



Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/infus

Surgical-VQLA++: Adversarial contrastive learning for calibrated robust visual question-localized answering in robotic surgery

Long Bai^{a,b,1}, Guankun Wang^{a,b,1}, Mobarakol Islam^{c,1}, Lalithkumar Seenivasan^{a,d}, An Wang^{a,b}, Hongliang Ren^{a,b,d,*}

^a Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong, China

^b Shenzhen Research Institute, The Chinese University of Hong Kong, Shenzhen, China

^c Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS), University College London, London, UK

^d Department of Biomedical Engineering, National University of Singapore, Singapore

ARTICLE INFO

Keywords:

Surgical education
Vision-language embedding
Adversarial contrastive learning
Image corruption
Visual-question answering

ABSTRACT

Medical visual question answering (VQA) bridges the gap between visual information and clinical decision-making, enabling doctors to extract understanding from clinical images and videos. In particular, surgical VQA can enhance the interpretation of surgical data, aiding in accurate diagnoses, effective education, and clinical interventions. However, the inability of VQA models to visually indicate the regions of interest corresponding to the given questions results in incomplete comprehension of the surgical scene. To tackle this, we propose the surgical visual question localized-answering (VQLA) for precise and context-aware responses to specific queries regarding surgical images. Furthermore, to address the strong demand for safety in surgical scenarios and potential corruptions in image acquisition and transmission, we propose a novel approach called Calibrated Co-Attention Gated Vision-Language (C²G-ViL) embedding to integrate and align multimodal information effectively. Additionally, we leverage the adversarial sample-based contrastive learning strategy to boost our performance and robustness. We also extend our EndoVis-18-VQLA and EndoVis-17-VQLA datasets to broaden the scope and application of our data. Extensive experiments on the aforementioned datasets demonstrate the remarkable performance and robustness of our solution. Our solution can effectively combat real-world image corruption. Thus, our proposed approach can serve as an effective tool for assisting surgical education, patient care, and enhancing surgical outcomes. Our code and data will be released at <https://github.com/longbai1006/Surgical-VQLAPlus>.

1. Introduction

Comprehensive responses to inquiries on surgery and surgical procedures should be ideally provided by proficient surgeons. However, these scarce highly skilled surgeons often face difficulties in finding time to respond to such queries due to their demanding academic and clinical workload [1,2]. While recorded surgical videos have been provided to medical students to facilitate learning through observation, observational learning does not cater to specific queries raised by individual students. Its effectiveness largely depends on the student's ability to infer conclusions from videos. To address this to an extent, Surgical-Visual Question Answering (VQA) [1] was introduced to answer questions on robotic surgery based on the given surgical scene. To help students better understand the predicted answer, we earlier introduced surgical Visual Question Localized-Answering (VQLA) [3],

which extended the VQA task further by localizing the regions in the image that are highly relevant to the predicted answer. Extending our earlier work, here, we propose VQLA++ to improve the performance and robustness of VQLA in robotic surgery.

Numerous VQA models have emerged in the computer vision domain in recent times, showcasing impressive performance by effectively capturing semantic associations between images and questions [4–7]. Advanced techniques such as multimodal fusion [8] and attention mechanisms [9] have also been explored and incorporated to enhance performance. For their potential use in assisting medical professionals in analyzing medical images and providing insightful answers to specific queries, VQA models have gained significant attention in the medical domain [10,11]. A medical VQA system that addresses the unique challenges in the medical domain, such as complex anatomical

* Corresponding author at: Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong, China.

E-mail addresses: b.long@link.cuhk.edu.hk (L. Bai), gkwang@link.cuhk.edu.hk (G. Wang), mobarakol.islam@ucl.ac.uk (M. Islam), lalithkumar_s@u.nus.edu (L. Seenivasan), wa09@link.cuhk.edu.hk (A. Wang), hren@ee.cuhk.edu.hk (H. Ren).

¹ Long Bai, Guankun Wang, and Mobarakol Islam are co-first authors.

<https://doi.org/10.1016/j.infus.2024.102602>

Received 3 October 2023; Received in revised form 28 May 2024; Accepted 25 July 2024

Available online 27 July 2024

1566-2535/© 2024 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

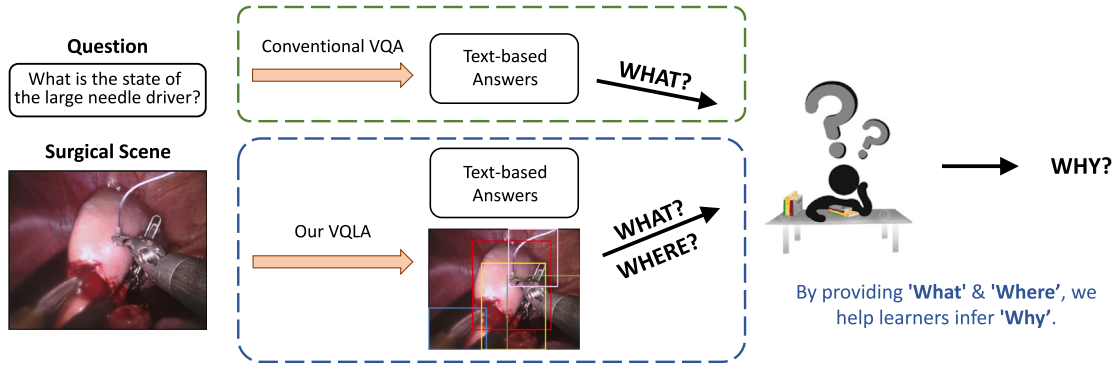


Fig. 1. Comparison of the conventional VQA and our VQLA model. By providing the answer to ‘What’ and ‘Where’, we can help learners to better infer ‘Why’ and achieve a better understanding of surgical scenes.

structures, diverse imaging modalities, and domain-specific terminology, has the potential to serve as a reliable tool that can complement medical experts in responding to questions posed by patients and students [1,12,13].

Surgical-VQA [1] has made a significant contribution to the medical surgical sub-domain by introducing VQA in robotic surgery. However, certain limitations hinder its effectiveness in assisting students with the comprehension of intricate surgical scenarios. In situations where expert guidance is unavailable for immediate responses to inquiries, understanding the rationale behind the answers often necessitates answers to both the “what” and “where” aspects. Traditional VQA approaches primarily focus on generating answers to the “what” question, neglecting the crucial dimension of “where” in the surgical scene the answer was derived from. This limitation poses challenges for students and trainees in deducing the reasons behind the predicted answer. For example, when a student queries the interaction between a surgical instrument and tissue, the VQA model can only provide a basic response without explicitly identifying the precise location of the tool and tissue within the surgical scene. Consequently, students still require assistance in comprehending the intricacies of complex surgical scenarios. Furthermore, the annotation of large amounts of textual data requires the involvement of expert surgeons, resulting in time-consuming and labor-intensive processes. To tackle these challenges, we earlier proposed the Surgical-VQLA [3], which aims to highlight the specific area relevant to the posed question, as depicted in Fig. 1. By integrating the answer to the “what” question with the localization results and observations of the surgical area, students can achieve a more comprehensive understanding of specific surgical procedures, instrument usage, and anatomical structures involved. Additionally, observing the surrounding surgical area provides valuable contextual information that facilitates students’ comprehension of the relationships and interactions among different elements in the surgical scene. By leveraging this integrated information, students can gain deeper insights into the reasoning behind surgical procedures. This enriched understanding enables them to infer the underlying causes and decisions, thereby enhancing their analytical capabilities and enabling more informed learning experiences.

In our earlier work [3], apart from defining the Surgical-VQLA system, we addressed the challenges of detection-based visual feature extractor, to solve the problem of non-end-to-end, non-global understanding, and slow inference speed. We also effectively balanced the fusion between two modalities using a gated module. Moreover, we further proposed CAT-ViL [14], which utilized a co-attention module to facilitate interaction between the two modalities, achieving improved fusion performance. However, when applied in medical and clinical settings, models may face the challenge of maintaining high robustness. Medical data often contain various types of noise, artifacts, or image quality degradation. Fig. 2 shows the effect that different levels of image corruption cause on the performance of the existing model

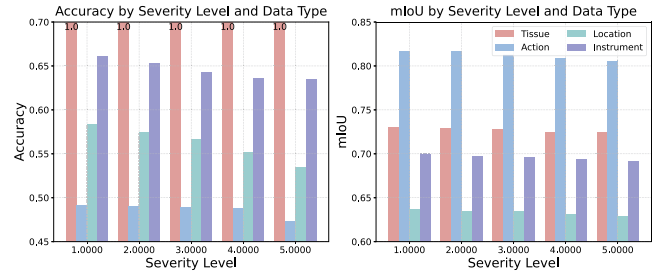


Fig. 2. Robustness results of Surgical-VQLA [3] for different types of samples at each severity level. The four types of samples include the tissue being operated on, the current state of the instruments, the location of the instruments in the operating area, and the identification of the instruments. We can observe that as the severity level of corruption increases, the performance of the model significantly decreases.

Surgical-VQLA [3]. There is a significant performance degradation as the level of image corruption increases. Moreover, these data are typically collected from different time points and institutions, leading to domain gaps [15]. Therefore, it is imperative to develop a robust model capable of handling the sensitivity to diverse interfering factors and delivering stable and reliable performance in real-world scenarios. Robust models can handle various types of corruption efficiently and help minimize the risk of misdiagnosis, enabling healthcare professionals to make more accurate and informed decisions regarding patient care [16]. To address these challenges, we propose a joint framework that integrates feature calibration and adversarial contrastive training techniques. This integration aims to enhance the model’s robustness and ensure its efficacy across various practical applications.

By introducing multimodal collaborated calibration and global contextual calibration mechanisms, we develop a new Surgical-VQLA++ framework. This model effectively aligns and normalizes multimodal representations and enhances robustness by capturing subtle feature perturbations through contrastive representation learning of adversarial examples. Through extensive experiments, our proposed solution has demonstrated significant improvements in performance and robustness. Our model surpasses the second-best model by achieving an improvement of 1.12% in accuracy and 1.55% in mIoU in overall performance. In robustness tests, it further exceeds the second-best by 1.64% in accuracy and 1.70% in mIoU, demonstrating the effectiveness of our proposed method.

In this work, our key contributions are:

- We present a Surgical-VQLA++ framework that establishes an instance-level connection between answering and localization.

The proposed model exhibits impressive performance and robustness, rendering it highly suitable for clinical applications. Furthermore, our feature extraction strategy promotes global understanding, compatibility with end-to-end solutions, and an efficient inference speed of 150.6 FPS.

- To enhance the alignment and interaction of multimodal representations, as well as globally calibrate contextual features, we propose a Calibrated Co-attention Gated Vision-Language (C²G-ViL) embedding module. This module explores optimal fusion weights and facilitates the alignment between different modalities.
- Building upon the DeiT backbone [17] for deep feature learning, we employ an adversarial contrastive training strategy specifically on the DeiT backbone. This strategic utilization aims to enhance the model's capacity to discern and capture intricate nuances or patterns, further boosting model performance and robustness. For multi-task convergence, we investigate various combinations of loss functions and move away from naive loss combinations.
- We present two comprehensive surgical datasets for VQLA tasks based on the public EndoVis18 [18] and EndoVis17 [19] datasets. We have expanded our dataset to include a total of 17,269 QA pairs by adding additional queries about surgical instruments. These QA pairs cover contents such as surgical organs, instruments, actions, and instrument locations. Each pair also comes with a corresponding bounding box to localize the answer.

