Alignment before Awareness: Towards Visual Question Localized-Answering in Robotic Surgery via Optimal Transport and Answer Semantics

Zhihong Zhu[1], Yunyan Zhang[2], Xuxin Cheng[1], Zhiqi Huang[1], Derong Xu[3,4], Xian Wu[2]*, Yefeng Zheng[2]
[1]Peking University, [2]Jarvis Research Center, Tencent YouTu Lab,
[3]University of Science and Technology of China, [4]City University of Hong Kong
zhihongzhu@stu.pku.edu.cn, kevinxwu@tencent.com

Abstract

The visual question localized-answering (VQLA) system has garnered increasing attention due to its potential as a knowledgeable assistant in surgical education. Apart from providing text-based answers, VQLA can also pinpoint the specific region of interest for better surgical scene understanding. Although recent Transformer-based models for VQLA have obtained promising results, they (1) conduct vanilla text-to-image cross attention, leading to unidirectional and coarse-grained alignment; (2) ignore exploiting the semantics of answers to further boost performance. In this paper, we propose a novel model termed OTAS, which first introduces optimal transport to achieve bidirectional and fine-grained alignment between images and questions, enabling more precise localization. Besides, OTAS incorporates a set of learnable candidate answer embeddings to query the probability of each answer class for a given image-question pair. Through Transformer attention, the candidate answer embeddings interact with the fused features of the image-question pair to make the answer decision. Extensive experiments on two widely-used benchmark datasets demonstrate the superiority of our model over state-of-the-art methods.

Keywords: Visual Question Localized-Answering, Optimal Transport, Answer Semantics

1. Introduction

Recorded surgical videos are a useful tool for medical students to learn the procedure (Sharma et al., 2021). However, students often have various questions regarding the surgical instruments, human tissues, and workflows shown in the video. As such, an automatic deep learning-based surgical visual question-answering (VQA) system (Seenivasan et al., 2022a) has been developed as a vital training and educational tool for junior surgeons, medical students, and patients (Hsieh and Lin, 2017; Lin et al., 2006). However, while such VQA systems can provide answers to learners (Li et al., 2019), they lack the ability to relate these answers to their localization at an instance level (Bai et al., 2023b). Surgical scenarios with various similar instruments and actions may further confuse the learners, whereas answers with localization can further assist learners in dealing with confusion (Seenivasan et al., 2022b). To this end, a surgical visual question localized-answering (VQLA) system has been proposed to improve surgical training and enhance scene understanding effectively. A typical example of the VQLA system is shown in Figure 1.

With the recent advent of the attention mechanism, the VQLA architecture's capability is greatly improved. Therein, Bai et al. (2023a) developed a detection-free Transformer-based VQLA model, enabling end-to-end real-time applications. Based on this, Bai et al. (2023c) introduced a gated vision-language embedding mechanism for effective fusion of heterogeneous features, obtaining state-of-the-art results. Despite the promising progress that existing VQLA models have achieved, we discover that they still suffer from two main issues:

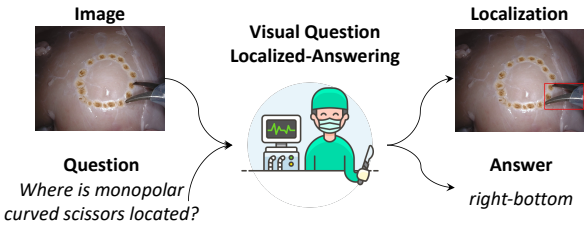

Figure 1: A typical example of the visual question localized-answer (VQLA) system, which takes an image and a question as input, and outputs both bounding box predictions and classification results.

(1) *Existing VQLA models mostly conduct vanilla one-way attention in robotic surgery scenarios.* To be specific, they focus solely on text-to-image attention while neglecting image-to-text attention. Intuitively, text and image information benefit from each other. Therefore, simply conducting unidirectional attention in VQLA is suboptimal. Besides, they lack the fine-grained correspondence between instruments/actions and questions, which is crucial for the localization task. Further efforts are needed to refine this coarse-grained interaction.

(2) *In existing methods, where answers are con-*

* Corresponding author

LREC-COLING 2024, pages 711–721

strained to specific categories, the labels are transformed into one-hot vectors. As a result, these vectors are orthogonal in the embedding space, neglecting the rich semantic information among answers. One main premise of our work is that answers are themselves words that may appear in various contexts and are thus semantically related to other words that appear in questions, and this relatedness can be leveraged. For example, in Figure 1, the word "*located*" in the question will have stronger semantic relevance to the orientation information "*right-bottom*" in the answer set compared to other answers (*e.g.*, "*tool manipulation*").

To solve the aforementioned issues, we propose a novel model via **O**ptimal **T**ransport and **A**nswer **S**emantics (OTAS) for VQLA. To solve the first issue, we introduce optimal transport (Kantorovich, 2006) to model the bidirectional and fine-grained alignment between images and questions. In detail, the alignment between images and questions is regarded as a transportation plan, where the distance between images and questions is measured by the transportation cost. By minimizing the transportation cost, the model could achieve bidirectional and fine-grained alignment between images and questions. The subsequent visualization studies further confirm our model's superiority in VQLA, especially in the localization task. To solve the second issue, we introduce a set of learnable candidate answer embeddings and let the image-question feature interact with the candidate answer embeddings by sending them through a Transformer decoder (Vaswani et al., 2017). In the decoder, the candidate answer embeddings work as a query to calculate their relationships with the fused image-question features to choose the final answer from a set of candidates. In this manner, our classification considers the interaction of answer semantics and the fused image-question features, which is different from existing VQLA methods. Experiment results show that our OTAS significantly outperforms previous models, and comprehensive analysis further verifies the advantages of our model.

The contributions of our work are three-fold: (1) We propose a novel model termed OTAS, which leverages optimal transport to achieve bidirectional and fine-grained alignment between questions and images. To the best of our knowledge, we make the first attempt to employ optimal transport in VQLA. (2) We propose to incorporate answer semantics to leverage its correlation with questions, which is achieved by a designed mechanism to learn and make use of candidate answer embedding through a Transformer decoder. (3) Extensive experiments show that our model achieves new state-of-the-art (SOTA) performance, which demonstrates the potential of AI-based VQLA systems in surgical training and surgical scene understanding.

