

Towards Flexible Evaluation for Generative Visual Question Answering

Huishan Ji, Qingyi Si, Zheng Lin*
 {jihuishan, siqingyi, linzheng}@iie.ac.cn
 Institute of Information Engineering,
 Chinese Academy of Sciences
 School of Cyber Security,
 University of Chinese Academy of Sciences
 Beijing, China

Weiping Wang
 wangweiping@iie.ac.cn
 Institute of Information Engineering,
 Chinese Academy of Science
 Beijing, China

Abstract

Throughout rapid development of multimodal large language models, a crucial ingredient is a fair and accurate evaluation of their multimodal comprehension abilities. Although Visual Question Answering (VQA) could serve as a developed test field, limitations of VQA evaluation, like the inflexible pattern of Exact Match, have hindered MLLMs from demonstrating their real capability and discourage rich responses. Therefore, **this paper proposes the use of semantics-based evaluators for assessing unconstrained open-ended responses on VQA datasets**. As characteristics of VQA have made such evaluation significantly different than the traditional Semantic Textual Similarity (STS) task, to systematically analyze the behaviour and compare the performance of various evaluators including LLM-based ones, we propose three key properties, i.e., **Alignment, Consistency and Generalization**, and a corresponding dataset Assessing VQA Evaluators (AVE) to facilitate analysis. In addition, this paper proposes a Semantically Flexible VQA Evaluator (SFVE) with meticulous design based on the unique features of VQA evaluation. Experimental results verify the feasibility of model-based VQA evaluation and effectiveness of the proposed evaluator that surpasses existing semantic evaluators by a large margin. The proposed training scheme generalizes to both the BERT-like encoders and decoder-only LLM. **Related codes and data available at Our Repository**.

CCS Concepts

• **Computing methodologies** → **Scene understanding**; *Activity recognition and understanding*; Information extraction.

Keywords

Visual Question Answering, Semantic Textual Similarity, Contrastive Learning, Evaluation Method

ACM Reference Format:

Huishan Ji, Qingyi Si, Zheng Lin* and Weiping Wang. 2024. Towards Flexible Evaluation for Generative Visual Question Answering. In *Proceedings of*

*Corresponding author: Zheng Lin, linzheng@iie.ac.cn.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

MM '24, October 28–November 1, 2024, Melbourne, VIC, Australia

© 2024 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-0686-8/24/10

<https://doi.org/10.1145/3664647.3681400>

the 32nd ACM International Conference on Multimedia (MM '24), October 28–November 1, 2024, Melbourne, VIC, Australia. ACM, New York, NY, USA, 12 pages. <https://doi.org/10.1145/3664647.3681400>

1 Introduction

Visual Question Answering (VQA) evaluates the multimodal comprehension abilities by posing questions about given images and comparing the model's responses with annotated answers [22, 24, 25, 39, 41, 47, 53]. However, current VQA evaluation metrics have made it tough for evaluating the rich responses of Multimodal Large Language Models (MLLMs).

Most VQA datasets comply with a triplet format and each sample consists of a question, image and annotation. Annotations are often a single word or phrase [25, 39, 47, 53, 54] or a set of ten candidate answers [22, 24, 41, 48]. Current evaluation metrics, Exact Match [39] (for samples without candidate answers) and VQA Score [6] (for samples with ten candidate answers), both require the responses to be identical in morphology with the annotation to be considered correct. Variations in tense, singular or plural forms and synonyms are not allowed, let alone sentence-style responses from MLLMs.

Traditional vision-language models treat VQA as a classification problem [44, 47, 51, 58, 59], where answers collected from the training set are used to establish pre-defined classes, and the possible responses are constrained to these classes. Thus the problem of evaluating multifarious responses does not exist. However, MLLMs treat VQA as a generative problem [8, 18, 30, 34] and generates assorted responses. Meanwhile, the growing trend that the MLLM community prefers zero-shot test, has made it even tougher for models to generate responses that are identical to the ground-truth answers. As shown in Figure 1, semantically equivalent but morphologically distinct responses are not accepted.

Although it is possible to force the model to output a single word with a harsh prompt, such remediation may potentially damage the performance and make it unfair for different MLLMs, especially for those with poor instruction-following ability and those that tend to response with long sentences. As MLLMs inherit the in-context learning capability of LLMs, it is feasible to introduce in-context examples to force short responses. However, since different MLLMs contain different in-context learning capability, such practice interferes a fair evaluation of MLLMs' multimodal comprehension performance. There thus is an urgent need for a metric that aligns well with human judgment and accommodates various response types while ensuring consistent evaluation despite variations in response morphology.

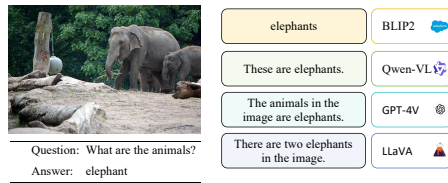


Figure 1: Responses from four MLLMs on a simple visual question. The responses are different in length, styles and complexity, which can all be considered correct but none of them exactly matches the annotated answer.