2. Related work

2.1. Grounded VQA in computer vision

VQA and answer grounding in computer vision have gained attention in recent times [20]. Antol et al. [21] introduced the VQA task to answer natural language questions based on the input images. Subsequently, numerous multimodal models incorporating advanced techniques were introduced in the computer vision domain to improve the performance on VQA tasks [4,22,23]. The potential use of the VQA model for critical tasks has emphasized the exploration of model interpretability [24]. Visual grounding, an emerging topic within the vision-language tasks [25–27], is being widely explored in the development of interpretable models and dependable scene understanding in VQA. Incorporating relevant visual evidence through language queries has been observed to help effectively guide the model's attention and contribute to achieving state-of-the-art (SOTA) performance in VQA tasks [28]. As a result, the effective execution of both question-answering and answer grounding has emerged as a prominent research area.

The integration of VQA and answer grounding techniques facilitates a more profound comprehension of visual content in computer vision research. Through the grounding process, answers can be linked to specific regions or objects within an image, providing precise localization information. This integration not only enables accurate answer generation but also offers explanations by connecting the selected answer to relevant visual evidence, thereby enhancing interpretability [29]. Achieving interpretability is particularly important in critical domains like healthcare, where building a trustworthy and robust understanding of computer vision systems is essential. Numerous studies have focused on developing advanced deep learning models integrating VQA and answer grounding. For instance, Fukui et al. [30] proposed a compact bilinear pooling model for VQA and visual grounding. In contrast, Zhu et al. [31] introduced a dual-decoder transformer network specifically designed for answer grounding in VQA. Zhang et al. [32] proposed an interpretable VQA approach by incorporating visual grounding through attention supervision mining. Reich et al. [33] presented a method to measure the faithfulness and plausibility of visual grounding in VQA. By guiding the model's attention to the most salient areas of an image relevant to the question, answer grounding enhances reasoning

abilities and improves answer accuracy. Mani et al. [34] integrated object proposals into VQA by utilizing pointing to indicate the location of the answer within the image. However, these approaches, designed for natural images, pose challenges in terms of inference speed and global contextual understanding due to their object proposal design. The medical domain, with its complex and sensitive medical imaging data, presents unique challenges. Adapting VQA and answering grounding tasks to the medical domain requires addressing concerns such as interpretability, transparency, and uncertainty, making it an active area of research.

2.2. Grounded VQA in the medical domain

Medical VQA has made significant contributions to the medical field, utilizing the fusion of images and questions to assist with clinical diagnosis and treatment. This includes aiding radiologists [35], enhancing clinical decision support [36], facilitating patient monitoring and assessment [1], supporting education and training [3], enabling clinical research and data analysis [37], and more. However, medical images possess unique characteristics that require specialized knowledge for accurate interpretation, posing additional challenges in developing medical VQA models. Recent advancements in medical VQA research have focused on addressing these challenges. Liu et al. [35] proposed contrastive pre-training and representation distillation, Naseem et al. [36] introduced knowledge-aware multimodal representation, and Liu et al. [12] presented an answer querying decoder. These works have significantly improved performance in their respective tasks.

Despite these advancements, answer grounding or VQLA in the medical domain remains relatively unexplored. Tascon et al. [38] discussed localized questions but cropped the target regions, which limits global understanding. They later improved their solution by localizing the question in the image using attention maps to enhance VQA performance [39]. Similarly, Vu et al. [40] employed multi-glimpse attention to localize image regions relevant to given questions. However, these methods did not consider the framework of multi-task learning, focusing only on how to improve the performance of VQA itself by localizing answers. Instead, the VQLA system can simultaneously output question-answers and localizations, offering several potential benefits. Another interesting study locates relevant video snippets from medical instructive videos to answer textual questions [41]. However, this study is not directly applicable to the tasks of VQA or VQLA as it solely focuses on locating matching spans in videos that are relevant to given questions.

The multi-task learning framework can leverage shared representations and interactions between question-answering and localization, improving performance in both tasks through joint training and information sharing [42,43]. Additionally, simultaneously outputting question-answering and localization can provide more comprehensive and informative results. The model can offer visual explanations or justifications for its answers, enhancing the interpretability and trustworthiness of the system.

3. Methodology

Our VQLA system architecture, as illustrated in Fig. 3, incorporates two feature extractors, namely the C²G-ViL embedding module and the pre-trained DeiT backbone. Additionally, it includes two prediction heads responsible for classification and bounding box generation tasks. To optimize the model's performance, we employ the adversarial contrastive training strategy during the training process.

3.1. Preliminaries

3.1.1. VisualBERT

VisualBERT [4] performs joint reasoning on vision and language tasks, by integrating Transformer [44] models and Faster RCNN [45], which extended the BERT [46] NLP model. Initially, tokens were generated from textual inputs, while visual features were extracted using Faster RCNN [45]. VisualBERT then employed element-wise addition operations on the token, segment, and positional embedding from the textual input, in addition to the feature, segment, and positional embedding from the visual input. These operations allowed both modalities to be combined effectively through concatenation. Subsequently, Transformer layers were utilized to boost interactions between the two modalities, which facilitated the learning of joint representations. Finally, prediction modules specific to various ViL tasks, such as VQA, visual answer justification, and visual reasoning were designed to perform predictions. The simplistic, efficient, and prompt design of multimodal fusion in VisualBERT effectively enhances subsequent modules to facilitate multimodal joint reasoning.

3.1.2. Multi-head attention

Multi-head attention (MHA) [44] is a mechanism that facilitates simultaneous attention to multiple regions of the input, enabling the model to capture diverse and complementary information effectively. Unlike relying solely on a single attention mechanism, multi-head attention involves the creation of multiple sets of attention weights, referred to as heads, which are computed independently. Each head attends to different portions of the input, thereby providing multiple perspectives and insights into the underlying data. The outputs generated by these heads are subsequently combined using linear transformations, thereby enhancing the model's capacity to capture intricate relationships and improve overall performance. In our formulation, we represent the input queries, keys, and values for each head i as Q_i , K_i , and V_i , respectively, where i ranges from 1 to h . Accordingly, let $Attn$ denote the single-head attention mechanism, and consequently, we can express this as follows:

$$h_i = Attn(W_{Q_i}Q_i, W_{K_i}K_i, W_{V_i}V_i) \quad (1)$$

A linear transformation is subsequently utilized for the aggregation of attention across multiple heads. W_o represents the trainable parameters in the multiple heads, and the MHA shall be formulated as:

$$h = MultiAttn(W_o [h_1 \parallel \dots \parallel h_h]) \quad (2)$$

in which $[\cdot \parallel \cdot]$ represents concatenation.

3.1.3. Adversarial examples

Adversarial examples are carefully constructed inputs that exploit vulnerabilities in the decision-making process of machine learning models to deceive them. These examples are specifically designed to closely resemble the original data while incorporating subtle modifications, known as perturbations. Perturbations refer to minor alterations introduced to the input data, often imperceptible to human observers, with the objective of inducing incorrect predictions or unexpected behavior from the model. Goodfellow et al. [47] proposed Fast Gradient Sign Method (FGSM) to generate adversarial examples. We follow [48] to perform perturbation on the fused embedding. Let x and G denote the input data and ground truth, respectively. Given a deep model F , the loss function can be defined as $\mathcal{L}(F(x+r), G)$, where r represents the perturbation. With enforcing a maximum constraint on r , we can obtain $\{\max_r \mathcal{L}(F(x+r), G)\}$ s.t. $\|r\|_\infty < \epsilon$. It can be approximated via the first-order approximation:

$$\mathcal{L}(F(x+r), G) \approx \mathcal{L}(F(x), G) + \nabla_x \mathcal{L}(F(x), G)^T r \quad (3)$$

Finally, we can obtain the perturbation as:

$$r = -\epsilon \text{sign}(\nabla_x \mathcal{L}(F(x), G)) \quad (4)$$

3.2. Feature extraction

Two parallel feature extractors are employed in this study to handle visual and textual representations separately. For visual representation, a visual feature extractor is utilized to obtain feature representations from original images. However, unlike the original VisualBERT [4], instead of adopting an object proposal-based strategy, we employ a ResNet18 [49] network that has been pre-trained on ImageNet [50]. This choice is made because object detection models solely focus on key objects and fail to consider surgical instruments and the surrounding environment, which play a crucial role in prediction accuracy [51]. Moreover, any errors in the detection process could significantly impact the final prediction results. In contrast, ResNet18 serves effectively as a global feature extractor, providing a holistic understanding of the target image compared to Faster RCNN [45], which is limited to local key objects. Additionally, pre-trained ResNet18 offers faster inference speed and an end-to-end solution.

In parallel to the visual feature extractor, a tokenizer is applied, specifically trained on the surgical dataset, to convert input questions into embedding matrices, which can subsequently be processed by the deep model. Following this, features from both modalities are fed into the C²G-ViL embedding module for further processing and fusion.

3.3. C²G-ViL embedding

In this study, we initially employ the VisualBERT strategy to obtain separate visual and text embedding. However, previous approaches have taken a simplistic approach to data fusion, utilizing concatenation [1,4], summation [52], or MLP layer [23]. This rudimentary fusion method fails to capture the active interaction between information and statistical representations from the two modalities, thereby limiting their effectiveness in subsequent tasks. To address this limitation, we propose a novel Calibrated Co-Attention Gated (C²G)-ViL embedding that aims to calibrate and aggregate the visual and textual representations.

Our C²G-ViL embedding comprises a co-attention interaction module, a multimodal collaborated calibration module, a global contextual calibration module, and a gated fusion module. Firstly, the co-attention module integrates guided-attention and self-attention mechanisms to enhance the interaction between modalities. By incorporating guided-attention blocks, our model is able to direct its attention towards key regions in images, further improving its understanding of visual content. Secondly, two calibration modules are engaged to ensure the robustness of models for reliable performance in clinical settings. These models need to accommodate variations in data, including diverse patient populations, imaging techniques, and image quality. Therefore, we introduce a process of representation calibration to enhance performance by reducing variations and discrepancies within the embedding space. This calibration process facilitates more accurate feature differentiation and dependable predictions [53]. Finally, we leverage a gate fusion operation to dynamically learn and control the fusion weights of different modalities. This dynamic fusion approach aims to achieve a more effective integration of multimodal representations, optimizing the overall performance of the model.