2. Related Work

Surgical Visual Question Localized-Answering. Due to the overwhelming burden of academic and clinical work, expert surgeons find it challenging to address the myriad questions posed by learners regarding surgical procedures (Sharma et al., 2021; Seenivasan et al., 2022a). As a partial solution, recorded surgical videos are shared with students, enabling them to learn through observation. However, this approach still falls short in addressing specific queries students may have. Recently, MedFuseNet (Sharma et al., 2021) was introduced to tackle medical visual question answering (VQA), expanding the realm of possibilities for developing reliable VQA models capable of assisting medical experts in addressing queries from learners. SurgicalVQA (Seenivasan et al., 2022a) was proposed to answer questionnaires on surgical tools, tool-tissue interactions and surgical phase based on the visual input. More recently, Bai et al. (2023a,c) introduced the Visual Question Localized-Answering (VQLA) model in the surgical domain, which can predict localized-answer based on a given input question and surgical scene, showcasing the application potential of AI-driven VQLA systems in surgical training and surgical scene understanding.

Optimal Transport. Optimal transport (Kantorovich, 2006) is a classic mathematical problem and was initially introduced to solve the problem of minimizing the cost when moving multiple items simultaneously. To reduce the computational complexity, Kusner et al. (2015) proposed a relaxed form of optimal transport. With the development of machine learning (Chen et al., 2024; Huang et al., 2024), optimal transport is widely applied to compare different distributions, such as structural matching (Chen et al., 2018; Marchisio et al., 2022; Zhu et al., 2023a), generative models (Liutkus et al., 2019; Rout et al., 2021), image matching (Zhang et al., 2020a; Qian et al., 2023), and cross-modal alignment (Chen et al., 2020; Zhou et al., 2023). In this paper, we use the optimal transport to achieve the bidirectional and fine-grained alignment between questions and images.

Label Semantics. Label semantics has been leveraged in many settings and tasks to improve performance and robustness (Zhu et al., 2024, 2023c,b; Xu et al., 2024). Mullenbach et al. (2018) proposed label-wise attention networks for datasets with very large structured label spaces. Rios and Kavuluru (2018) extended the attention mechanism for zero-shot settings. Chalkidis et al. (2020) used BERT to embed the labels. In this paper, we introduce a set of answer embeddings to capture the rich semantic information between answers and

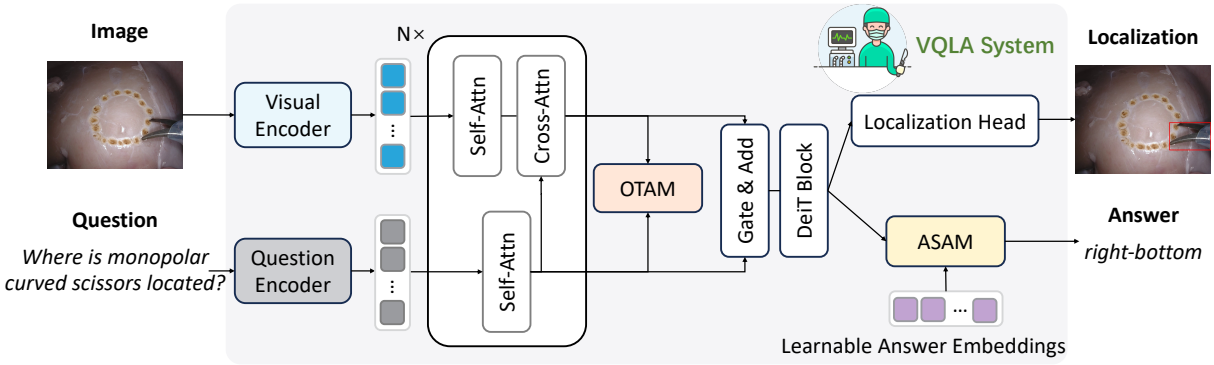

Figure 2: The architecture of proposed OTAS. "Attn" represents "Attention", "OTAM" and "ASAM" represent "Optimal Transport-based Alignment Module" and "Answer Semantics-aware Module", respectively.

questions, with a flatter and smaller label space.

3. Methodology

This section details our proposed OTAS, whose architecture is shown in Figure 2. OTAS consists of a visual-question encoder (§3.1), a coarse-grained alignment module (§3.2), an optimal transport-based alignment module (OTAM) (§3.3), a standard DeiT module (§3.4), and prediction heads (§3.5) including answer semantics-aware module (ASAM) and localization head. Finally, we introduce the training objective (§3.6) of our proposed OTAS.

3.1. Visual-Question Encoder

Following Bai et al. (2023a,c), to ensure a fair comparison, we adopt ResNet18 (He et al., 2016) as our visual feature extractor to obtain visual embeddings. The question embeddings are obtained using a specialized pre-trained tokenizer (Seenivasan et al., 2022a). Note that both visual and question dimensions are reduced to $\mathbb{R}^{n\times d}$ for subsequent interaction, where n denotes the length of tokens (*i.e.*, sub-words for text or patches for an image), and d denotes the dimension of hidden states.

3.2. Coarse-grained Alignment Module

We keep the vanilla N-layer one-way attention module following Bai et al. (2023a) to facilitate coarse-grained alignment between images and questions. Concretely, each co-attention layer consists of self-attention for question embeddings, and self-attention followed by cross-attention for image embeddings. The self-attention comprises a multi-head attention layer, a feed-forward layer, and ReLU activation. Given image/question embeddings from the previous layer, the self-attention leverages it to generate *query* $\mathbf{q} \in \mathbb{R}^{d_q}$, *key* $\mathbf{K} \in \mathbb{R}^{d_K}$ and *value* $\mathbf{V} \in \mathbb{R}^{d_V}$ matrices. Formally, the attention of each head $\mathbf{h}_i$ is calculated as:

$$\mathbf{h}_i = A(\mathbf{W}_i^{(q)}\mathbf{q}, \mathbf{W}_i^{(K)}\mathbf{K}, \mathbf{W}_i^{(V)}\mathbf{V}), \quad (1)$$

where $\mathbf{W}_i^{(q)} \in \mathbb{R}^{p_q\times d_q}$, $\mathbf{W}_i^{(K)} \in \mathbb{R}^{p_K\times d_K}$ and $\mathbf{W}_i^{(V)} \in \mathbb{R}^{p_V\times d_V}$ are learnable parameters, and $A(\cdot)$ represents single-head attention aggregation. A linear conversion is then applied for the attention aggregation from multiple heads: $\mathbf{h} = MA(\mathbf{W}_o[\mathbf{h_1}||\cdots||\mathbf{h}_M])$, where $\mathbf{W}_o \in \mathbb{R}^{p_o\times Mp_V}$ is a learnable parameter, $MA(\cdot)$ represents multi-head attention aggregation, M is the number of heads in the current layer, and $||$ represents concatenation.