To compare different evaluators, traditional Semantic Textual Similarity (STS) task measures the difference between predicted scores and human annotation results from a single aspect of semantic relevance. However, both intuition and our experiments (refer to Section 3.1 and 5.3) suggest that there is a significant difference between the evaluation of correctness in VQA responses and the traditional assessment of media-text relevance in STS.

Therefore, to systematically evaluate the performance of an evaluator, with the unique characteristics in the task of VQA response evaluation taken into consideration, we propose three quantitative key properties, i.e., **Alignment**, **Consistency** and **Generalization**. **Alignment** stands for the overall correspondence of predicted scores with human annotation. **Consistency** measures how well an evaluator accommodates semantically equivalent responses of different morphology and length. **Generalization** indicates the variance of performance on different sources of data. Further, to facilitate comprehensive analysis of the performance and behaviour of evaluators, we provide a human-annotated dataset Assessing VQA Evaluators (AVE) that grades the correctness of model responses towards ground-truth labels on VQA datasets. AVE is further augmented by ChatGPT and WordNet [42] to increase the diversity.

As pilot experiment shows (refer to Section 5.3), formulaic metrics (BLEU [45], ROUGE [33], METEOR [10]) and model-based metrics [14, 21, 32, 46] perform poorly on the VQA response evaluation task. Therefore, we propose a novel evaluator that is trained with meticulously designed pretraining tasks. The tasks are designed for improving the embedding representation of VQA text, which utilizes contrastive learning to guide the evaluator to capture the fine-grained difference within a text pair and ignore the noise in morphology and length. Experiments demonstrate that the proposed pretraining tasks significantly improve the performance of our evaluator on the AVE dataset, making the evaluator’s prediction aligns much better with human judgement.

The contribution of this paper can be concluded as follows:

- This paper addresses the dilemma, where rich responses of MLLMs hinder fair evaluation under current metric, by proposing semantic-similarity-based evaluation that applies to various VQA responses.
- This paper proposes three quantitative key properties in VQA response evaluation based on its characteristics, and a high-quality human-annotated dataset, AVE, for assessing different evaluators comprehensively. In addition, we evaluate the performance of various types of existing semantic similarity evaluators on the proposed AVE dataset.

- Experimental results demonstrate the feasibility of applying model-based methods to the flexible evaluation of VQA responses as well as the effectiveness of our proposed evaluator. Our evaluator significantly surpasses existing methods, including ChatGPT and the SOTA embedding model Voyage-lite-02-Instruct¹ by a large margin. Our training scheme generalizes to both encoder-only and decoder-only models.

2 Related Work

2.1 Visual Question Answering

As the answer space of most open-ended VQA datasets is limited and the same answer applies for multiple questions (the most common 2,000 answers in the training set of VQA v2[22] is able to cover about 94% questions in its validation set), early methods [44, 47, 51, 58, 59] treat VQA as a classification task, which adopt answers in the training set as class labels, unable to predict unseen classes. Generative methods on VQA [8, 18, 30, 34, 35] treat VQA as a generation task and facilitates much more variant responses.

2.2 Semantic Textual Similarity

Current semantic evaluation tasks include Semantic Textual Similarity (STS) [4, 5] that assesses to what extent the two sentences are related, Paraphrase Identification [50, 61] that decides whether two texts express the same meaning, and Natural Language Inference [17, 56] that determines the logical relationship between texts. The essence of these tasks lies in quantifying the degree of semantic equivalence between sentences, which is a fundamental challenge due to the complexity and variability of natural language. Methods in STS include formulaic methods like BLEU [45] and model-based ones [21, 32, 43, 46, 64]. The former mainly relies on n-gram or other statistic features between the candidate and reference to calculate the overlap and import penalty for noise. The latter utilizes models as encoders to extract the information and compare between the candidate and reference. Early model-based evaluator [64] compares the similarity of each pair each time, which is computation-consuming. Later works [21, 32, 43, 46] first generate embeddings separately for the candidate and reference, then simply calculate the cosine similarity between them as the similarity score. Due to the style of STS, either formulaic or model-based methods pay more attention to the overall similarity and are less capable of detecting fine-grained semantic difference, as shown in our experiments.