3.3.1. Co-attention cross-model interaction

The co-attention block utilized in this study aims to capture the relationships and dependencies between two inputs. Motivated by the works [23,54], we introduce a co-attention learning block to enhance cross-modality interaction, facilitating information exchange and leveraging the complementary features and contextual cues present in the multimodal inputs [14]. Each co-attention layer incorporates a self-attention block for each modality to learn contextually rich representations. Moreover, we employ a guided-attention block to guide the visual input based on the text input, enabling effective identification and focus on target image regions. The MHA block [44] described in Section 3.1.2

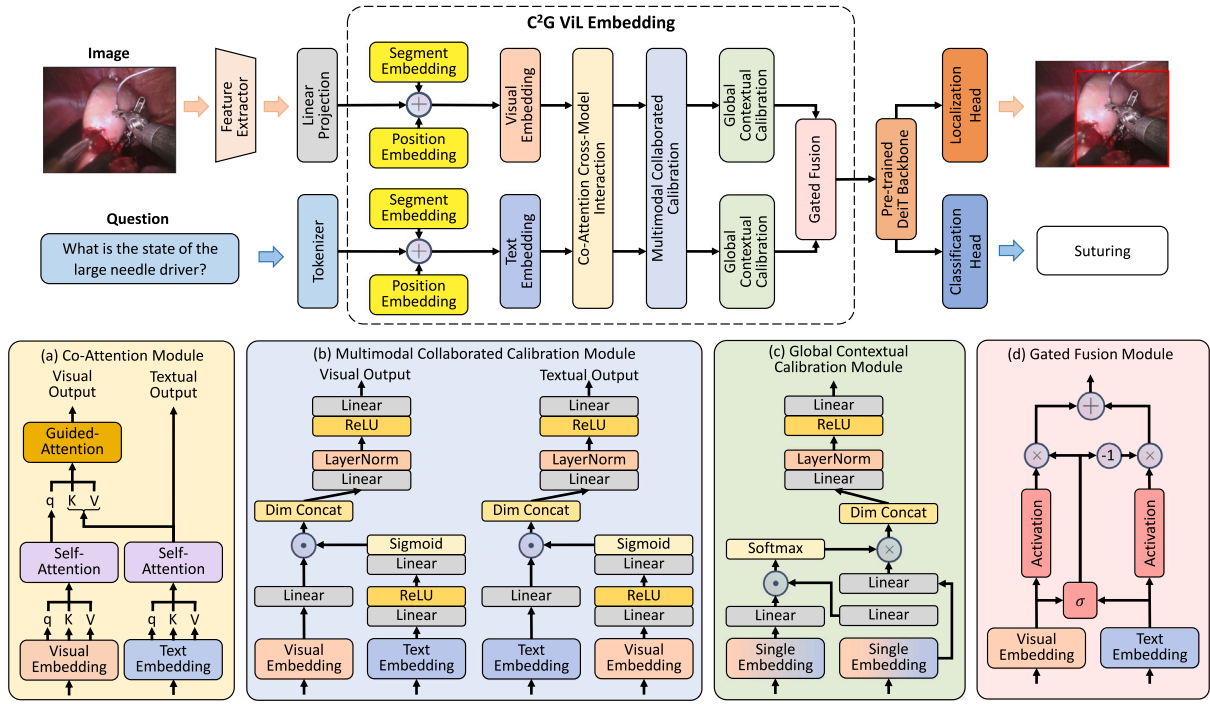


Fig. 3. The overall network architecture of our Surgical-VQLA++ framework. The network comprises a visual feature extractor, a customizedly trained tokenizer, C²G-ViL embedding module (embedding setup, co-attention cross-model interaction, multimodal collaborated calibration, global contextual calibration, gated fusion), pre-trained DeiT backbone, and the parallel prediction heads for question answering and answer localization. \odot , \otimes , and \oplus represent the element-wise dot product, cross product, and summation operation, respectively. Dim Concat denotes dimensional concatenation.

is a typical example. The guided-attention and self-attention blocks share the same architecture, although their inputs differ. Specifically, to allow the visual input to attend to relevant information from the text input, the guided-attention block takes queries from the visual input and keys and values from the text input:

$$h_i = \text{Attn}(\mathbf{W}_{Q_i} \mathbf{Q}_i^{\text{Visual}}, \mathbf{W}_{K_i} \mathbf{K}_i^{\text{Text}}, \mathbf{W}_{V_i} \mathbf{V}_i^{\text{Text}}) \quad (5)$$

The network structure of the co-attention module is depicted in Fig. 3(a). Through this mutual interaction, the model acquires a comprehensive understanding of the instructive interplay from textual input to visual input, resulting in improved performance in joint reasoning. Empirically, we employ six co-attention layers to establish the interactive relationship between the two modalities. The co-attention mechanism effectively enables the model to harness the synergistic effects between the inputs, enhancing the overall representation and facilitating more accurate and informed decision-making.

3.3.2. Multimodal collaborated calibration

The multimodal collaborated calibration consists of two identical modules, in which one calibrates the visual embedding using text embedding, and the other calibrates the text embedding using visual embedding, as shown in Fig. 3(b). This module aims to align the representation from different modalities by establishing a reliable correspondence between them.

To calibrate the visual embedding based on the text embedding, as depicted in Fig. 3(b), the two input embeddings are first projected into N calibration heads by the first two linear layers, where, each head corresponds to embedding dimension $\frac{D}{N}$. The text embedding is then propagated through a linear layer with Sigmoid activation function. This operation shall effectively align the text embedding with the visual embedding. The output representation can be adjusted according to the attention weights. Then, the element-wise product is conducted between the visual and text embedding, before the concatenation of multiple calibration heads. The subsequent fully connected layer, LayerNorm, and ReLU activation will further shape the output of the calibrated embedding.

Calibrating text embedding using visual inputs follows the same steps, with the only difference being the swap of the two modalities. Therefore, we can utilize the mutual statistical correlation to quantify the relationship between patterns. The resulting calibrated embedding shall capture the alignment and refinement information from both visual and text modalities.

3.3.3. Global contextual calibration

Following the calibration of the joint embedding, the module depicted in Fig. 3(c) aims to refine and enhance the global contextual semantics within each modality. To achieve this, we utilize the pairwise bilinear pooling technique to refine and explore the contextual information [55]. The global contextual calibration module comprises two structurally identical components, with each component operating on the visual and textual embedding, respectively. We denote the input embedding from the same modality as E_n and E_m . They are directly divided into N calibration heads with embedding dimension $\frac{D}{N}$ in each head.

$$E_m = \left\{ E_{t,i} \in \mathbb{R}^{\frac{D}{N}} \right\}_{i=1}^N$$

$$E_n = \left\{ E_{v,i} \in \mathbb{R}^{\frac{D}{N}} \right\}_{i=1}^N \quad (6)$$

The formulation of the bilinear pooling can be expressed as follows:

$$E_{gcc} = \text{Softmax} \left[\theta^T L_1(E_n) \odot L_2(E_m) \right] \times L_3(E_m) \quad (7)$$

in which θ is the learnable weight vector, and L denotes the linear layer. $L_1, L_2 \in \mathbb{R}^{\frac{D}{N} \times \frac{D}{N}}$, and $L_3 \in \mathbb{R}^{\frac{D}{N} \times D}$. Then the projected embedding from different heads will be concatenated to obtain an updated embedding E'_{gcc} . The final output embedding shall be shaped through:

$$E''_{gcc} = E_n + L_5 \left[\text{ReLU} \left[\text{LN} \left(L_4(E'_{gcc}) \right) \right] \right] \quad (8)$$

in which $L_4, L_5 \in \mathbb{R}^{D \times D}$, and LN denotes LayerNorm. Consequently, we perform context calibration on each modality separately based on the statistical information, improving global reasoning capabilities.

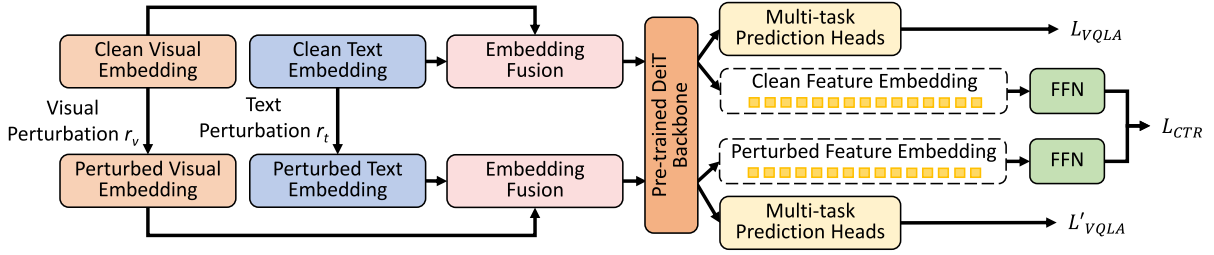


Fig. 4. Overview of the adversarial contrastive training strategy. The visual and text embedding are perturbed separately. After being propagated through the embedding fusion and pre-trained DeiT backbone, the contrastive loss is applied to the clean and perturbed feature embedding.

3.3.4. Gated fusion

A mere concatenation or linear combination of multiple modality inputs lacks the ability to effectively facilitate the learning of cross-modality representations [56]. To address this limitation, we have introduced a gating mechanism to regulate the inputs from different modalities following our preliminary conference version [3]. The specific node δ dynamically adjusts the combination weights for the visual and textual embedding inputs in the subsequent step, allowing the gating module to learn the fusion representation and weights directly from the training data. The \tanh activation function is employed to encode each modality individually. Let R_v and R_t denote the visual and text inputs, respectively. The formulation of the gated module can be expressed as follows:

$$I = k * \tanh(\omega_v \cdot I_v) + (1 - k) * \tanh(\omega_t \cdot I_t) \quad (9)$$

$$k = \delta(\omega \cdot [I_v \parallel I_t])$$

in which $\omega, \omega_v, \omega_t$ are learnable parameters. As presented in Fig. 3(d), this differential solution effectively couples the multimodal embedding inputs, aiming to find a better intermediate representation for the combination of the two modalities.

3.4. Contrastive training with adversarial examples

Despite their intent to deceive machine learning models, adversarial examples can paradoxically enhance the robustness of the model through the adversarial training process. Exposing the model to adversarial examples during training can force it to learn more generalized and robust representations resilient to small perturbations in the input space. Therefore, as presented in Fig. 4, we utilize the adversarial contrastive training strategy to augment the learning process, which encourages the model to consider a broader range of potential inputs and variations, thereby enhancing the robustness and generalization capabilities.

Specifically, we compute the gradients of text and visual embedding separately, and apply the adversarial perturbation described in Eq. (4). r_t and r_v denote the perturbations applied on text and visual embedding, respectively:

$$r_t = -\alpha \cdot \epsilon \cdot \text{sign}(\nabla_{x_t} \mathcal{L}(F(x_t), G)) \quad (10)$$

$$r_v = -\beta \cdot \epsilon \cdot \text{sign}(\nabla_{x_v} \mathcal{L}(F(x_v), G))$$

in which α and β represent the perturbation weight on each modality. Subsequently, after the embedding fusion, the clean and perturbed embedding shall be propagated through the backbone network for deep feature learning. Following [14], we adopt the DeiT-base [17] pre-trained on ImageNet [50] as the backbone.

To facilitate efficient representation learning, we incorporate a regularization approach based on contrastive learning, leveraging both clean and perturbed embedding examples. This regularization technique aims to promote the learning of robust representations that encompass both intraclass similarities and interclass differences. The clean and perturbed example pairs are propagated through both the pre-trained DeiT backbone and a feed-forward network (FFN) to obtain

the clean feature matrix P_i and perturbed feature matrix P_j^r . The cosine similarity is employed to get the distance between P_i and P_j^r :

$$\cos(P_i, P_j^r) = \frac{(P_i)(P_j^r)^T}{(P_i)^2 \cdot (P_j^r)^2} \quad (11)$$

According to [57], given a batch of B clean examples, we generate B adversarial examples, resulting in a total of $2B$ examples. The contrastive loss is thereby established as follows:

$$\mathcal{L}_{CTR}(x; r) = -\log \frac{\exp(\cos(P_i, P_j^r)/C)}{\sum_{\gamma=1}^{2B} \exp(\cos(P_i, P_j^r)/C)} \quad (12)$$

in which C represents the regularization temperature. $\|\cdot\|_{0|\gamma=i}$ means the result is 0 when $\gamma = i$, otherwise it remains unchanged. By incorporating both clean feature examples and adversarial examples within a contrastive learning framework, the model is able to achieve strong generalization capabilities that extend beyond the training data. This integration enables the development of robust and disentangled feature representations that effectively capture the intrinsic characteristics of the data. Furthermore, it enhances the model's robustness and resilience to various sources of variation.