The cross-attention module also contains the above components, but its input is from both two modalities, in which $\mathbf{q}$ is from visual embeddings and $\mathbf{K}$, $\mathbf{V}$ are from question embeddings:

$$\mathbf{h}_i = A(\mathbf{W}_i^{(q)}\mathbf{q}_v, \mathbf{W}_i^{(K)}\mathbf{K}_q, \mathbf{W}_i^{(V)}\mathbf{V}_q). \quad (2)$$

By this, visual embeddings are guided by question embeddings, thus enabling text-to-image attention.

3.3. Optimal Transport-based Alignment Module

Although §3.1 establishes an initial alignment between text and visual embeddings, it is unidirectional and coarse-grained. In this work, we present an innovative perspective to apply optimal transport (Kantorovich, 2006) to achieve bidirectional and fine-grained alignment. Next, we detail the Optimal Transport-based Alignment Module (OTAM).

Optimal transport is a classic mathematical problem, which considers both the position and weight of each element in a distribution, thereby capturing subtle changes and local structures between distributions (Peyré et al., 2019). Given an initial state $\alpha = \{\alpha_1, \cdots, \alpha_p\}$ before transportation, a final state $\beta = \{\beta_1, \cdots, \beta_q\}$ after transportation, and the unit cost function $C(\alpha_i, \beta_j)$ representing the unit transport cost from i-th position in α to the j-th position in β, the objective of optimal transport is

to develop a transport plan $\mathbf{T}$ to minimize the total transport cost $\mathcal{D}(\alpha, \beta)$, where each element $\mathbf{T}_{i,j}$ denotes the amount of mass transported from α_i to β_j. The total cost $\mathcal{D}(\alpha, \beta)$ is calculated as follows:

$$\begin{aligned} \mathcal{D}(\alpha, \beta) &= \min_{\mathbf{T} \geq 0} \sum_{i=1}^{p} \sum_{j=1}^{q} \mathbf{T}_{i,j} \cdot C(\alpha_i, \beta_j), \\ s.t. \sum_{j=1}^{q} \mathbf{T}_{i,j} &= m_i, \forall i \in \{1, \ldots, p\}, \\ \sum_{i=1}^{p} \mathbf{T}_{i,j} &= \hat{m}_j, \forall j \in \{1, \ldots, q\}, \end{aligned} \tag{3}$$

where each point $\alpha_i \in \mathbb{R}^d$ (resp. $\beta_j \in \mathbb{R}^d$) has a weight $m_i \in [0, \infty)$ (resp. $\hat{m}_j \in [0, \infty)$).

For the visual and question embeddings $\mathbf{F}_v$ and $\mathbf{F}_q$, we further employ optimal transport to measure the distance between them. The corresponding transport cost $\mathcal{D}(\mathbf{F}_v, \mathbf{F}_q)$ is calculated as follows:

$$\begin{aligned} \mathcal{D}(\mathbf{F}_v, \mathbf{F}_q) &= \min_{\mathbf{T} \geq 0} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{T}_{i,j} \cdot C(\mathbf{F}_{v,i}, \mathbf{F}_{q,j}), \\ s.t. \sum_{j=1}^{n} \mathbf{T}_{i,j} &= m_{v,i}, \forall i \in \{1, \ldots, n\}, \\ \sum_{i=1}^{n} \mathbf{T}_{i,j} &= m_{q,j}, \forall j \in \{1, \ldots, n\}. \end{aligned} \tag{4}$$

Here we leverage cosine similarity to define the unit cost function $C(\mathbf{F}_{v,i}, \mathbf{F}_{q,j})$:

$$C(\mathbf{F}_{v,i}, \mathbf{F}_{q,j}) = 1 - \cos(\mathbf{F}_{v,i}, \mathbf{F}_{q,j}), \tag{5}$$

as the cosine similarity between $\mathbf{F}_{v,i}$ and $\mathbf{F}_{q,j}$ increases, the corresponding unit cost decreases.

We then define the initial states of these two embeddings as all-one vectors normalized by their lengths. For the optimal transport problem, several solutions including Sinkhorn (Cuturi, 2013) and IPOT (Xie et al., 2020) incur great time complexity, we follow Kusner et al. (2015) to calculate the related moving distance which removes the second constraint to obtain the lower bound of the accurate solution. Then the optimal solution for each $\mathbf{F}_{v,i}$ is to move all its mass to the closest $\mathbf{F}_{q,j}$, and the transportation matrix becomes:

$$\mathbf{T}_{i,j} = \begin{cases} \frac{1}{n}, & if \quad j = \arg\min_{j'} C(\mathbf{F}_{v,i}, \mathbf{F}_{q,j'}), \\ 0, & otherwise. \end{cases} \tag{6}$$

Now the transport cost $\mathcal{D}(\mathbf{F}_v, \mathbf{F}_q)$ becomes:

$$\begin{aligned} \mathcal{D}(\mathbf{F}_v, \mathbf{F}_q) &= \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{T}_{i,j} \cdot C(\mathbf{F}_{v,i}, \mathbf{F}_{q,j}) \\ &= \frac{1}{n} \sum_{i=1}^{n} \min_{j} C(\mathbf{F}_{v,i}, \mathbf{F}_{q,j}). \end{aligned} \tag{7}$$

Similarly, the transport cost $D(\mathbf{F}_q, \mathbf{F}_v)$ from $\mathbf{F}_q$ to $\mathbf{F}_v$ can be derived like Eq. 7:

$$\mathcal{D}(\mathbf{F}_q, \mathbf{F}_v) = \frac{1}{n} \sum_{j=1}^{n} \min_{i} C(\mathbf{F}_{v,i}, \mathbf{F}_{q,j}). \tag{8}$$

The final transport loss $\mathcal{L}_{OT}$ is defined as follows:

$$\mathcal{L}_{OT} = \frac{\mathcal{D}(\mathbf{F}_v, \mathbf{F}_q) + \mathcal{D}(\mathbf{F}_q, \mathbf{F}_v)}{2}. \tag{9}$$

By this, visual and question embeddings interact deeply in a bidirectional and fine-grained manner.