2.3 Multimodal Comprehension Evaluation of MLLMs

As a developed realm and valuable resource of high-quality data, VQA has been applied in the evaluation of MLLMs. Several works [13, 48, 62] propose multiple-choice datasets to make evaluation simple and straight. For example, MME [13] proposes a smart quantitative analysis of MLLMs with manually designed instruction-answer pairs that strictly limit responses to be yes or no. Therefore, all MLLMs are evaluated relatively fairly. However, although MME

¹As of the time of submission, Voyage-lite-02-Instruct achieves the best performance on the task of Semantic Textual Similarity (STS).

is insightful and effective, such detour avoids the problem of evaluating open-ended response directly. It ignores previous huge amount of VQA data and costs additional human annotation, limiting the scale of the dataset and making it tough for expanding. MM-vet [63], LVLM-eHub [49] and ConvBench [36] classify VQA into the integration of multiple key abilities, and manually annotates corresponding VQA samples of their required abilities. Then, they use ChatGPT for evaluation (which we show to be less capable in evaluating the correctness of open-ended responses, refer to Section 5.3). The classification of VQA abilities is insightful and aids to the probing of specific abilities of MLLMs. Yet MM-vet requires high-quality annotation to identify the VQA abilities each question requires and are thus limited to a small amount too.

3 Semantic Evaluation of VQA

With the rapid development of MLLMs, current metrics in VQA response evaluation are too stubborn to assess the rich generation and hinder evaluating MLLMs' performance with existing VQA datasets. Meanwhile, as mentioned in Section 2.2, current semantic evaluation models and tasks are inconsistent with the goal of flexible VQA response evaluation.

Therefore, this paper proposes the task of semantic evaluation of VQA, aiming at introducing flexible similarity-based soft evaluation with continuous scores into the assessment of VQA responses, contrary to inflexible metrics like Exact Match or VQA Score that require identical morphology of responses towards ground-truth labels. Such flexible evaluation enables to assess the rich generation from MLLMs and thus enables to use existing VQA datasets for probing MLLMs' multimodal comprehension ability.

3.1 Characteristics of VQA Evaluation

The proposed task of semantic VQA response evaluation shares significant difference with existing semantic evaluation tasks like STS and contains its own characteristics.

Discrimination Granularity. As mentioned in Section 1, traditional semantic evaluation tasks typically focus on the overall meaning in texts, rather than capturing the fine-grained detailed difference. However, the core of semantic VQA evaluation is comparing the response with annotated answer under the same question², where both texts share large overlap in meaning as the questions are same, demanding fine-grained semantic discrimination.

Text Length. As VQA answers are generally much shorter than texts in STS³, n-gram based formulaic metrics like BLEU [45] will be more easily affected by the context in response. Model-based metrics are also vulnerable to such length shift, as their training data barely cover similar pattern.

Distribution Shift. The texts in STS datasets [2, 4, 5] come from general domains, like news and social media, while different VQA

datasets comply to different sub-tasks, like knowledge [41] or reasoning [25]. Such distribution shift causes inconsistent evaluation on responses from different VQA datasets.

3.2 Three Key Properties in VQA Evaluation

To systematically evaluate the performance of a VQA evaluator, we propose three quantitative key properties, i.e., **Alignment**, **Consistency** and **Generalization**.

Alignment. Alignment assesses the overall performance of similarity scores predicted by evaluators with that of human annotation, in the metric of Spearman's Rank Correlation following similar setting in previous works[4, 5, 21, 32].

Consistency. A smart evaluator shall catch the key information in responses and ignore the noise text, e.g., the response of *elephants* shall be scored equally with *Theses are elephants* under the question of *What are the animals?*. Therefore, Consistency measures how close the different responses sharing the same meaning are scored.

Generalization. Considering various VQA datasets focus on various sub-tasks and come from various sources, Generalization depicts how well an evaluator is able to handle text from different domains. Refer to Section 3.4 for quantitative definitions.

3.3 A Dataset Assessing VQA Evaluators

To comprehensively compare and analyze the behaviour of different evaluators on VQA responses, taking the proposed three key properties into consideration, we propose a dataset Assessing VQA Evaluators (AVE). By collecting multiple MLLMs' responses on multiple datasets, the proposed dataset simulates a real scene of applying evaluators to evaluate the quality of various VQA responses. In order to compare the evaluators' scoring results with human judgement, we provide human annotation of the semantic correctness of responses towards ground-truth answers. The construction process of AVE is shown as follows:

Response Collection. First, we collect responses of five models, LLaVA [35], BLIP2 [30], mPLUG-Owl [60], OFA-large [55], Qwen-VL [8] on the validation set of four datasets, OKVQA[41], A-OKVQA[48], VQA v2[22] and GQA[25] (balanced testdev set).

Sampling Results. Second, we sample in the responses while controlling the sampling amount of each dataset to be the same. In addition, samples that are answered correctly, i.e., the response is identical with the ground-truth answer, are excluded.

Human Annotation. Third, three annotators are asked to measure the semantic similarity⁴ of each sampled response towards the ground-truth label and annotate an integral similarity score from 0 to 10, under certain rules (refer to Appendix for more details). Then the scores are averaged over the three annotators.

Description Generation. Fourth, in order to simulate MLLM responses with sentences instead of words or phrases, we select responses that are shorter than three words for augmentation⁵. The

²Considering the polysemy and ambiguity of words and phrases, the question text is indispensable for evaluating the semantic correctness.