3.5. Prediction heads and loss functions

Following our preliminary conference version [3], the parallel prediction heads comprise a linear layer for classification and a 3-layer FFN for bounding box regression. Besides, our preliminary version [3] employed the CE loss as our QA loss, and the summation of \mathcal{L}_1 -norm and GIoU loss [58] as our localization loss, as follows:

$$\mathcal{L}'_{VQLA}(x; G) = \mathcal{L}_{CE} + (\mathcal{L}_{GIoU} + \mathcal{L}_1) \quad (13)$$

However, in the context of question-answering (QA), there exists a significant challenge of class imbalance that prevents effective convergence when using the cross-entropy (CE) loss. To address this issue, we apply the focal loss [59] to tackle the imbalanced QA prediction. In the case of localization, Carion et al. [60] previously utilized a combination of \mathcal{L}_1 -norm and GIoU loss for object detection. However, our extensive comparative experiments have shown that the use of only the GIoU loss can achieve even better performance.

Moreover, to effectively balance the weights between the two losses and address the trade-off inherent in multi-task learning, it is crucial to explore and determine the optimal weight parameters. Rather than performing this exploration manually, we leverage the uncertainty loss [61] for multi-task training. By treating dropout as a Bayesian approximation, the uncertainty loss shall sample multiple dropout masks. This process leads to multiple predictions for a given input, enabling the estimation of task uncertainties. These uncertainties are then used to weigh the losses during training. Tasks with lower uncertainties shall be assigned higher weights and vice versa. Therefore, the model can dynamically allocate weights and resources to the localization and question-answering tasks during training. The resulting VQLA loss can be formulated as follows:

$$\mathcal{L}_{VQLA}(x; G) = \frac{1}{2\sigma_1^2} \mathcal{L}_{Focal} + \log \sigma_1 + \frac{1}{2\sigma_2^2} \mathcal{L}_{GIoU} + \log \sigma_2 \quad (14)$$

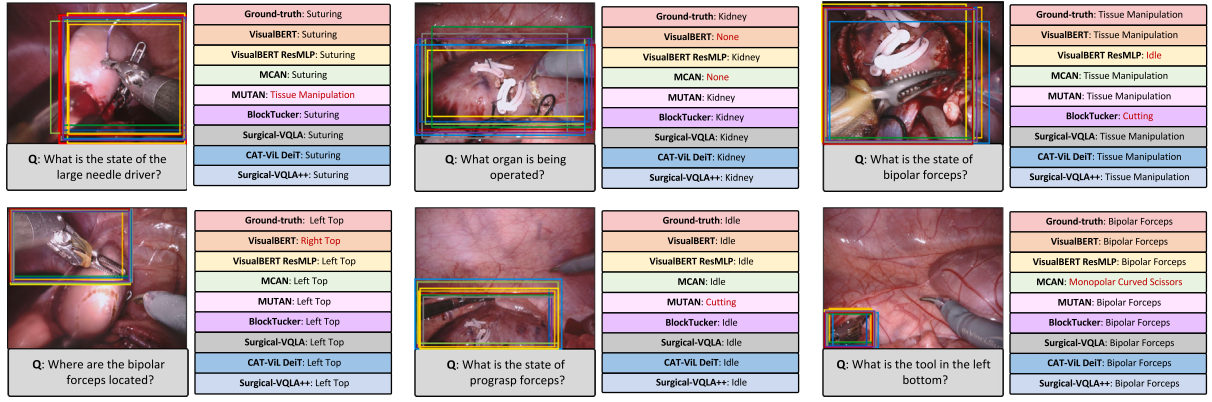


Fig. 5. Qualitative comparison of our Surgical-VQLA++ against SOTA solutions. The color of the detection bounding box corresponds to the color of the answers, in which red represents the ground truth and light blue represents our proposed solution. On the four types of question-answer pairs (tissue, action, location, instrument), our method achieves the best performance on both answering and localization tasks.

Table 1

Comparison experiments between the preliminary version Surgical-VQLA [3] and the SOTA version Surgical-VQLA++.

| Models | Localization | Classification | Uncertainty | EndoVis-18-VQLA | | | EndoVis-17-VQLA | | |
|-------------------|-------------------------------|----------------|-------------|-----------------|---------------|---------------|-----------------|---------------|---------------|
| | | | | Acc | F-Score | mIoU | Acc | F-Score | mIoU |
| Surgical-VQLA [3] | $\mathcal{L}_1 + \text{GIoU}$ | CE | × | 0.6206 | 0.2742 | 0.7731 | 0.4562 | 0.3317 | 0.7652 |
| Surgical-VQLA++ | GIoU | Focal | ✓ | 0.6568 | 0.3180 | 0.7982 | 0.5071 | 0.4476 | 0.7758 |

where the scale value σ_1 and σ_2 are learnable parameters. Therefore, our final loss is the summation of VQLA loss on clean examples (\mathcal{L}_{VQLA}), VQLA loss on perturbed examples (\mathcal{L}'_{VQLA}), and the contrastive loss (\mathcal{L}_{CTR}):

$$\mathcal{L} = \mathcal{L}_{VQLA}(x; G) + \mathcal{L}_{VQLA}(x + r; G) + \mathcal{L}_{CTR}(x; r) \quad (15)$$

4. Experiments

4.1. Datasets

EndoVis-18-VQLA is a publicly available robotic surgery dataset derived from the MICCAI Challenge 2018 [18]. The VQLA labels for this dataset are accessible through [3]. Initially, the annotations only included surgical actions, target tissues, and instrument positions. To enhance the dataset's utility, we have expanded it by adding questions specifically targeting instrument names. This expansion aims to improve the overall usefulness and diversity of the dataset. The training comprises 1560 frames, and the test set has 436 frames. The number of QA pairs in the training set are expanded from 9014 to 12741, the test set is extended from 2769 to 3820 QA pairs. The dataset splitting follows the guidelines (Training Video ID: 2, 3, 4, 6, 7, 9, 10, 11, 12, 14, 15; Test Video ID: 1, 5, 16) in [1] to ensure proper isolation between video sequences and prevent any potential data leakage.

EndoVis-17-VQLA, derived from the MICCAI Challenge 2017 [19], is used for external validation. It is important to note that we *train* our model solely on the EndoVis-18 training set and then *test* its performance on the EndoVis-18 test set and EndoVis-17 dataset. The annotations for this dataset are also provided by [3], and we apply the same dataset expansion as mentioned above. With 97 frames, this dataset is extended from 472 to 708 QA pairs. The purpose of using this dataset is to showcase the generalizability of our model when facing domain gaps.

Table 2

Comparison experiments of our Surgical-VQLA++ against SOTA solutions. We implement all the model on the extended datasets, so the reported results will be different from the results reported in their original publications. GF: Gated Fusion, CA: Co-Attention layer, MCC: Multimodal Collaborated Calibration, GCC: Global Contextual Calibration, C²G: C²G-ViL Embedding.

| Models | EndoVis-18-VQLA | | | EndoVis-17-VQLA | | |
|--------------------------------------|-----------------|---------------|---------------|-----------------|---------------|---------------|
| | Acc | F-Score | mIoU | Acc | F-Score | mIoU |
| VisualBERT [4] | 0.6203 | 0.3061 | 0.7433 | 0.4647 | 0.2471 | 0.7336 |
| VisualBERT ResMLP [1] | 0.6433 | 0.3218 | 0.7576 | 0.4633 | 0.2674 | 0.7387 |
| MCAN [23] | 0.6366 | 0.3191 | 0.7827 | 0.4506 | 0.2640 | 0.7647 |
| VQLA-DeiT [17] | 0.6183 | 0.2827 | 0.7343 | 0.4379 | 0.2324 | 0.7224 |
| MUTAN [62] | 0.6343 | 0.2930 | 0.7764 | 0.4817 | 0.2900 | 0.7690 |
| MFH [63] | 0.6241 | 0.3149 | 0.7679 | 0.4774 | 0.3418 | 0.7499 |
| BlockTucker [64] | 0.6391 | 0.3077 | 0.7693 | 0.4520 | 0.3059 | 0.7640 |
| CAT-ViL DeiT [14] | 0.6290 | 0.2807 | 0.7776 | 0.4887 | 0.2812 | 0.7657 |
| Surgical-VQLA [3] | 0.6456 | 0.3219 | 0.7212 | 0.4322 | 0.2416 | 0.7321 |
| Surgical-VQLA++ w/o C ² G | 0.6249 | 0.3069 | 0.7578 | 0.4548 | 0.3325 | 0.7441 |
| Surgical-VQLA++ w/o CA | 0.6398 | 0.2998 | 0.7808 | 0.4407 | 0.3887 | 0.7569 |
| Surgical-VQLA++ w/o GF | 0.6349 | 0.2814 | 0.7742 | 0.4581 | 0.3464 | 0.7677 |
| Surgical-VQLA++ w/o MCC | 0.6362 | 0.3187 | 0.7740 | 0.4909 | 0.3401 | 0.7675 |
| Surgical-VQLA++ w/o GCC | 0.6314 | 0.3210 | 0.7845 | 0.4477 | 0.3282 | 0.7655 |
| Surgical-VQLA++ | 0.6568 | 0.3180 | 0.7982 | 0.5071 | 0.4476 | 0.7758 |

4.2. Implementation details

Our proposed solution is compared against the following models: VisualBERT² [4], VisualBERT ResMLP³ [1], VQLA-DeiT⁴ [17], MCAN⁵ [23], MUTAN⁶ [62], MFH⁷ [63], BlockTucker⁶ [64], CAT-ViL DeiT⁸ [14], and our earlier version Surgical-VQLA⁹ [3]. We employ the same loss \mathcal{L}_{VQLA} for all models to keep a fair comparison. Following

² github.com/uclanlp/visualbert

³ github.com/lalithjets/surgical_vqa

⁴ github.com/rwightman/pytorch-image-models

⁵ github.com/MILVLG/mcan-vqa

⁶ github.com/Cadene/block.bootstrap.pytorch

⁷ github.com/MILVLG/openvqa

⁸ github.com/longbai1006/CAT-ViL

⁹ github.com/longbai1006/Surgical-VQLA

Table 3

Comparison of various test samples on Surgical-VQLA, CAT-ViL DeiT, and our Surgical-VQLA++. The four types of samples include the tissue being operated on, the current state of the instruments, the instrument location in the operating area, and the instrument identification. Please note that there are no QA pairs related to tissues in the EndoVis-17-VQLA dataset.

| Model | Data type | EndoVis-18-VQLA | | | EndoVis-17-VQLA | | |
|-------------------|------------|-----------------|---------------|---------------------------------------|-----------------|---------------|---------------------------------------|
| | | Acc | F-Score | mIoU | Acc | F-Score | mIoU |
| CAT-ViL DeiT [14] | Tissue | 1.0000 | 1.0000 | 0.6910 \pm 0.1558 | / | / | / |
| Surgical-VQLA [3] | | 1.0000 | 1.0000 | 0.7253 \pm 0.1178 | / | / | / |
| Surgical-VQLA++ | | 1.0000 | 1.0000 | 0.8124 \pm 0.1234 | / | / | / |
| CAT-ViL DeiT [14] | Action | 0.5426 | 0.1646 | 0.8279 \pm 0.0919 | 0.3305 | 0.1898 | 0.7656 \pm 0.0988 |
| Surgical-VQLA [3] | | 0.5049 | 0.1676 | 0.8174 \pm 0.0999 | 0.3051 | 0.2882 | 0.7479 \pm 0.0701 |
| Surgical-VQLA++ | | 0.5771 | 0.2022 | 0.8655 \pm 0.0859 | 0.2966 | 0.3336 | 0.7972 \pm 0.0967 |
| CAT-ViL DeiT [14] | Location | 0.5585 | 0.4802 | 0.7491 \pm 0.2122 | 0.5297 | 0.4988 | 0.7182 \pm 0.2521 |
| Surgical-VQLA [3] | | 0.6117 | 0.5616 | 0.6411 \pm 0.1908 | 0.5424 | 0.5269 | 0.6450 \pm 0.1787 |
| Surgical-VQLA++ | | 0.6064 | 0.4678 | 0.7515 \pm 0.2145 | 0.6441 | 0.6181 | 0.7187 \pm 0.2535 |
| CAT-ViL DeiT [14] | Instrument | 0.6791 | 0.2266 | 0.7911 \pm 0.1421 | 0.5763 | 0.4664 | 0.8176 \pm 0.1369 |
| Surgical-VQLA [3] | | 0.6722 | 0.2067 | 0.7039 \pm 0.1358 | 0.5593 | 0.4002 | 0.7166 \pm 0.1290 |
| Surgical-VQLA++ | | 0.6543 | 0.2194 | 0.7820 \pm 0.1441 | 0.5805 | 0.4173 | 0.8116 \pm 0.1399 |

their official repositories, we implement all the model on the extended datasets, so the reported results will be different from the results reported in their original publications.