3.4. Standard DeiT Module

Following previous works, DeiT (Touvron et al., 2021) serves as the backbone of our OTAS. Before feeding into the DeiT, we follow Zhang et al. (2020b); Wu et al. (2021); Bai et al. (2023a,c); Cheng et al. (2023) to generate a gating matrix Λ to regulate the fusion of $\mathbf{F}_v$ and $\mathbf{F}_q$:

$$\Lambda = \text{sigmoid}(\mathbf{W}_{\Lambda}^{1} \mathbf{F}_v + \mathbf{W}_{\Lambda}^{2} \mathbf{F}_q), \tag{10}$$

where $\mathbf{W}_{\Lambda}^{1}$ and $\mathbf{W}_{\Lambda}^{2}$ are two learnable matrices. After fusion, the output $\mathbf{F}$ is computed as follows:

$$\mathbf{F} = \mathbf{F}_v + \Lambda \mathbf{F}_q. \tag{11}$$

Subsequently, the fused embeddings $\mathbf{F}$ are fed into the pre-trained DeiT-Base module before the final prediction heads. The pre-trained DeiT-Base can learn fused representations and resolve ambiguous groundings from multimodal information.

3.5. Prediction Heads

Classification Head. Given an input image-question pair and a set of candidate answers, our classification head, also denoted as answer semantics-aware module (ASAM), predicts whether each candidate answer matches the corresponding image-question pair. The candidate with the highest probability is selected as the final answer. Specifically, we adopt a two-layer Transformer decoder followed by a linear projector as our classifier head, and introduce a set of learnable candidate answer embeddings together with the fused image-question embedding $\mathbf{F}$ as the input.

Let us first denote $\mathbf{A} \in \mathbb{R}^{C \times d}$ as the candidate answer embedding matrix, where C is the number of answer classes and d is the dimension of hidden states. Similar to §3.2, $\mathbf{A}$ is randomly initialized and will be updated during training through a self-attention module, a cross-attention module, and a feed-forward network in order. The self-attention computes the relationships between different answer embeddings by using $\mathbf{A}$ to construct all the *query*, *key*, and *value* matrices. The cross-attention

cares about the relationships between the answer embeddings $\mathbf{A}$ and the fused image-question embeddings $\mathbf{F}$. It thus utilizes $\mathbf{A}$ as the *query* and $\mathbf{F}$ as the *key* and *value* to compute the attention and further updates the answer embeddings by combining the attended image-question features. Mathematically, denoting the answer embeddings at the l-th layer as $\mathbf{A}_l$, it will be updated from the output of the previous layer $\mathbf{A}_{l-1}$:

$$\begin{aligned} \mathbf{A}_l &= MA(\mathbf{A}_{l-1}, \mathbf{A}_{l-1}, \mathbf{A}_{l-1}), \\ \mathbf{A}_l &= MA(\mathbf{A}_l, \mathbf{F}, \mathbf{F}), \\ \mathbf{A}_l &= FFN(\mathbf{A}_l), \end{aligned} \tag{12}$$

where $l = 1, \cdots, L$ and L is the number of Transformer decoder layers.

In this manner, the image-question embeddings are injected into the answer embeddings and used to refine the latter. The refined C are fed into the final linear projection layer followed by a softmax function to predict the probabilities of preset answer candidates, which is formulated as follows:

$$\mathbf{p} = \text{softmax}(\mathbf{W}_A \mathbf{A}_L + \mathbf{b}), \tag{13}$$

where $\mathbf{W}_A$ and $\mathbf{b}$ are learnable parameters, and $\mathbf{p}$ consists of C probabilities corresponding to C answer candidates. The answer with the highest probability is chosen as the predicted answer.

Localization Head. For the localization head, we adopt the Detection with Transformers (DETR) (Carion et al., 2020). Concretely, we utilize a feed-forward network comprising a 3-layer perceptron, ReLU activation, and a linear projection layer to model the coordinates of the bounding boxes.

3.6. Training Objective

For the classification task, we employ a straightforward cross-entropy loss $\mathcal{L}_{CE}$. For the detection task, we combine $\mathcal{L}_1$ loss and GIoU (Rezatofighi et al., 2019) loss to boost performance. The GIoU loss focuses on both overlapping regions and non-overlapping regions, defined as follows:

$$\mathcal{L}_{GIoU} = 1 - \left(\frac{|b_g \cap b_p|}{|b_g \cup b_p|} - \frac{|B(b_g, b_p) \setminus b_g \cup b_p|}{|B(b_g, b_p)|} \right), \tag{14}$$

where b_g represents the ground truth bounding box, b_p represents the predicted bounding box, $|\cdot|$ indicates the area, and B represents the operation of finding the largest box containing both b_g and b_p.

The overall training objective is formulated as:

$$\mathcal{L} = \mathcal{L}_{CE} + (\mathcal{L}_{GIoU} + \mathcal{L}_1) + \lambda \mathcal{L}_{OT}, \tag{15}$$

where λ is a trade-off hyper-parameter.

4. Experiments

4.1. Datasets and Metrics

Following previous works, EndoVis-18 (Allan et al., 2020) and EndoVis-17 (Allan et al., 2019) are taken as testbeds.[1] EndoVis-18 includes 1,560 frames and 9,014 QA pairs for training, 447 frames and 2,769 QA pairs for testing. To further prove the generalization ability of our model, we follow Bai et al. (2023c) and Bai et al. (2023a) to directly apply the model trained on EndoVis-18 to EndoVis-17, which includes 97 frames with 472 QA pairs. For metrics, we adopt accuracy (Acc), F-Score, and mean intersection over union (mIoU) (Rezatofighi et al., 2019) to evaluate the model performance.