³About 97.9% answers in VQA v2 [22], 97.7% in OKVQA [41], 99.9% in GQA [25] are shorter than three words. The average length of text in STS-12[4] is 12.5, which is much longer.

⁴We considered multiple aspects of measuring the correctness of a response towards the ground-truth answer, yet at last we come to the single aspect of semantic similarity for annotation. Refer to Appendix for more explanation.

⁵For longer responses, we manually condense it.

augmentation is conducted in two ways. The first way comes from using ChatGPT (refer to Appendix for prompts) to convert each pair of question and response into three descriptions and asking ChatGPT to select two descriptions that are closest to the original question-answer pair as augmented responses. For example, the question of *What are the animals?* and the response of *elephants* are fed into ChatGPT, and it generates descriptions like *The animals are elephants*. The second way comes from applying manually designed answer templates (refer to Appendix) to increase the diversity of descriptions rather than fully relying on ChatGPT. Now each sample contains three descriptions, i.e., the two ChatGPT-augmented descriptions and a manual-template-fitted one.

Synonym Generation. Fifth, we use WordNet [42] to locate a synonym for each answer. For cases where multiple synonyms exist, we choose the most common synonym by countering frequencies of words in Brown Corpus [20]. Meanwhile, we ask ChatGPT to introduce a shift in morphology to simulate cases that the outputs are merely different in tenses or singular or plural forms. The augmented answer is deemed to contain the same meaning with small disturbance on the style.

Manual Filter. At last, to ensure high-quality of the dataset, the three annotators also conduct manual filter (refer to Appendix for rules) to eliminate ambiguous samples, especially those generated from the fifth stage.

Generally speaking, the whole AVE dataset consists of three parts generated from above: Part 1 contains original answers and original responses (without augmentation). Part 2 contains original answers and the generated three descriptions of responses. Part 3 contains tense-shifted answers and original responses. The whole procedure is depicted in Figure 2.

Meanwhile, the AVE dataset can also be clustered by the involved four datasets that each sample belongs to, OKVQA, A-OKVQA, VQA v2 and GQA, merging the Part 1 to Part 3 together and classifies by the sources of data only. Refer to Section 3.4 for how they are used.

The total sample amount of the final AVE dataset is 3,592, with each sample containing four types of augmentation results, as described above. The dataset is then split into a validation set and a test set with the ratio of 3:7. The distribution of annotated scores is shown in Appendix, which is relatively smooth. To evaluate the inter-annotator agreement, following previous works [11, 19, 23], we apply Krippendorff's alpha [28] and obtain a result of 0.713.

3.4 The Proposed Evaluation Indicators

In AVE dataset, a sample consists of a question q_i , a ground-truth label a_i , a source dataset label d_i , a response r_i and a human-annotated score s_i . The task of VQA response evaluation can be defined as: given a question-answer pair, an evaluator $f(q_i, a_i, r_i)$ is expected to predict the annotated similarity score s_i with the output o_i .

$$o_i = \cos(f(q_i, r_i), f(q_i, a_i)) \quad (1)$$

$$\text{score}_f = \text{Spearman}(O, S) \quad (2)$$

where \cos stands for cosine similarity, score_f is the performance score of the evaluator f , with O and S indicating the lists of all predicted scores and annotated scores respectively. Note that the

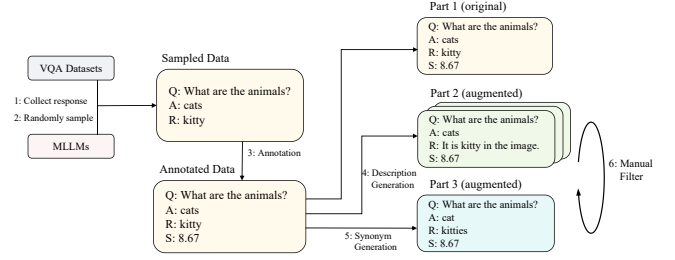


Figure 2: The construction procedure of AVE. After randomly sampled from the outputs of models, each sample is manually annotated with a score and automatically augmented by generated descriptions and a variation on the answer word while remaining almost the same correctness as a VQA response. Part 1 to 3 denote different augmentation methods.

metrics used for evaluating evaluators is Spearman's rank coefficient of correlation (Spearman).

The key properties of Alignment, Consistency and Generalization introduced in Section 3.2 are computed as follows:

Alignment. We use the average result of an evaluator on all parts of the proposed AVE dataset as alignment:

$$\text{Alignment} = \frac{1}{N_{\text{Parts}}} \sum_{i=1}^{N_{\text{Parts}}} \text{score}_{f_i} \quad (3)$$

where N_{Parts} is the number of Parts in the AVE dataset and N_{sets} is the number of involved VQA datasets in AVE, according to different type of division. The score_{f_i} is the spearman score of evaluator f on the i th part of AVE.