All models are trained using the Python PyTorch framework on NVIDIA RTX 3090 GPU. Adam optimizer [65] is employed for 80 epochs with a batch size of 64 and a learning rate of 1×10^{-5} .

4.3. Results

We first compare our proposed method with our preliminary version, Surgical-VQLA [3]. The results of this comparison are presented in Table 1 and Fig. 5. To ensure a fair evaluation, we re-implement the Surgical-VQLA model using our extended dataset. Through the introduction of our Surgical-VQLA++ framework, we observe improvements in the metrics for both datasets, ranging from 1.06% to 11.59%. These enhancements affirm the significant improvement our work has made in enhancing the VQLA performance of the model.

Next, we perform a quantitative evaluation against SOTA methods, as shown in Table 2. In order to maintain fairness, all models utilize uncertainty loss, employing focal loss for QA prediction and GIoU for the localization task. Our proposed methodology demonstrates exceptional performance across accuracy and mIoU metrics, surpassing all baseline models. This substantiates the efficacy and superiority of our approach. Notably, the performance of Surgical-VQLA exhibits improvement compared to the results presented in Table 1, which suggests the broad applicability of the adopted loss function strategy.

Moreover, we assess the effects of removing the (i) co-attention layers, (ii) the gated fusion module, (iii) multimodal collaborated calibration, (iv) global contextual calibration, and (v) C²G-ViL embedding, respectively. Experimental results in Table 2 conducted on the EndoVis-18-VQLA dataset reveal that the absence of any of these modules leads to a significant degradation in model performance. Similarly, in the external validation set, EndoVis-17-VQLA, both accuracy and mIoU decrease significantly. These observations indicate that all three proposed modules positively contribute to achieving the best results.

Furthermore, we provide a detailed evaluation of the model's performance across different types of test samples, as shown in Table 3. These samples are divided based on the types of questions, including tissue recognition, action state, instrument location, and instrument identification. Firstly, all models achieve full scores in tissue recognition as this task setup is not complicated. Secondly, our model achieves the best performance on action state recognition and instrument location questions, while it performs unsatisfactorily in instrument identification. However, although the model does not perform the best in instrument-related questions during text dialogue, its localization module maintains a high level of bounding box detection performance. These findings underscore the effectiveness of our framework in handling diverse surgical scenarios and questions, thus enhancing its applicability and reliability in real-world settings.

4.4. Ablation study

Effects of the visual feature extractors. In Table 4, a comparative analysis is presented, evaluating the model's performance before and after the incorporation of object proposals using global feature extractor ResNet18 [49]. Apart from the significant improvement in inference speed brought by the end-to-end architecture, our model also demonstrates superior quantitative performance. This can be attributed to the global feature extractor's ability to avoid directing attention to incorrect regions as a result of erroneous detections.

Effects of loss function combinations. In Table 5, we present the analysis of various combinations of loss functions and their impact on the performance of our proposed model. When compared to using GIoU alone, the combination of \mathcal{L}_1 and GIoU does not consistently improve the evaluation metrics, indicating that it is not a superior strategy. Conversely, the integration of focal loss exhibits notable superiority over CE Loss. Additionally, the incorporation of uncertainty loss [61] further enhances the model's performance, especially when combined with GIoU and focal loss. Furthermore, we investigate several advanced IoU functions in Table 6 using the aforementioned combination strategies. The results reveal that, compared with DIoU and CIoU [66], GIoU loss is the most suitable choice for our multitask learning framework.

Effects of different fusion strategies. In Table 7, we investigate the impact of different ViL embedding while keeping the DeiT backbone and feature extractors consistent. The comparisons include Concatenation [4], JCA [67], MMHCA [68], MAT [69], GMU [56], Self-Attention [44], Guided-Attention [23], and Co-Attention (T2V: Use text to guide vision [23]; V2T: Use vision to guide text; Bi: Bidirectional guidance). Additionally, we explore our Surgical-VQLA++ solution using the attention mechanisms mentioned above for a comprehensive comparison. The study demonstrates that our ViL embedding strategy outperforms other sophisticated methods. Moreover, in comparison to other forms of attention, utilizing text information to guide visual feature embedding significantly helps the model focus on crucial regions within an image, resulting in outstanding performance in VQLA tasks.

Effects of the number of co-attention layers. In Table 8, we evaluate the impact of different co-attention configurations on our model's performance. After careful analysis, we identify layer number = 6 as the most suitable co-attention configuration for our final model. Deviations from this optimal value result in a slight decrease in performance, with a significant degradation observed when the layer number exceeds 8.

Effects of the weights in adversarial perturbation. In Table 9, we systematically investigate the influence of weight parameters α and β in contrastive training on the overall performance of our model. Through a grid search, we determine that setting $\alpha = 1$ and $\beta = 0.5$ yields the

Table 4

Ablation study on different visual feature extractors.

| Model | Visual feature extractor | | | EndoVis-18-VQLA | | | EndoVis-17-VQLA | | |
|-------|--------------------------|--------------------|-----------------|-----------------|---------------|---------------|-----------------|---------------|---------------|
| | Detection | Feature extraction | Inference Speed | Acc | F-Score | mIoU | Acc | F-Score | mIoU |
| Ours | × | ResNet [49] | 150.6 FPS | 0.6568 | 0.3180 | 0.7982 | 0.5071 | 0.4476 | 0.7758 |
| | FRCNN [45] | ResNet [49] | 18.09 FPS | 0.6340 | 0.3720 | 0.7218 | 0.4251 | 0.2946 | 0.7050 |

Table 5Ablation study on different loss function combinations. For QA prediction, we use the CE loss and focal loss [59]. For the localization task, we explore the effects of GloU [58] and \mathcal{L}_1 + GloU [60]. The integration of uncertainty loss [61] is also introduced.

| Localization | Classification | Uncertainty | EndoVis-18-VQLA | | | EndoVis-17-VQLA | | |
|-----------------------|----------------|-------------|-----------------|---------------|---------------|-----------------|---------------|---------------|
| | | | Acc | F-Score | mIoU | Acc | F-Score | mIoU |
| GIoU | CE | × | 0.6199 | 0.3048 | 0.7613 | 0.4929 | 0.2839 | 0.7519 |
| \mathcal{L}_1 +GIoU | CE | × | 0.6295 | 0.2939 | 0.7737 | 0.5113 | 0.2976 | 0.7570 |
| GIoU | Focal | × | 0.6433 | 0.3793 | 0.7885 | 0.4379 | 0.4396 | 0.7709 |
| \mathcal{L}_1 +GIoU | Focal | × | 0.6250 | 0.3610 | 0.7725 | 0.5042 | 0.3865 | 0.7394 |
| GIoU | CE | ✓ | 0.6267 | 0.2917 | 0.7620 | 0.5170 | 0.3709 | 0.7625 |
| \mathcal{L}_1 +GIoU | CE | ✓ | 0.6211 | 0.2947 | 0.7795 | 0.5127 | 0.3563 | 0.7691 |
| GIoU | Focal | ✓ | 0.6568 | 0.3180 | 0.7982 | 0.5071 | 0.4476 | 0.7758 |
| \mathcal{L}_1 +GIoU | Focal | ✓ | 0.6371 | 0.2998 | 0.7712 | 0.5071 | 0.3078 | 0.7530 |

Table 6

Ablation study on different IoU loss functions.

| Localization | Classification | Uncertainty | EndoVis-18-VQLA | | | EndoVis-17-VQLA | | |
|------------------|----------------|-------------|-----------------|---------------|---------------|-----------------|---------------|---------------|
| | | | Acc | F-Score | mIoU | Acc | F-Score | mIoU |
| IoU | Focal | ✓ | 0.6122 | 0.3634 | 0.4097 | 0.4361 | 0.4408 | 0.4104 |
| CIoU [66] | Focal | ✓ | 0.6255 | 0.3236 | 0.7859 | 0.4590 | 0.4486 | 0.7640 |
| DIoU [66] | Focal | ✓ | 0.6333 | 0.3271 | 0.7561 | 0.4618 | 0.2572 | 0.7552 |
| GIoU [58] | Focal | ✓ | 0.6568 | 0.3180 | 0.7982 | 0.5071 | 0.4476 | 0.7758 |

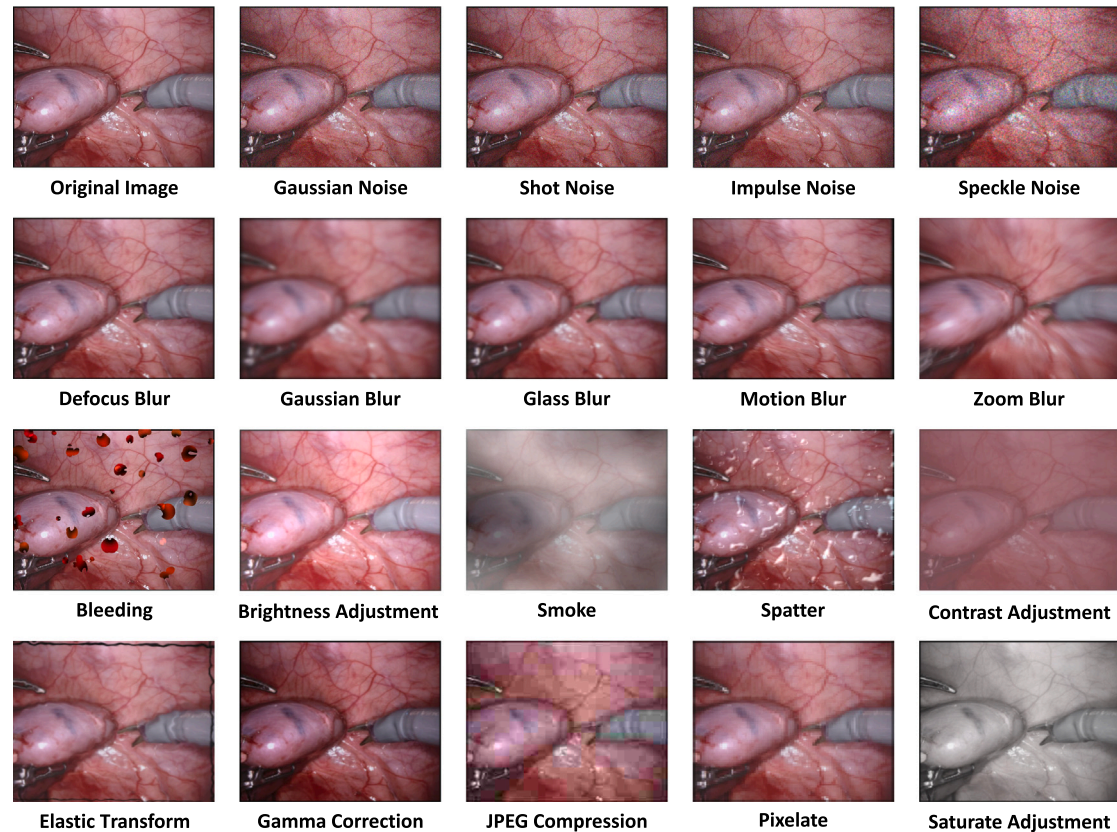
**Fig. 6.** Corrupted data visualization for the robustness assessment. We simulate four different types of image corruption, including noise, blur, occlusion, and digital damage.