4.2. Implementation Details

Following previous works, the dimensions of visual embedding, question embedding, and candidate answer embedding are 768 on both datasets. Our model is trained using the Adam optimizer (Kingma and Ba, 2014) for 80 epochs with a learning rate of 1×10^{-5} and a batch size of 64. For the hyper-parameter λ in Eq. 15, we set it to 0.5. All the experiments are conducted on an Nvidia V100 GPU. The experimental results are averaged over five runs with different random seeds to keep statistically stable. Codes are based on PyTorch (Paszke et al., 2019) and Transformers[2] (Wolf et al., 2020).

4.3. Main Results

We compare our OTAS with nine state-of-the-art (SOTA) baselines, including MUTAN (Ben-Younes et al., 2017), MFH (Yu et al., 2018), VisualBERT (Li et al., 2019), MCAN (Yu et al., 2019), Block-Tucker (Ben-Younes et al., 2019), VQA-DeiT (Touvron et al., 2021), VisualBERT ResMLP (Seenivasan et al., 2022a), GVLE-LViT (Bai et al., 2023c), and CAT-ViL DeiT (Bai et al., 2023a). Additionally, we conduct a robustness experiment compared to GVLE-LViT and CAT-ViL DeiT to assess the model's stability when the test data is corrupted. We apply 18 types of corruption to the test data, with the severity level ranging in severity levels from 1 to 5 following Hendrycks and Dietterich (2019).

The performance comparison and robustness experiments are shown in Table 1 and Figure 3, from which we have the following observations:

(1) *OTAS gains significant and consistent improvements on all metrics and datasets.* Specifically, on EndoVis-18, it overpasses the previous SOTA GVLE-LViT by 2.84%, 4.62% and 2.39% on

[1]The VQLA annotations are availabe at https://github.com/longbai1006/Surgical-VQLA.

[2]https://github.com/huggingface/transformers

Model	EndoVis-18			EndoVis-17		
	Acc ↑	F-Score ↑	mIoU ↑	Acc ↑	F-Score ↑	mIoU ↑
MUTAN (Ben-Younes et al., 2017)	62.83	33.95	76.39	42.42	34.82	72.18
MFH (Yu et al., 2018)	62.83	32.54	75.92	41.03	35.00	72.16
VisualBERT (Li et al., 2019)	62.68	33.29	73.91	40.05	33.81	70.73
MCAN (Yu et al., 2019)	62.85	33.38	75.26	41.37	29.32	70.29
BlockTucker (Ben-Younes et al., 2019)	62.01	32.86	76.53	42.21	35.15	72.88
VQA-DeiT (Touvron et al., 2021)	61.04	31.56	73.41	37.97	28.58	69.09
VisualBERT R (Seenivasan et al., 2022a)	63.01	33.90	73.52	41.90	33.70	71.37
GVLE-LViT (Bai et al., 2023c)	66.59	36.14	76.25	45.76	24.89	72.75
CAT-ViL DeiT (Bai et al., 2023a)	64.52	33.21	77.05	44.91	36.22	73.22
OTAS (Ours)	**68.48**	**37.81**	**78.89**	**48.62**	**37.50**	**75.83**

Table 1: Comparisons with state-of-the-art methods on EndoVis-18 and EndoVis-17 datasets. The best results and the second-best results are highlighted in **bold** and in underline, respectively.

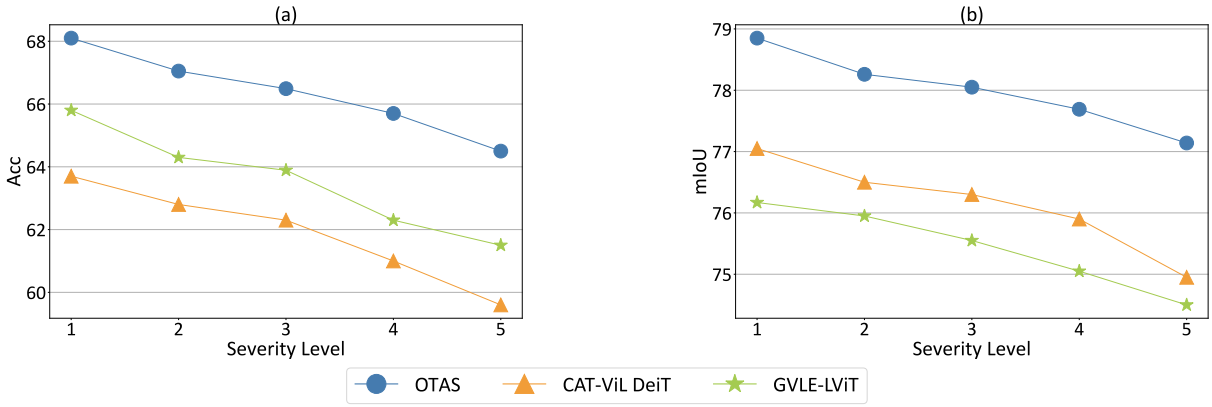

Figure 3: Robustness experiments on the EndoVis-18 dataset. (a) Acc on the classification (question-answering) task. (b) mIoU on the localization task.

Acc, F-Score and mIoU, respectively; on EndoVis-17, it overpasses CAT-ViL DeiT by 6.25%, 3.53% and 3.56% on Acc, F-Score and mIoU, respectively. This is because our model leverages the bidirectional and fine-grained alignment between images and questions, allowing the two modalities to provide crucial and subtle clues for each other. Besides, our designed answer semantics-aware module can effectively capture the semantics between questions and candidate answers, providing indicative clues for the final answer prediction.

(2) *From Figure 3, we can observe that as the severity of image corruption increases, the performance of all models degrades.* Notably, our proposed OTAS demonstrates remarkable stability in the face of corruption, consistently outperforming other models across all severity levels. The outstanding robustness of our model shows great potential for real-world applications.

4.4. Model Analysis

To understand our OTAS in more depth, we perform comprehensive studies to answer the following research questions (RQs): (1) Does each proposed module contribute to the overall performance of the model? (2) How do different OT solutions influence performance? (3) Should the OT-based alignment module precede or follow the coarse-grained alignment module? (4) Is bidirectional cross-modal information interaction in the OT-based alignment module necessary? (5) Is there an impact of answer embedding size in the answer semantics-aware module on performance? (6) How do pre-trained models impact the initialization of question and answer embeddings? (7) Qualitatively, does our method truly stand out in the VQLA systems?