Consistency. Consistency measures how close the responses of the same meaning with different morphology are evaluated. We regard the variance of the same sample among the three parts of AVE as consistency:

$$\text{Consistency} = \log \left(1 / \left(\frac{1}{N_{\text{samples}}} \sum_{j=1}^{N_{\text{samples}}} \text{var}(o_{j1}, o_{j2}, o_{j3}) \right) \right) \quad (4)$$

where N_{samples} is the amount of samples and var denotes calculating the variance. Then, o_{j1}, o_{j2}, o_{j3} are the predicted scores of the evaluator on Part 1, 2, 3 for the same sample j , respectively.

Generalization. Generalization measures the difference of performance on various datasets, and we define it as the variance of the performance on each involved VQA dataset:

$$\text{Generalization} = \log 1 / \text{var}(\text{align}_{\text{OKVQA}}, \text{align}_{\text{A-OKVQA}}, \text{align}_{\text{VQA v2}}, \text{align}_{\text{GQA}}) \quad (5)$$

where $\text{align}_{\text{dataset}}$ is the mean Alignment score on the AVE data belonging to the corresponding VQA dataset.

4 Semantically Flexible VQA Evaluator

With the three key properties of an ideal evaluator taken into consideration, we propose a novel evaluator based on meticulously designed pretraining tasks.

4.1 Pretraining Tasks

To guide the model to be sensitive to the key information between answer and response, this paper introduces several pretraining tasks to enhance the embedding. Data for augmentation come from a random sampling of VQA data in the training sets of OKVQA[41], A-OKVQA[48], TDIUC[27], VG-QA[29], GQA[25] and VQA v2 [22], which end to a total amount of 105, 311 samples. All augmented samples are mixed for training.

NLI data. In previous works [21, 32], models perform well with the natural language inference datasets SNLI [17] and MNLI [56], where each sample includes a premise, an entailment and a contradiction. In addition, these NLI datasets are all manually constructed, ensuring the high quality of their data, and the premise shares limited overlap with entailment compared with the sentence pairs in back-translation datasets. Therefore, to ensure the fundamental discriminating ability of models to capture overall meaning of sentences, we adopt NLI data and regard the premise-entailment pairs as positive pairs and premise-contradiction pairs as negative pairs.

Candidate answers. To make the best of available VQA datasets, for datasets with ten candidate answers, OKVQA, A-OKVQA and VQA v2, we consider candidate answers as correct answers as well. Then, for each sample, the most common candidate answer and a less common one are used to form a positive pair, with a random answer sampled from the answer space as negative.

Synonym and Antonym. In VQA response evaluation, semantically similar answers shall receive similar scores. We replace the answer with a synonym by WordNet [42], and if the antonym of an answer exists, we then pair up the answer and antonym as a negative pair, else we pair up the answer and a randomly sampled answer from the answer space as a negative pair. In addition, we use ChatGPT to produce synonyms as well, as ChatGPT is able to capture contextual information in the question and thus generates more accurate synonyms.

Generated descriptions. To simulate the output of MLLMs, for each sample, we provide ChatGPT with its question and answer to generate three descriptions with small disturbance of the same meaning. Then, we construct positive samples by pairing up the original answer and each generated description. For negative samples, we replace the answer in the generated description with a randomly sampled answer. The goal is to pull the embedding representation of a natural language description close to its simple form of a single answer, so that responses with different length but carrying similar meanings will receive similar scores. In addition, the negative pair is constructed by replacing the key answer word in the description, therefore guiding the model to be sensitive to the key words and to ignore the noise.

4.2 Model Framework

Following previous works [15, 16, 21, 32] on the STS task [1–5, 12], we use cosine similarity for distance calculation between embeddings [21, 32]. As shown in Figure 3, we adopt the simple contrastive learning framework [15] and contrastive learning with in-batch hard negatives [21].

The backbone encoder in this paper is RoBERTa [37]. In order to gain better generalization and comprehension ability, we apply the decoder-only LLM LLAMA2 [52] with the prompt[32] of *Summarize the text [text] in a single word*. Then, the hidden states of the first generated new token is considered as the embedding vector.

The contrastive learning with in-batch hard negatives loss [21] is defined as follows:

$$loss_{ibn} = -\log \frac{e^{sim(h_i, h_i^+)/\tau}}{\sum_{j=1}^N (e^{sim(h_i, h_i^+)/\tau} + e^{sim(h_i, h_i^-)/\tau})} \quad (6)$$

where h_i is the embedding representation of sample i , h_i^+ and h_i^- respectively denote the representation of the positive sample and in-batch hard negative sample of sample i .

5 Experiments

5.1 Implementation Details

Experiments in this paper is based on *transformer* package[57] on Pytorch. We use AdamW [38] optimizer, and the hyper-parameters of AdamW, betas, eps and weight-decay are set to 0.9, 0.999, 1e-8 and 0.01. We use a cosine scheduler and the batch size and peak learning rate for encoders are 128, 1e-5 for RoBERTa-base [37], VisualBERT [31] and LXMERT [51], 32, 6e-6 for RoBERTa-large and 8, 4e-6 for LLAMA2 [52].