Table 7

Ablation study on different fusion strategies. All experiments employ the same feature extractor, DeiT backbone, and prediction heads. “Attn”: Attention.

| Models | EndoVis-18-VQLA | | | EndoVis-17-VQLA | | |
|-------------------------|-----------------|---------------|---------------|-----------------|---------------|---------------|
| | Acc | F-Score | mIoU | Acc | F-Score | mIoU |
| Concatenation [4] | 0.6183 | 0.2827 | 0.7343 | 0.4379 | 0.2324 | 0.7224 |
| JCA [67] | 0.6221 | 0.2895 | 0.7542 | 0.4463 | 0.2558 | 0.7465 |
| MMHCA [68] | 0.6295 | 0.2798 | 0.7700 | 0.4364 | 0.3006 | 0.7579 |
| MAT [69] | 0.6303 | 0.2872 | 0.7632 | 0.4393 | 0.2942 | 0.7547 |
| GMU [56] | 0.6361 | 0.3490 | 0.7196 | 0.4167 | 0.2360 | 0.7326 |
| Self-Attn [44] | 0.6262 | 0.2810 | 0.7714 | 0.4449 | 0.3030 | 0.7548 |
| Guided-Attn [23] | 0.6295 | 0.2983 | 0.7766 | 0.4633 | 0.3018 | 0.7582 |
| Co-Attn (Bi) | 0.6262 | 0.2915 | 0.7661 | 0.4463 | 0.3033 | 0.7482 |
| Co-Attn (V2T) | 0.6308 | 0.3037 | 0.7510 | 0.4040 | 0.2201 | 0.7341 |
| Co-Attn (T2V) | 0.6366 | 0.2858 | 0.7791 | 0.4407 | 0.2569 | 0.7563 |
| Ours with Self-Attn | 0.6484 | 0.3447 | 0.7749 | 0.4576 | 0.2545 | 0.7543 |
| Ours with Guided-Attn | 0.6403 | 0.2817 | 0.7886 | 0.4845 | 0.3183 | 0.7609 |
| Ours with Co-Attn (Bi) | 0.6251 | 0.3110 | 0.7679 | 0.4393 | 0.2768 | 0.7498 |
| Ours with Co-Attn (V2T) | 0.6233 | 0.3602 | 0.7609 | 0.4703 | 0.4666 | 0.7517 |
| Ours with Co-Attn (T2V) | 0.6568 | 0.3180 | 0.7982 | 0.5071 | 0.4476 | 0.7758 |

Table 8

Ablation study on the number of co-attention layers.

| Co-attention layer | EndoVis-18-VQLA | | | EndoVis-17-VQLA | | |
|--------------------|-----------------|---------------|---------------|-----------------|---------------|---------------|
| | Acc | F-Score | mIoU | Acc | F-Score | mIoU |
| 2 | 0.6211 | 0.2754 | 0.7829 | 0.4520 | 0.3285 | 0.7706 |
| 4 | 0.6260 | 0.3301 | 0.7829 | 0.4379 | 0.3554 | 0.7685 |
| 6 | 0.6568 | 0.3180 | 0.7982 | 0.5071 | 0.4476 | 0.7758 |
| 8 | 0.6252 | 0.3683 | 0.7772 | 0.4025 | 0.4555 | 0.7664 |
| 10 | 0.5641 | 0.2126 | 0.7049 | 0.4633 | 0.2487 | 0.7167 |

Table 9

Ablation on the weight parameters of the adversarial perturbations. α and β denote the perturbation weight on text and visual embedding. The bolded weights are the parameters we use in all the above experiments. “w/o CTAS” denotes our model without contrastive training using adversarial samples.

| α | β | EndoVis-18-VQLA | | | EndoVis-17-VQLA | | |
|----------|---------|-----------------|---------------|---------------|-----------------|---------------|---------------|
| | | Acc | F-Score | mIoU | Acc | F-Score | mIoU |
| w/o CTAS | | 0.6419 | 0.3378 | 0.7787 | 0.4689 | 0.3658 | 0.7705 |
| 0.5 | 1 | 0.6398 | 0.3243 | 0.7872 | 0.4562 | 0.4366 | 0.7716 |
| 0.5 | 0.5 | 0.6201 | 0.3118 | 0.7881 | 0.4548 | 0.4185 | 0.7675 |
| 0.5 | 0.1 | 0.6416 | 0.3376 | 0.7892 | 0.4802 | 0.3043 | 0.7592 |
| 1 | 1 | 0.6359 | 0.2938 | 0.7675 | 0.4884 | 0.4373 | 0.7599 |
| 1 | 0.5 | 0.6568 | 0.3180 | 0.7982 | 0.5071 | 0.4476 | 0.7758 |
| 1 | 0.1 | 0.6529 | 0.3066 | 0.7838 | 0.4703 | 0.3994 | 0.7621 |
| 2 | 1 | 0.6584 | 0.3566 | 0.7786 | 0.4590 | 0.4025 | 0.7477 |
| 2 | 0.5 | 0.6232 | 0.3599 | 0.7829 | 0.4619 | 0.3999 | 0.7711 |
| 2 | 0.1 | 0.6317 | 0.3560 | 0.7827 | 0.4958 | 0.3960 | 0.7682 |

best performance. These weight parameters are crucial for achieving optimal results in our proposed framework. Besides, as illustrated in the first row of Table 9, we report the results obtained when the contrastive training strategy is excluded. The performance exhibits a significant degradation, providing further evidence to support the effectiveness of our proposed training strategy.

4.5. Robustness against image corruption

We conducted a comprehensive robustness assessment to evaluate how effectively our model handles challenging image corruption. We used a benchmark consisting of 19 different types of image corruption, with 18 sourced from a widely recognized benchmark [70] and 1 simulating blood occlusion on the camera lens during surgery [71]. The visualization of these nineteen image corruption techniques is shown in Fig. 6. All corruption methods are categorized into five severity levels to observe the model’s performance under varying degrees of corruption. Below is a brief introduction to these four categories and 19 types of image corruption.

The ‘Noise’ category represents various types of noise that may occur during image capture, storage, and transmission. For instance, low-light conditions typically cause Gaussian noise. Impulse noise, a color variation of salt-and-pepper noise, can be generated by bit errors. Shot noise is a type of electronic noise. Speckle noise is an additive noise that varies based on the intensity of the original pixel.

The ‘Blur’ category represents image blur that can occur due to defocusing or motion during capture and transmission. Defocus blur appears when the image is out of focus. Gaussian Blur occurs when surgical images are excessively de-noised. Glass Blur usually results from optical distortions or errors in compression/ transmission. Motion blur occurs when the camera moves very fast. Zoom blur appears when a camera rapidly moves towards an object.

The ‘Occlusion’ category represents reductions in image quality due to various objective conditions in real surgical scenes, such as abnormal lighting and obstructions. Bleeding mimics blood occlusion on the lens in surgical scenes. Brightness varies with surgical light intensity. Smoke, commonly seen during cauterization in surgery, shrouds the object. Spatter simulates water occluding the camera.

The ‘Digital’ category represents digital damage that can occur during image collection and transmission due to hardware issues. Contrast is influenced by the color of the tissues and lighting conditions. Elastic transform can stretch or contract the surgical area. Gamma Correction adjusts the brightness and contrast of the surgical scene through a nonlinear transformation. JPEG is a compression artifact effect produced by the lossy compression of surgical images. Pixelation occurs when upsampling low-resolution surgical images. Saturate Adjustment changes the purity of image colors.

Then, we directly conduct inference on the corrupted test dataset and calculate the average results across the different corruption techniques for each severity level. The outcomes of our comprehensive robustness experiments, categorized by the severity level, are depicted in Fig. 7. A consistent deterioration in performance is observed as the severity level of corruption escalates. Furthermore, Table 10 provides an overview of the average results for each corruption type. Our Surgical-VQLA++ showcases notable performance advantages across various corruption scenarios. It is worth highlighting that our methods consistently achieve higher accuracy and mIoU compared to both the original Surgical-VQLA and CAT-ViL approaches. While our method and CAT-ViL DeiT exhibit comparable performance in the QA task of EndoVis-17-VQLA, our approach still maintains an overall advantage. In this context, the incorporation of feature calibration and adversarial contrastive training effectively enhances the robustness of our model in dealing with image corruption challenges.

4.6. Discussion and limitations

Overall, our approach has demonstrated substantial improvements in various comparative experiments. It effectively enhances the performance of question-answering and answer localization tasks and exhibits exceptional adaptability to domain shift and image corruption challenges. This success can be attributed to our proposed C²G-ViL module, which aligns multimodal representations and globally calibrates contextual features, thereby enhancing the informative interactions between modalities during feature fusion. Additionally, our exploration of optimal fusion weights and contrastive training strategies enables our models to recognize and capture subtle discrepancies or patterns.

However, we have also observed poor F-Score results for our methods compared to other sophisticated approaches, particularly in certain instances of the VQLA task. This indicates that our work may underperform in specific scenarios. Therefore, as a potential avenue for future research, it would be worthwhile to explore more optimal information fusion strategies to improve the precision and recall of our model while balancing all the relevant metrics. Besides, considering the increasing demand for specialized query understanding in surgery and surgical

Table 10

Robustness comparison on Surgical-VQLA, CAT-ViL DeiT, and our Surgical-VQLA++. We compute the average results on each corruption type.

