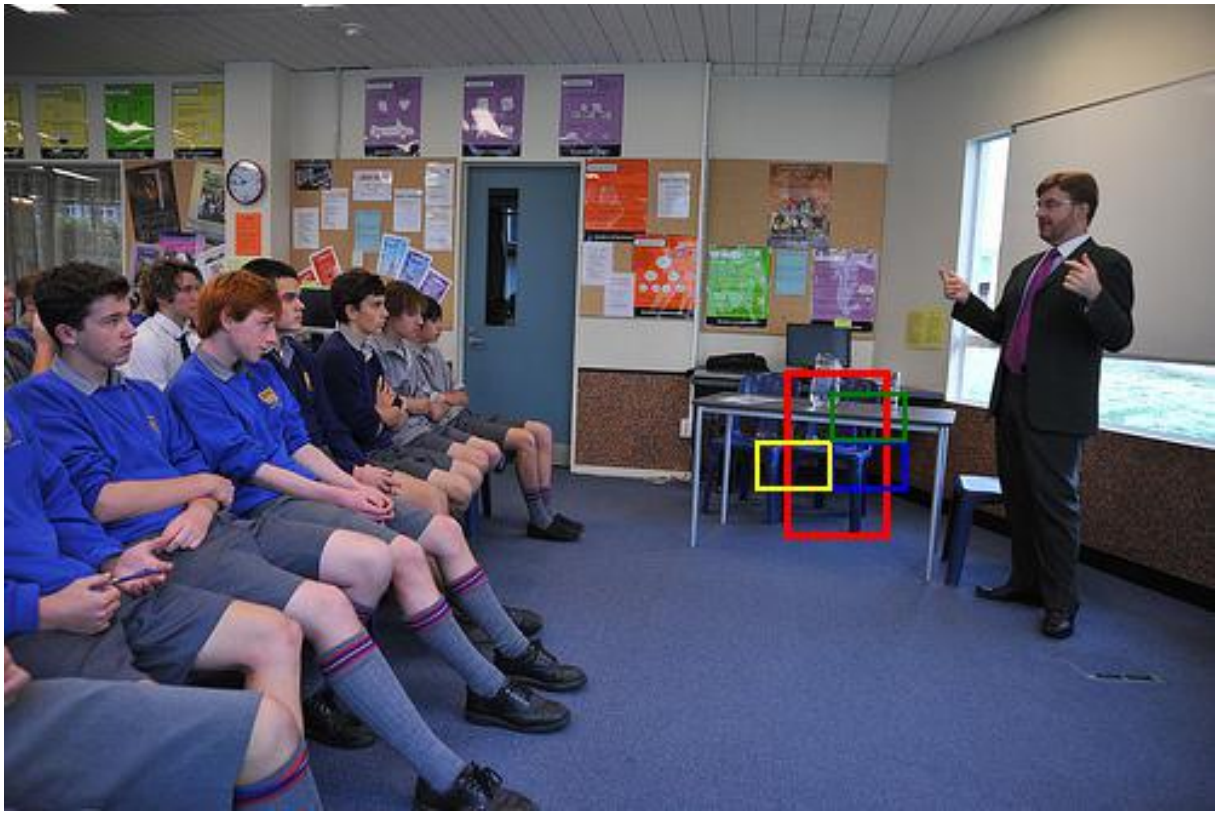


Figure 12: Question: Which blue chair behind the table?

Figure 13: Question: Which hose is sticking out of the wall?