5.2 Baselines

To assess to what extent existing models are competent for the VQA response evaluation, this paper collects four types of common methods for semantic similarity evaluation and refer to them as: formulaic, PLM, LLM, and API.

- Formulaic methods contain BLEU [45], ROUGE [33] and METEOR [10]. These methods base on n-grams for assessing the overlap. As VQA answers are usually short, we also report 2-gram results for BLEU and ROUGE.
- PLM refers to the Pretrained Language Models, which are generally small in sizes and typically in BERT-like encoder-only structures. SBERT (Sentence BERT) [46] embeds texts into vectors with BERT and apply cosine similarity to measure the distance as textual similarity. SIMCSE [21] provides both unsupervised and supervised methods, and this paper selects the supervised and better-performing one trained on NLI datasets for comparison. BGE [14] follows a multi-task learning scheme that collects and pretrains on multifarious datasets for better generalization. Angle [32] aims to mitigate the gradient saturation issue encountered when using cosine distance by projecting vectors onto the complex plane and introducing an angular loss.
- LLM refers to large language models. This paper selects four of the well-performing LLMs, Baichuan2 [9], Qwen [7], LLAMA-2 [52] and Mistral [26].
- API refers to the remote usage of models by API online, including ChatGPT and Text-embedding-v3-large from OpenAI, and Voyage-lite-02-instruct from Voyage AI. Refer to Appendix for prompts. The latter two are embedding models

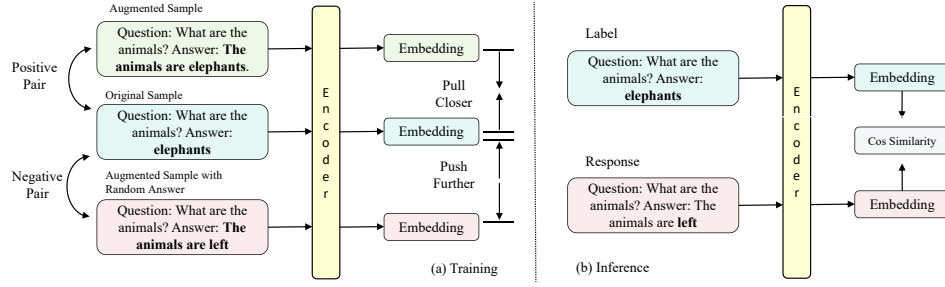


Figure 3: Framework of contrastive learning in the proposed Semantically Flexible VQA Evaluator (SFVE). The original sample is augmented into two variations and form a positive pair and a negative pair. The example in the figure shows the procedure of the pretraining task *Generated descriptions*. In the positive pair, the semantics of the sentence is considered same as the original, while in the negative pair, as the answer word is replaced with a random answer, the sentence contains unmatched meaning with the original.

Types	Methods	Alignment ↑				Consistency ↑	Generalization ↑	STS Avg. ↑
		Part 1	Part 2	Part 3	Avg.			
Formulaic	BLEU-2	-1.7	-2.6	3.2	-0.4	4.35	11.30	50.1
	BLEU-4	-2.9	-3.5	2.0	-1.5	4.23	11.25	47.6
	ROUGE-2	-2.8	-4.5	1.2	-2.0	5.79	11.82	53.9
	ROUGE-L	4.9	-0.1	3.3	2.7	6.73	10.78	48.3
	METEOR	12.4	4.4	15.3	10.7	7.25	9.34	53.4
PLM	RoBERTa-large (w/o CL)	11.9	0.7	23.4	12.0	7.91	10.36	27.9
	SBERT	47.7	44.3	40.6	44.2	8.79	8.55	76.8
	SIMCSE	44.9	44.7	41.7	43.7	9.37	8.30	83.8
	BGE	42.3	36.5	41.0	39.9	8.93	8.25	84.9
	Angle	43.4	38.2	40.2	40.6	9.01	7.78	86.4
LLM	Baichuan2-7b	28.1	27.8	31.8	29.2	5.30	8.86	64.6
	Qwen-7b	25.9	26.3	24.2	25.5	9.07	10.10	68.3
	LLaMA2-7b	32.7	27.9	34.6	31.7	7.45	8.53	61.9
	Mistral-7b	16.8	14.5	20.7	17.3	4.61	8.49	72.1
API	ChatGPT	21.2	15.2	24.6	20.3	5.21	8.35	73.7
	Text-embedding-v3-large	32.5	28.6	36.3	32.5	9.40	8.00	82.3
	Voyage-lite-02-instruct	29.1	28.9	29.3	29.1	11.81	6.78	86.3
SFVE (ours)	SFVE-base	58.4	57.1	53.7	56.4	9.12	8.34	81.2
	SFVE-large	58.1	57.5	56.0	57.2	9.53	8.67	82.0
	SFVE-LLaMA2-7b	60.2	57.0	57.2	58.1	9.46	8.87	77.9