Answer 1: Each proposed module contributes to improving the model performance. The core contributions of this work are the optimal transport-based alignment module (OTAM) and the answer semantics-aware module (ASAM). From Table 2, we observe that removing either module results in a sharp decline in the model's performance across all metrics and datasets, which demonstrates the effectiveness of our proposed modules. Interestingly, OTAM has a greater impact on the localization task, whereas ASAM significantly influences the question-answering task. This further validates our motivation for developing these two modules. Since

Model	EndoVis-18			EndoVis-17		
	Acc ↑	F-Score ↑	mIoU ↑	Acc ↑	F-Score ↑	mIoU ↑
OTAS (Ours)	**68.48**	**37.81**	**78.89**	**48.62**	**37.50**	**75.83**
w/o OTAM	67.65(↓0.83)	37.06(↓0.75)	77.68(↓1.21)	47.35(↓1.27)	36.94(↓0.56)	74.16(↓1.67)
+ More Parameters	67.76(↓0.72)	37.22(↓0.59)	77.81(↓1.08)	47.48(↓1.14)	37.02(↓0.48)	74.40(↓1.43)
w/o CAM	67.83(↓0.65)	37.28(↓0.53)	77.93(↓0.96)	47.54(↓1.08)	37.10(↓0.40)	74.44(↓1.39)
w/o ASAM	67.41(↓1.07)	36.91(↓0.90)	77.97(↓0.92)	47.10(↓1.52)	36.71(↓0.79)	74.25(↓1.58)

Table 2: Results of ablation experiments. "OTAM" and "ASAM" represents "Optimal Transport-based Alignment Module" and "Answer Semantics-aware Module", respectively.

Model	EndoVis-18			
	Acc ↑	F-Score ↑	mIoU ↑	Speed ↓
Sinkhorn	65.57	35.95	76.56	567
IPOT	68.53	37.80	78.92	543
Ours	68.48	37.81	78.89	280

Table 3: Comparisons with different OT solutions on the EndoVis-18 dataset. Speed denotes the average training time (ms) for each batch.

Model	EndoVis-18		
	Acc ↑	F-Score ↑	mIoU ↑
OTAM -> CAM	67.90	37.28	77.56
CAM -> OTAM (Q->V + V->Q)	68.48	37.81	78.89
w/o Q->V	67.95	37.37	78.15
w/o V->Q	68.06	37.40	78.23

Table 4: Comparisons with different OT positions and structures on the EndoVis-18 dataset. "CAM" represents "Coarse-grained alignment module". Q->V (resp. V->Q) represents executing OT from question embeddings to visual embeddings (resp. visual embeddings to question embeddings).

OTAM introduces more parameters, a natural question is whether the additional parameters involved in OTAS contribute to the final performance. To this end, we remove OTAM and expand the number of layers of the co-attention module to six layers to validate that the proposed OTAM rather than the extra parameters contribute to performance improvement. We refer it to w/o OTAM + More Parameters in Table 2. We observe that though more parameters slightly improve the performance, there is still a significant gap w.r.t. the proposed OTAS, which verifies that the improvements indeed come from OTAM rather than the involved parameters.

Answer 2: Our proposed OTAM achieves a nice trade-off between model performance and training efficiency. Due to the higher time complexity of Sinkhorn (Cuturi, 2013) and IPOT (Xie et al., 2020) algorithms, in this paper, we utilize the relaxed moving distance to calculate the lower bound of the original problem. To verify the effectiveness of the relaxed moving distance, we replace it with Sinkhorn and IPOT, respectively. The corresponding results are shown in Table 3. We observe that OTAM maintains a nearly comparable performance to the exact solution IPOT and has a significant speed advantage, which demonstrates the superiority of the relaxed moving distance.

Answer 3: The transition of information from coarse-grained to fine-grained is effective in the VQLA system. As a plug-in module, OTAM offers flexibility in its placement within the network. Therefore, we aim to investigate the rationality and effectiveness of its current design. From the results in Table 4, we observe that putting coarse-grained alignment module (CAM) after OTAM results in performance deterioration. Specifically, the accuracy decreases by 0.58% in the question-answering task, and the mIoU decreases by 1.33% in the localization task. The probable reason is that coarse alignment in CAM may destroy the subtle cross-modality information learned by the OTAM. It is also interesting to explore a concurrent cross-modal attention mechanism in the future.

Answer 4: The bidirectional cross-modal information interaction in the OT-based alignment module enables comprehensive information exchange between the two modalities. We also investigate the impact of bidirectional information interaction in the OT-based alignment module. As observed from the last two rows of Table 4, both text-to-image and image-to-text information exchanges enhance the final performance, underscoring the necessity of bidirectional information interaction.

Answer 5: The size of the answer embeddings in answer semantics-aware module has a certain impact on performance. We also explore the impact of answer embedding dimensions. Figure 5 illustrates that as the dimension of answer embeddings increases, the model's performance improves while the best result is obtained when the embedding size is approximately 1024. However, increasing the embedding size escalates computational costs, while the performance saturates

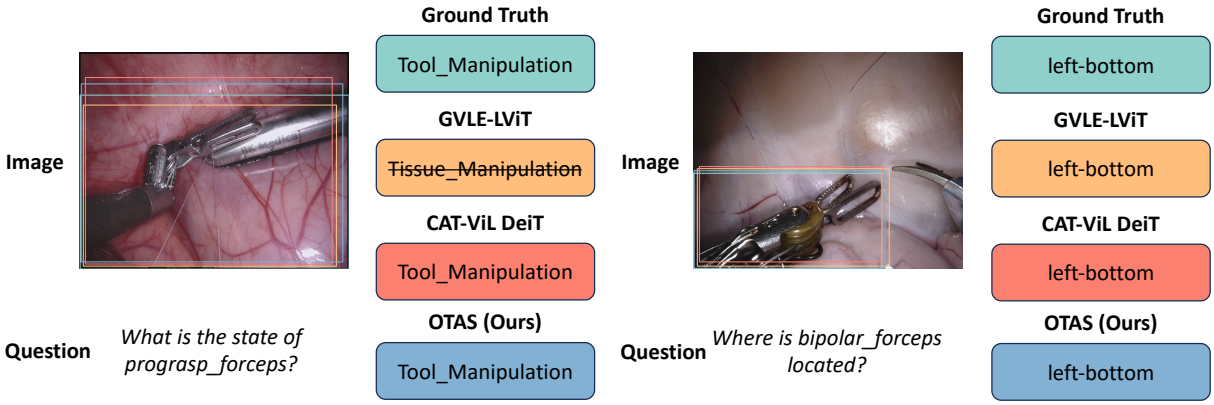

Figure 4: Two case studies on the VQLA task. Our OTAS ( ) displays state-of-the-art (SOTA) performance in generating the answers and location against GVLE-LViT ( ) (Bai et al., 2023c) and CAT-ViL DeiT ( ) (Bai et al., 2023a). The ground truth bounding box and its answer are marked in . ~~Strikethrough~~ words present the incorrect answers. Zoom in for better viewing.