| Dataset | | EndoVis-18-VQLA | | | | | | EndoVis-17-VQLA | | | | | |
|-----------|------------|-----------------------|----------------------|----------------------|-----------------------|----------------------|----------------------|-----------------------|----------------------|----------------------|-----------------------|----------------------|----------------------|
| Metric | | Acc | | | mIoU | | | Acc | | | mIoU | | |
| Model | | Surgical-VQLA [3] | CAT-ViL DeiT [14] | Surgical-VQLA++ | Surgical-VQLA [3] | CAT-ViL DeiT [14] | Surgical-VQLA++ | Surgical-VQLA [3] | CAT-ViL DeiT [14] | Surgical-VQLA++ | Surgical-VQLA [3] | CAT-ViL DeiT [14] | Surgical-VQLA++ |
| Loss | | \mathcal{L}'_{VQLA} | \mathcal{L}_{VQLA} | \mathcal{L}_{VQLA} | \mathcal{L}'_{VQLA} | \mathcal{L}_{VQLA} | \mathcal{L}_{VQLA} | \mathcal{L}'_{VQLA} | \mathcal{L}_{VQLA} | \mathcal{L}_{VQLA} | \mathcal{L}'_{VQLA} | \mathcal{L}_{VQLA} | \mathcal{L}_{VQLA} |
| Noise | Gauss. | 0.5927 | 0.6180 | 0.6396 | 0.7657 | 0.7753 | 0.7929 | 0.4435 | 0.4663 | 0.4814 | 0.7581 | 0.7638 | 0.7775 |
| | Shot | 0.5905 | 0.6213 | 0.6464 | 0.7657 | 0.7760 | 0.7949 | 0.4386 | 0.4711 | 0.4740 | 0.7605 | 0.7640 | 0.7790 |
| | Impulse | 0.6023 | 0.6233 | 0.6407 | 0.7672 | 0.7768 | 0.7951 | 0.4369 | 0.4712 | 0.4749 | 0.7576 | 0.7638 | 0.7802 |
| | Speckle | 0.5719 | 0.5915 | 0.6096 | 0.7398 | 0.7427 | 0.7599 | 0.4404 | 0.4756 | 0.4339 | 0.7300 | 0.7390 | 0.7508 |
| Blur | Defocus | 0.6188 | 0.6299 | 0.6523 | 0.7760 | 0.7787 | 0.7932 | 0.4433 | 0.4686 | 0.4938 | 0.7515 | 0.7588 | 0.7711 |
| | Gauss. | 0.6006 | 0.6157 | 0.6474 | 0.7483 | 0.7471 | 0.7678 | 0.4485 | 0.4720 | 0.4647 | 0.7090 | 0.7281 | 0.7446 |
| | Glass | 0.6236 | 0.6584 | 0.6561 | 0.7763 | 0.7804 | 0.7938 | 0.4565 | 0.4799 | 0.4949 | 0.7607 | 0.7632 | 0.7736 |
| | Motion | 0.6135 | 0.6314 | 0.6567 | 0.7730 | 0.7771 | 0.7904 | 0.4480 | 0.4743 | 0.4972 | 0.7527 | 0.7610 | 0.7729 |
| | Zoom | 0.5880 | 0.6249 | 0.6292 | 0.7293 | 0.7690 | 0.7683 | 0.4362 | 0.4734 | 0.4556 | 0.7330 | 0.7485 | 0.7557 |
| Occlusion | Bleeding | 0.5944 | 0.5978 | 0.6080 | 0.7361 | 0.7600 | 0.7736 | 0.4391 | 0.4468 | 0.4576 | 0.7083 | 0.7336 | 0.7573 |
| | Brightness | 0.6275 | 0.6242 | 0.6444 | 0.7656 | 0.7725 | 0.7921 | 0.4417 | 0.4708 | 0.4975 | 0.7488 | 0.7566 | 0.7728 |
| | Smoke | 0.6333 | 0.6213 | 0.6318 | 0.7597 | 0.7606 | 0.7687 | 0.4302 | 0.4729 | 0.4636 | 0.7239 | 0.7390 | 0.7405 |
| | Spatter | 0.5898 | 0.6018 | 0.6190 | 0.7411 | 0.7536 | 0.7797 | 0.4353 | 0.4756 | 0.4921 | 0.7195 | 0.7395 | 0.7598 |
| Digital | Contrast | 0.6307 | 0.6139 | 0.6255 | 0.7382 | 0.7271 | 0.7677 | 0.4364 | 0.4624 | 0.4469 | 0.6966 | 0.7240 | 0.7314 |
| | Elastic | 0.5959 | 0.6059 | 0.6226 | 0.7303 | 0.7465 | 0.7628 | 0.4404 | 0.4672 | 0.5014 | 0.6906 | 0.7230 | 0.7355 |
| | Gamma | 0.6207 | 0.6342 | 0.6484 | 0.7782 | 0.7798 | 0.7923 | 0.4453 | 0.4683 | 0.4927 | 0.7531 | 0.7599 | 0.7721 |
| | Jpeg | 0.5935 | 0.6078 | 0.6195 | 0.7513 | 0.7565 | 0.7798 | 0.4343 | 0.4956 | 0.4867 | 0.7530 | 0.7527 | 0.7734 |
| | Pixelate | 0.6219 | 0.6311 | 0.6556 | 0.7697 | 0.7747 | 0.7917 | 0.4538 | 0.4699 | 0.4802 | 0.7610 | 0.7600 | 0.7776 |
| | Saturate | 0.6095 | 0.6327 | 0.6437 | 0.7682 | 0.7684 | 0.7821 | 0.4516 | 0.4757 | 0.4621 | 0.7552 | 0.7591 | 0.7699 |
| Mean | | 0.6063 | 0.6203 | 0.6367 | 0.7568 | 0.7644 | 0.7814 | 0.4421 | 0.4715 | 0.4764 | 0.7381 | 0.7494 | 0.7629 |

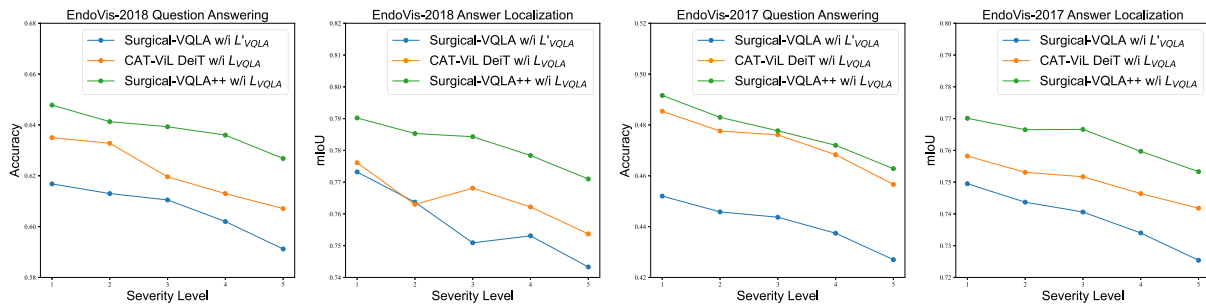


Fig. 7. Robustness comparison on each severity level. We process the data using 19 image corruption approaches with 5 severity levels, and average the predictions on each severity level.

procedures, it is necessary to acknowledge the limitation of our current method, which may not fully meet the accuracy requirements of the VQLA task, especially when facing domain shifts. In addition to developing domain adaptation and continual learning strategies, one potential direction for future work could involve exploring gradient perturbation techniques for multimodal fusion features. These advancements would help better handle domain shifts and image corruption challenges while simultaneously enhancing prediction accuracy.

Furthermore, current Multimodal Large Language Models (MLLMs) have become crucial for general multimodal understanding tasks due to their large parameter counts and extensive training data, which enable them to display advanced capabilities [72,73]. These models can understand and produce complex responses that mimic human-like interactions. However, MLLMs often respond slowly and require high-performance GPUs for training and deployment. For domain-specific issues, MLLMs need targeted fine-tuning based on specific data, and they may also encounter problems with hallucinations and the repetitive output phenomenon [74,75]. In contrast, Surgical VQLA++ is specifically designed to provide accurate and context-aware responses to queries about particular surgical images, which is vital in medical settings where the correctness of information can greatly influence diagnosis and treatment decisions. In surgical environments, decisions must frequently be made swiftly based on the latest visual information. Surgical-VQLA++ supports real-time decision-making in these critical situations and uses adversarial contrastive learning to improve its robustness against image corruptions, making it suitable for medical applications. Moreover, while the surgical report or caption generation task can only provide general textual descriptive output, our VQLA model can deliver specific answers to given questions and also provide corresponding localization results, enhancing both the precision and utility of visual information processing in surgical contexts.

5. Conclusion

In this study, we introduced Surgical-VQLA++, an advanced model that enhances the robustness and VQLA performance in the context of robotic surgery. Our framework integrates C²G-ViL embedding with adversarial contrastive learning, focusing on precise multimodal information alignment and robust feature extraction. The application of these methodologies ensures that our model not only accurately identifies and localizes important surgical details but also remains resilient against real-world image corruptions. Our extensive evaluations across expanded versions of the EndoVis-18-VQLA and EndoVis-17-VQLA datasets have demonstrated significant improvements in both accuracy and robustness compared to prior models. These achievements underscore the potential of Surgical-VQLA++ to assist in surgical education and clinical decision-making by providing more insightful analyses of surgical scenes.

Moving forward, our potential future works aim to further enhance the capabilities of VQLA in surgical settings. We plan to integrate VQLA systems with mixed reality or simulations to improve surgical education, quickly addressing the queries of students and junior surgeons

and reducing the reliance on limited teaching resources. Additionally, deploying VQLA systems on surgical robots or navigation tools will offer real-time visual assistance, thereby improving the efficiency and safety of surgical operations. To tackle the challenges in VQLA data collection, such as domain shift and temporal discontinuity, we also aim to incorporate auxiliary information from MLLMs and apply continual learning techniques. This approach will help in fine-tuning well-trained VQLA models across diverse datasets, ensuring high performance on both existing and new data and fostering the development of a unified VQLA framework. These initiatives are expected to enhance prediction accuracy and expand the applications of VQLA in surgical environments.

CRedit authorship contribution statement

Long Bai: Conceptualization, Investigation, Methodology, Validation, Writing – original draft. **Guankun Wang:** Investigation, Methodology, Validation, Writing – original draft. **Mobarakol Islam:** Conceptualization, Investigation, Methodology, Writing – original draft. **Lalithkumar Seenivasan:** Investigation, Methodology, Writing – original draft. **An Wang:** Investigation, Writing – original draft. **Hongliang Ren:** Conceptualization, Resources, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

We have shared a link to my code and data in the Abstract.

Acknowledgments

This work was supported by Hong Kong Research Grants Council (RGC) CRF C4026-21GF, GRF 14203323, GRF 14216022, GRF 14211420, and NSFC/RGC Joint Research Scheme N_CUHK420/22; Shenzhen-Hong Kong-Macau Technology Research Programme (Type C) STIC Grant 202108233000303. M. Islam was funded by EPSRC Grant [EP/W00805X/1].

References

- [1] L. Seenivasan, M. Islam, A. Krishna, H. Ren, Surgical-VQA: Visual question answering in surgical scenes using transformer, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2022, pp. 33–43.
- [2] D. Sharma, S. Purushotham, C.K. Reddy, MedFuseNet: An attention-based multimodal deep learning model for visual question answering in the medical domain, *Sci. Rep.* 11 (1) (2021) 1–18.