Table 1: The comparison of performance on our proposed AVE dataset. The STS Avg. denotes the average scores over STS 2012 to STS 2016 [1–5], SICK-R [40] STS-B [12], providing a reference of methods’ general discriminating ability. RoBERTa-large [37] (w/o CL) refers to the original pretrained checkpoint without contrastive learning. SFVE-base, SFVE-large and SFVE-LLaMA2-7b are RoBERTa-base, RoBERTa-large and LLaMA2-7b trained by contrastive learning on our proposed pretraining tasks. The specific model checkpoints in experiments are as follows: SBERT[46]: SROBERTa-NLI-large, BGE[14]:BAAI-bge-large-en, SIMCSE[21]: RoBERTa-NLI-large, Angle [32]: RoBERTa-large.

Settings	Alignment ↑				Consistency ↑	Generalization ↑
	Part 1	Part 2	Part 3	Avg.		
All tasks	58.4	57.1	53.7	56.4	9.12	8.34
w/o NLI data	53.3	52.3	50.9	52.2	8.71	8.50
w/o Candidate answers	57.3	56.5	53.0	55.6	9.33	8.54
w/o Synonym and Antonym	42.1	40.3	38.8	40.4	9.19	8.41
w/o generated descriptions	56.9	47.0	52.3	52.1	8.07	8.93
w/o All tasks	12.5	3.1	20.0	11.8	10.58	9.90
NLI data only	44.3	42.0	33.1	39.8	9.46	10.00
Candidate answers only	37.4	29.8	39.6	35.6	8.10	8.11
Synonym and Antonym only	53.8	43.7	50.6	49.4	7.89	9.08
Generated descriptions only	42.8	49.1	42.1	44.6	8.17	7.67

Table 2: Ablation experiments of designed pretraining tasks on RoBERTa-base. The row of All tasks represents the best performance of RoBERTa-base with all pretraining tasks, and the row of w/o All tasks contains results from testing on the RoBERTa-base checkpoint without further training. w/o represents without the corresponding pretraining task, contrary to the setting in lower part of the table where the model is trained only on a single task each time.

that produce text embedding of given text, which are then used to calculate similarity score by cosine distance.

5.3 Main Experiments

Table 1 exhibits the performance comparison of various evaluators. The results are on the test set of AVE, with specific scores on each

part of the dataset, as described in Section 3.3. To promote a comprehensive assessment of existing methods, this paper compares the performance with four common types of methods for semantic evaluation, as introduced in Section 5.2. The last row of Types contains our results from training with the proposed pretraining tasks on the corresponding model. Then, the column of Alignment contains the separate results on each of the three parts in AVE datasets and their average.

5.3.1 Performance of Formulaic Methods. Formulaic methods, i.e., BLEU, ROUGE and METEOR perform poorly in Alignment scores, and some of them drop below 0, indicating adverse scores to the human annotation. Such phenomenon is expected, as the n-gram matching strategy of BLEU and ROUGE is unable to handle the synonyms or variations in tenses and singular or plural forms. For METEOR, however, it applies port stem [10] and synonym matching to preprocess the 1-gram in both the candidate and reference, restoring words to stems and thus performs better.

In addition, it is interesting to notice that although the Alignment of BLEU and ROUGE are much lower than that of METEOR, their Generalization scores are much higher. There are two reasons to this anomaly. First, BLEU and ROUGE fail to handle the task well and their prediction can be considered random, thus the sources of data do not affect the results, just like RoBERTa-large (w/o CL). Second, these n-gram evaluators do not involve semantics, therefore the sources of data that causes word distribution shift matter less.

5.3.2 Performance of PLMs. The BERT-like models pretrained for textual similarity prediction, i.e., SBERT, SIMCSE, BGE, AngLe (the latter four models), show much better performance than RoBERTa-large (w/o CL) and formulaic methods, indicating the basic textual similarity tasks are helpful to the VQA response evaluation task, but they fail to align well with human judgement, compared to SFVE results under the same structure of BERT-large. In addition, the performance on Part 1 and 3 of PLM models are similar, and the major gap lies in the capability of processing long responses.

5.3.3 Performance of LLMs. For LLMs (refer to Appendix for the detailed prompt) including ChatGPT, they fail to gain satisfactory results on AVE. Naturally, LLM performs better than RoBERTa w/o CL, and Generalization scores are slightly higher than PLMs, which we attribute to the better generalization ability of LLMs. Although LLMs obtain acceptable results on STS tasks, just like the formulaic methods, they encounter significant performance drop on the VQA response evaluation. Such phenomenon verifies the significant difference between the evaluation of STS and VQA responses and the necessity in the task of VQA evaluation.