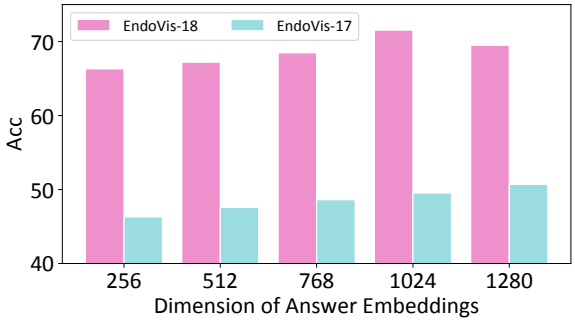

Figure 5: Experiments on the dimension of answer embeddings on two benchmark datasets.

Model	Initiation	EndoVis-18		
		Acc ↑	F-Score ↑	mIoU ↑
OTAS w/o ASAM	-	67.41	36.91	77.97
OTAS (Ours)	Random	68.48	37.81	78.89
OTAS	ClinicalBERT	70.72	40.65	79.54
OTAS	PubMedBERT	71.62	39.06	79.60

Table 5: Comparisons with different initiation strategies of question and answer embeddings.

quickly. As a trade-off, our model adopts 768-dimensional answer embeddings.

Answer 6: The incorporation of pre-trained models can further improve the model performance. One of our core contributions is the answer semantics-aware module (ASAM), which introduces a set of learnable answer embeddings. These embeddings effectively reveal the rich semantic associations between questions and answers through interaction with fused image-question features. Here we investigate the power of encoders pre-trained on the large medical corpus. To maintain consistency in the embedding space, we substitute both the question encoder and the answer encoder. As shown in Table 5, while comparing with the models that adopt ClinicalBERT (Alsentzer et al., 2019) or PubMedBERT (Gu et al., 2021) as question and answer encoders, we observe that guiding fused image-question features with domain knowledge generally works better, *e.g.*, results of using ClinicalBERT or PubMedBERT outperform training with no answer semantics. Besides, our proposed method achieves consistent improvements over baselines while being lightweight.

Answer 7: Our model demonstrates qualitative superiority over other state-of-the-art methods. To better understand the proposed model, we show two cases in Figure 4. In the left case, our model outperforms the other two competitive models by being closer to the ground truth bounding box in the localization task. Moreover, GVLE-LViT incorrectly predicts "*Tissue_Manipulation*" in the question-answering task. Our model effectively captures the semantic association between the answer and the question, enabling accurate predictions. In the right case, despite the simplicity of the question-answering task, our method still achieves competitive results in the localization task, showcasing the superiority of our proposed model.

5. Conclusion

In this paper, we proposed a novel model termed OTAS for VQLA tasks, which creatively introduced optimal transport to achieve bidirectional and fine-grained alignment between questions and images. Moreover, we incorporated answer semantic information into the answer class prediction process to correlate the answering embeddings with the

fused image-question features, which improved the accuracy significantly. Experiment results on two benchmark datasets showed that our model significantly outperformed previous models.

Future Work. Future work will extend beyond classification-based Question Answering (QA) tasks and delve into generative QA tasks.

6. Bibliographical References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

Long Bai, Mobarakol Islam, and Hongliang Ren. 2023a. CAT-ViL: Co-attention gated vision-language embedding for visual question localized-answering in robotic surgery. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 397–407. Springer.

Long Bai, Mobarakol Islam, and Hongliang Ren. 2023b. Revisiting distillation for continual learning on visual question localized-answering in robotic surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 68–78. Springer.

Long Bai, Mobarakol Islam, Lalithkumar Seenivasan, and Hongliang Ren. 2023c. Surgical-VQLA: Transformer with gated vision-language embedding for visual question localized-answering in robotic surgery. *arXiv preprint arXiv:2305.11692*.

Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. 2017. MUTAN: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2612–2620.

Hedi Ben-Younes, Remi Cadene, Nicolas Thome, and Matthieu Cord. 2019. BLOCK: Bilinear superdiagonal fusion for visual question answering and visual relationship detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8102–8109.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with Transformers. In *European Conference on Computer Vision*, pages 213–229. Springer.

Ilias Chalkidis, Manos Fergadiotis, Sotiris Kotitsas, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. An empirical study on large-scale multi-label text classification including few and zero-shot labels. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7503–7515.

Liqun Chen, Yizhe Zhang, Ruiyi Zhang, Chenyang Tao, Zhe Gan, Haichao Zhang, Bai Li, Dinghan Shen, Changyou Chen, and Lawrence Carin. 2018. Improving sequence-to-sequence learning via optimal transport. In *International Conference on Learning Representations*.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: Universal image-text representation learning. In *European Conference on Computer Vision*, pages 104–120. Springer.

Zhaorun Chen, Zhuokai Zhao, Zhihong Zhu, Ruiqi Zhang, Xiang Li, Bhiksha Raj, and Huaxiu Yao. 2024. AutoPRM: Automating procedural supervision for multi-step reasoning via controllable question decomposition. *NAACL*.

Xuxin Cheng, Zhihong Zhu, Yaowei Li, Hongxiang Li, and Yuexian Zou. 2023. DAS-CL: Towards multimodal machine translation via dual-level asymmetric contrastive learning. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 337–347, New York, NY, USA. Association for Computing Machinery.

Marco Cuturi. 2013. Sinkhorn Distances: Lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*, 26.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare*, 3(1):1–23.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*.