- [3] L. Bai, M. Islam, L. Seenivasan, H. Ren, Surgical-VQLA: Transformer with gated vision-language embedding for visual question localized-answering in robotic surgery, 2023, arXiv preprint [arXiv:2305.11692](#).
- [4] L.H. Li, M. Yatskar, D. Yin, C.J. Hsieh, K.W. Chang, Visualbert: A simple and performant baseline for vision and language, 2019, arXiv preprint [arXiv:1908.03557](#).
- [5] S. Uppal, S. Bhagat, D. Hazarika, N. Majumder, S. Poria, R. Zimmermann, A. Zadeh, Multimodal research in vision and language: A review of current and emerging trends, *Inf. Fusion* 77 (2022) 149–171.
- [6] S. Zhang, M. Chen, J. Chen, F. Zou, Y.-F. Li, P. Lu, Multimodal feature-wise co-attention method for visual question answering, *Inf. Fusion* 73 (2021) 1–10.
- [7] W. Zhang, J. Yu, W. Zhao, C. Ran, Dmrfnet: deep multimodal reasoning and fusion for visual question answering and explanation generation, *Inf. Fusion* 72 (2021) 70–79.
- [8] N.H. Nguyen, D.T. Vo, K. Van Nguyen, N.L.T. Nguyen, OpenViVQA: Task, dataset, and multimodal fusion models for visual question answering in Vietnamese, *Inf. Fusion* 100 (2023) 101868.
- [9] W. Zhang, J. Yu, H. Hu, H. Hu, Z. Qin, Multimodal feature fusion by relational reasoning and attention for visual question answering, *Inf. Fusion* 55 (2020) 116–126.
- [10] T. Do, B.X. Nguyen, E. Tjiputra, M. Tran, Q.D. Tran, A. Nguyen, Multiple meta-model quantifying for medical visual question answering, in: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V* 24, Springer, 2021, pp. 64–74.
- [11] Y. Khare, V. Bagal, M. Mathew, A. Devi, U.D. Priyakumar, C. Jawahar, Mm-bert: Multimodal bert pretraining for improved medical vqa, in: *2021 IEEE 18th International Symposium on Biomedical Imaging, ISBI, IEEE, 2021*, pp. 1033–1036.
- [12] Y. Liu, Z. Wang, D. Xu, L. Zhou, Q2atransformer: Improving medical vqa via an answer querying decoder, in: *International Conference on Information Processing in Medical Imaging*, Springer, 2023, pp. 445–456.
- [13] S. Takada, R. Togo, T. Ogawa, M. Haseyama, Estimation of visual contents based on question answering from human brain activity, in: *2020 IEEE International Conference on Image Processing, ICIP, IEEE, 2020*, pp. 61–65.
- [14] L. Bai, M. Islam, H. Ren, 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*, Springer, 2023, pp. 397–407.
- [15] L. Bai, M. Islam, H. Ren, Revisiting distillation for continual learning on visual question localized-answering in robotic surgery, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2023, pp. 68–78.
- [16] M. Islam, Z. Li, B. Glocker, Robustness stress testing in medical image classification, 2023, arXiv preprint [arXiv:2308.06889](#).
- [17] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, in: *International Conference on Machine Learning, PMLR, 2021*, pp. 10347–10357.
- [18] M. Allan, S. Kondo, S. Bodenstedt, S. Leger, R. Kadkhodamohammadi, I. Luengo, F. Fuentes, E. Flouty, A. Mohammed, M. Pedersen, et al., 2018 Robotic scene segmentation challenge, 2020, arXiv preprint [arXiv:2001.11190](#).
- [19] M. Allan, A. Shvets, T. Kurmann, Z. Zhang, R. Duggal, Y.-H. Su, N. Rieke, I. Laina, N. Kalavakonda, et al., 2017 Robotic instrument segmentation challenge, 2019, arXiv preprint [arXiv:1902.06426](#).
- [20] S. Manmadhan, B.C. Kooor, Visual question answering: a state-of-the-art review, *Artif. Intell. Rev.* 53 (2020) 5705–5745.
- [21] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C.L. Zitnick, D. Parikh, Vqa: Visual question answering, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2425–2433.
- [22] T.M. Le, V. Le, S. Gupta, S. Venkatesh, T. Tran, Guiding visual question answering with attention priors, in: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 4381–4390.
- [23] Z. Yu, J. Yu, Y. Cui, D. Tao, Q. Tian, Deep modular co-attention networks for visual question answering, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6281–6290.
- [24] M. Parelli, D. Mallis, M. Diomataris, V. Pitsikalis, Interpretable visual question answering via reasoning supervision, in: *2023 IEEE International Conference on Image Processing, ICIP, IEEE, 2023*, pp. 2525–2529.
- [25] C. Deng, Q. Wu, Q. Wu, F. Hu, F. Lyu, M. Tan, Visual grounding via accumulated attention, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7746–7755.
- [26] W. Su, P. Miao, H. Dou, G. Wang, L. Qiao, Z. Li, X. Li, Language adaptive weight generation for multi-task visual grounding, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 10857–10866.
- [27] H. Jiang, Y. Lin, D. Han, S. Song, G. Huang, Pseudo-q: Generating pseudo language queries for visual grounding, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 15513–15523.
- [28] A. Urooj, H. Kuehne, K. Duarte, C. Gan, N. Lobo, M. Shah, Found a reason for me? weakly-supervised grounded visual question answering using capsules, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8465–8474.
- [29] G. Kv, A. Mittal, Reducing language biases in visual question answering with visually-grounded question encoder, in: *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII* 16, Springer, 2020, pp. 18–34.
- [30] A. Fukui, D.H. Park, D. Yang, A. Rohrbach, T. Darrell, M. Rohrbach, Multimodal compact bilinear pooling for visual question answering and visual grounding, 2016, arXiv preprint [arXiv:1606.01847](#).
- [31] L. Zhu, L. Peng, W. Zhou, J. Yang, Dual-decoder transformer network for answer grounding in visual question answering, *Pattern Recognit. Lett.* 171 (2023) 53–60.
- [32] Y. Zhang, J.C. Niebles, A. Soto, Interpretable visual question answering by visual grounding from attention supervision mining, in: *2019 IEEE Winter Conference on Applications of Computer Vision (Wacv)*, IEEE, 2019, pp. 349–357.
- [33] D. Reich, F. Putze, T. Schultz, Measuring faithful and plausible visual grounding in VQA, 2023, arXiv preprint [arXiv:2305.15015](#).
- [34] A. Mani, N. Yoo, W. Hinthorn, O. Russakovsky, Point and ask: Incorporating pointing into visual question answering, 2020, arXiv preprint [arXiv:2011.13681](#).
- [35] B. Liu, L.M. Zhan, X.M. Wu, Contrastive pre-training and representation distillation for medical visual question answering based on radiology images, in: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part II* 24, Springer, 2021, pp. 210–220.
- [36] U. Naseem, M. Khushi, A.G. Dunn, J. Kim, K-PathVQA: Knowledge-aware multimodal representation for pathology visual question answering, *IEEE J. Biomed. Health Inf.* (2023).
- [37] J.J. Lau, S. Gayen, A. Ben Abacha, D. Demner-Fushman, A dataset of clinically generated visual questions and answers about radiology images, *Scientific data* 5 (1) (2018) 1–10.
- [38] S. Tascon-Morales, P. Márquez-Neila, R. Sznitman, Consistency-preserving visual question answering in medical imaging, in: *International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2022, pp. 386–395.
- [39] S. Tascon-Morales, P. Márquez-Neila, R. Sznitman, Localized questions in medical visual question answering, 2023, arXiv preprint [arXiv:2307.01067](#).
- [40] M.H. Vu, T. Löfstedt, T. Nyholm, R. Sznitman, A question-centric model for visual question answering in medical imaging, *IEEE Trans. Med. Imaging* 39 (9) (2020) 2856–2868.
- [41] B. Li, Y. Weng, B. Sun, S. Li, Towards visual-prompt temporal answering grounding in medical instructional video, 2022, arXiv preprint [arXiv:2203.06667](#).
- [42] Y. Cai, X. Sui, G. Gu, Multi-modal multi-task feature fusion for RGBT tracking, *Inf. Fusion* 97 (2023) 101816.
- [43] W. Zhang, G. Yang, N. Zhang, L. Xu, X. Wang, Y. Zhang, H. Zhang, J. Del Ser, V.H.C. de Albuquerque, Multi-task learning with multi-view weighted fusion attention for artery-specific calcification analysis, *Inf. Fusion* 71 (2021) 64–76.
- [44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, *Adv. Neural Inf. Process. Syst.* 30 (2017).
- [45] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, *Adv. Neural Inf. Process. Syst.* 28 (2015).
- [46] J. Devlin, M.W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2018, arXiv preprint [arXiv:1810.04805](#).
- [47] I.J. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, 2014, arXiv preprint [arXiv:1412.6572](#).
- [48] L. Pan, C.W. Hang, A. Sil, S. Potdar, Improved text classification via contrastive adversarial training, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, 36, (10) 2022, pp. 11130–11138.
- [49] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.
- [50] J. Deng, W. Dong, R. Socher, L.J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: *2009 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2009, pp. 248–255.
- [51] L. Seenivasan, S. Mithran, M. Islam, H. Ren, Global-reasoned multi-task learning model for surgical scene understanding, *IEEE Robot. Autom. Lett.* (2022).
- [52] Y. Wu, S. Zhao, S. Qi, J. Feng, H. Pang, R. Chang, L. Bai, M. Li, S. Xia, W. Qian, et al., Two-stage contextual transformer-based convolutional neural network for airway extraction from CT images, 2022, arXiv preprint [arXiv:2212.07651](#).
- [53] H. Yuan, M. Wang, D. Ni, L. Xu, Detecting human-object interactions with object-guided cross-modal calibrated semantics, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, 36, (3) 2022, pp. 3206–3214.
- [54] C.F.R. Chen, Q. Fan, R. Panda, Crossvit: Cross-attention multi-scale vision transformer for image classification, in: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 357–366.
- [55] J.H. Kim, K.W. On, W. Lim, J. Kim, J.W. Ha, B.T. Zhang, Hadamard product for low-rank bilinear pooling, 2016, arXiv preprint [arXiv:1610.04325](#).

- [56] J. Arevalo, T. Solorio, M. Montes-y Gómez, F.A. González, Gated multimodal units for information fusion, 2017, arXiv preprint [arXiv:1702.01992](#).
- [57] T. Chen, S. Kornblith, M. Norouzi, G. Hinton, A simple framework for contrastive learning of visual representations, in: *International Conference on Machine Learning*, PMLR, 2020, pp. 1597–1607.
- [58] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, S. Savarese, 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*, 2019, pp. 658–666.
- [59] T.Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2980–2988.
- [60] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: *European Conference on Computer Vision*, Springer, 2020, pp. 213–229.
- [61] A. Kendall, Y. Gal, R. Cipolla, Multi-task learning using uncertainty to weigh losses for scene geometry and semantics, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7482–7491.
- [62] H. Ben-Younes, R. Cadene, M. Cord, N. Thome, Mutan: Multimodal tucker fusion for visual question answering, in: *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2612–2620.
- [63] Z. Yu, J. Yu, C. Xiang, J. Fan, D. Tao, Generalized multimodal factorized high-order pooling for visual question answering, *IEEE Trans. Neural Netw. Learn. Syst.* 29 (12) (2018) 5947–5959.
- [64] H. Ben-Younes, R. Cadene, N. Thome, M. Cord, Block: Bilinear superdiagonal fusion for visual question answering and visual relationship detection, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, 33, (01) 2019, pp. 8102–8109.
- [65] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint [arXiv:1412.6980](#).
- [66] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, D. Ren, Distance-iou loss: Faster and better learning for bounding box regression, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, 34, (07) 2020, pp. 12993–13000.
- [67] R.G. Praveen, W.C. de Melo, N. Ullah, H. Aslam, O. Zeeshan, T. Denorme, M. Pedersoli, A.L. Koerich, S. Bacon, P. Cardinal, et al., A joint cross-attention model for audio-visual fusion in dimensional emotion recognition, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2486–2495.
- [68] M.I. Georgescu, R.T. Ionescu, A.I. Miron, O. Savencu, N.C. Ristea, N. Verga, F.S. Khan, Multimodal multi-head convolutional attention with various kernel sizes for medical image super-resolution, in: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 2195–2205.
- [69] Z. Wu, L. Liu, Y. Zhang, M. Mao, L. Lin, G. Li, Multimodal crowd counting with mutual attention transformers, in: *2022 IEEE International Conference on Multimedia and Expo, ICME, IEEE, 2022*, pp. 1–6.
- [70] D. Hendrycks, T. Dietterich, Benchmarking neural network robustness to common corruptions and perturbations, 2019, arXiv preprint [arXiv:1903.12261](#).
- [71] L.C. Garcia-Peraza-Herrera, L. Fidon, C. D'Ettorre, D. Stoyanov, T. Vercauteren, S. Ourselin, Image compositing for segmentation of surgical tools without manual annotations, *IEEE Trans. Med. Imaging* 40 (5) (2021) 1450–1460.
- [72] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F.L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, 2023, arXiv preprint [arXiv:2303.08774](#).
- [73] J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou, J. Zhou, Qwen-vl: A frontier large vision-language model with versatile abilities, 2023, arXiv preprint [arXiv:2308.12966](#).
- [74] A. Gunjal, J. Yin, E. Bas, Detecting and preventing hallucinations in large vision language models, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, 38, (16) 2024, pp. 18135–18143.
- [75] X. Shen, R. Zhang, X. Zhao, J. Zhu, X. Xiao, PMG: Personalized multimodal generation with large language models, 2024, arXiv preprint [arXiv:2404.08677](#).