Min-Chai Hsieh and Yu-Hsuan Lin. 2017. VR and AR applications in medical practice and education. *Hu Li Za Zhi*, 64(6):12–18.

Yunpeng Huang, Yaonan Gu, Jingwei Xu, Zhihong Zhu, Zhaorun Chen, and Xiaoxing Ma. 2024. Securing reliability: A brief overview on enhancing in-context learning for foundation models. *arXiv preprint arXiv:2402.17671*.

Leonid V Kantorovich. 2006. On the translocation of masses. *Journal of Mathematical Sciences*, 133(4):1381–1382.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. 2015. From word embeddings to document distances. In *International Conference on Machine Learning*, pages 957–966. PMLR.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Henry C Lin, Izhak Shafran, David Yuh, and Gregory D Hager. 2006. Towards automatic skill evaluation: Detection and segmentation of robot-assisted surgical motions. *Computer Aided Surgery*, 11(5):220–230.

Antoine Liutkus, Umut Simsekli, Szymon Majewski, Alain Durmus, and Fabian-Robert Stöter. 2019. Sliced-Wasserstein flows: Nonparametric generative modeling via optimal transport and diffusions. In *International Conference on Machine Learning*, pages 4104–4113. PMLR.

Kelly Marchisio, Ali Saad-Eldin, Kevin Duh, Carey Priebe, and Philipp Koehn. 2022. Bilingual lexicon induction for low-resource languages using graph matching via optimal transport. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2545–2561.

James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. 2018. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Proc. of NeurIPS*.

Gabriel Peyré, Marco Cuturi, et al. 2019. Computational Optimal Transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Jianjun Qian, Shumin Zhu, Chaoyu Zhao, Jian Yang, and Wai Keung Wong. 2023. OTFace: Hard samples guided optimal transport loss for deep face representation. *IEEE Transactions on Multimedia*, 25:1427–1438.

Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 658–666.

Anthony Rios and Ramakanth Kavuluru. 2018. Few-shot and zero-shot multi-label learning for structured label spaces. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2018, pages 3132–3142.

Litu Rout, Alexander Korotin, and Evgeny Burnaev. 2021. Generative modeling with optimal transport maps. In *International Conference on Learning Representations*.

Lalithkumar Seenivasan, Mobarakol Islam, Adithya K Krishna, and Hongliang Ren. 2022a. Surgical-VQA: Visual question answering in surgical scenes using Transformer. In *International Conference on Medical Image Computing and Computer Assisted Intervention*, pages 33–43. Springer.

Lalithkumar Seenivasan, Sai Mitheran, Mobarakol Islam, and Hongliang Ren. 2022b. Global-reasoned multi-task learning model for surgical scene understanding. *IEEE Robotics and Automation Letters*, 7(2):3858–3865.

Dhruv Sharma, Sanjay Purushotham, and Chandan K Reddy. 2021. MedFuseNet: An attention-based multimodal deep learning model for visual question answering in the medical domain. *Scientific Reports*, 11(1):19826.

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*, pages 10347–10357. PMLR.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez,

Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010.

T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al. 2020. Transformers: State-of-the-art natural language processing. In *Proc. of EMNLP*.

Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. 2021. Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6153–6166.

Yujia Xie, Xiangfeng Wang, Ruijia Wang, and Hongyuan Zha. 2020. A fast proximal point method for computing exact Wasserstein distance. In *Uncertainty in Artificial Intelligence*, pages 433–453. PMLR.

Derong Xu, Ziheng Zhang, Zhihong Zhu, Zhenxi Lin, Qidong Liu, Xian Wu, Tong Xu, Xiangyu Zhao, Yefeng Zheng, and Enhong Chen. 2024. Editing factual knowledge and explanatory ability of medical large language models. *arXiv preprint arXiv:2402.18099*.

Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6281–6290.

Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. 2018. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE Transactions on Neural Networks and Learning Systems*, 29(12):5947–5959.

Chi Zhang, Yujun Cai, Guosheng Lin, and Chunhua Shen. 2020a. DeepEMD: Few-shot image classification with differentiable earth mover's distance and structured classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12203–12213.

Zhuosheng Zhang, Kehai Chen, Rui Wang, Masao Utiyama, Eiichiro Sumita, Zuchao Li, and Hai Zhao. 2020b. Neural machine translation with universal visual representation. In *International Conference on Learning Representations*.

Yan Zhou, Qingkai Fang, and Yang Feng. 2023. CMOT: Cross-modal mixup via optimal transport for speech translation. *arXiv preprint arXiv:2305.14635*.

Zhihong Zhu, Xuxin Cheng, Zhiqi Huang, Dongsheng Chen, and Yuexian Zou. 2023a. Enhancing code-switching for cross-lingual SLU: A unified view of semantic and grammatical coherence. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7849–7856, Singapore. Association for Computational Linguistics.

Zhihong Zhu, Xuxin Cheng, Zhiqi Huang, Dongsheng Chen, and Yuexian Zou. 2023b. Towards unified spoken language understanding decoding via label-aware compact linguistics representations. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12523–12531, Toronto, Canada. Association for Computational Linguistics.

Zhihong Zhu, Xuxin Cheng, Hongxiang Li, Yaowei Li, and Yuexian Zou. 2024. Dance with labels: Dual-heterogeneous label graph interaction for multi-intent spoken language understanding. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24, page 1022–1031, New York, NY, USA. Association for Computing Machinery.

Zhihong Zhu, Weiyuan Xu, Xuxin Cheng, Tengtao Song, and Yuexian Zou. 2023c. A dynamic graph interactive framework with label-semantic injection for spoken language understanding. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

7. Language Resource References

Max Allan, Satoshi Kondo, Sebastian Bodenstedt, Stefan Leger, Rahim Kadkhodamohammadi, Imanol Luengo, Felix Fuentes, Evangello Flouty, Ahmed Mohammed, Marius Pedersen, et al. 2020. 2018 Robotic Scene Segmentation Challenge. *arXiv preprint arXiv:2001.11190*.

Max Allan, Alex Shvets, Thomas Kurmann, Zichen Zhang, Rahul Duggal, Yun-Hsuan Su, Nicola Rieke, Iro Laina, Niveditha Kalavakonda, Sebastian Bodenstedt, et al. 2019. 2017 Robotic Instrument Segmentation Challenge. *arXiv preprint arXiv:1902.06426*.