The performance of embedding models (the latter two models) on STS is higher than LLMs but the VQA response evaluation performance is still low. The reason to their incompetence on AVE, as we speculate, is that these models focus more on retrieving and capturing the general meaning of given texts than discovering the fine-grained difference between given pair of texts. In addition, such focus of capturing the general meaning has also empowered them with the ability to ignore noise in morphology and text length, thus gaining high scores of Consistency despite the low Alignment scores.

5.3.4 Performance of SFVE. The section of SFVE (ours) in the table presents our results on AVE. The pretraining tasks effectively improve the Alignment scores of all three models and bring moderately better Consistency and Generalization performance. From the prospective of model sizes, the 125M Roberta-base demonstrates similar capability with the 355M Roberta-large with merely a gap of 0.8%. The same applies for the 7b LLAMA2, which surpasses RoBERTa-large by 0.9%. Giant increase in model sizes brings minor improvement in scores. We believe the reason is that the similarity measure, either in STS or AVE, is relatively simple for models to comprehend and implement, where a simple structure with limited parameters achieves excellent performance with proper training.

Therefore, for practical usage of evaluators during the training of generative VQA models, considering the significantly larger computation cost in LLAMA2-7b than RoBERTa, we recommend utilizing SFVE-RoBERTa base or large for a rough validation of model performance each certain steps or epoch, and use SFVE-LLAMA2 for more accurate evaluation near the best steps or epochs.

5.4 Ablation Experiments

To analyze the influence of each pretraining task, Table 2 provides ablation results by removing a pretraining task each time and by training on a single task alone. From the table it is clear that all pretraining tasks contribute to the final performance more or less.

The most important task is *Synonym and Antonym*, which causes a drop of 16.0% in Alignment scores on average and damages Consistency as well. In addition, when trained only on such data, the model performs the best. We believe the importance of training on *Synonym and Antonym* task lies in aligning the representation of synonyms and increasing the difference towards antonyms and other answers.

The second influencing pretraining task is *Generated descriptions*, without which the model can not directly learn to align the representation between semantically similar texts with different length. Yet the removal of it does not substantially damage the results on other parts than Part 2, which consists of long responses.

Meanwhile, the removal of *NLI data* matters almost the same as *Generated descriptions*. As mentioned before, NLI data focuses more on the coarse-grained meaning between text pairs while AVE requires a finer semantic discrimination. However, for a model that barely handles the task (shown in the row of *w/o All tasks*), we believe the easier data in NLI aid to fertilizing the basic capability in semantic evaluation. Yet the NLI data alone is insufficient, as shown in the row of *NLI data only*.

5.5 Practical Application

To demonstrate the practical values of our proposed evaluator in flexible VQA response evaluation, we collect responses of multiple MLLMs and compare the results with different evaluators by overall scores and case study.

As shown in Table 3, VQA score is clearly incompetent for assessing assorted responses from MLLMs. Since all responses from mPLUG-Owl are sentences, VQA Score even comes to 0. In the comparison of LXMERT and mPLUG-Owl, both BGE and SBERT indicate LXMERT generates better responses than mPLUG-Owl. However, taking the case study in Figure 4 into consideration, we









	<p>Question: What is this dish being cooked in?</p> <p>Answer: wok</p> <table><tr><th></th><th>SFVE (ours)</th><th>SBERT</th></tr><tr><td>LXMERT pan</td><td>79.4</td><td>83.6</td></tr><tr><td>mPLUG-Owl The dish is being cooked in a large black wok.</td><td>80.3</td><td>71.2</td></tr></table>		SFVE (ours)	SBERT	LXMERT pan	79.4	83.6	mPLUG-Owl The dish is being cooked in a large black wok.	80.3	71.2		<p>Question: How many people have skateboards?</p> <p>Answer: 2</p> <table><tr><th></th><th>SFVE (ours)</th><th>SBERT</th></tr><tr><td>LXMERT 2</td><td>100.0</td><td>100.0</td></tr><tr><td>mPLUG-Owl There are 2 people with skateboards in the image.</td><td>91.1</td><td>81.8</td></tr></table>		SFVE (ours)	SBERT	LXMERT 2	100.0	100.0	mPLUG-Owl There are 2 people with skateboards in the image.	91.1	81.8		<p>Question: What cultural clothing are the women wearing?</p> <p>Answer: japanese</p> <table><tr><th></th><th>SFVE (ours)</th><th>SBERT</th></tr><tr><td>LXMERT none</td><td>13.9</td><td>54.9</td></tr><tr><td>mPLUG-Owl The women in the image are wearing traditional chinese clothing.</td><td>47.7</td><td>57.5</td></tr></table>		SFVE (ours)	SBERT	LXMERT none	13.9	54.9	mPLUG-Owl The women in the image are wearing traditional chinese clothing.	47.7	57.5		<p>Question: What kind of resort are these people at?</p> <p>Answer: ski resort</p> <table><tr><th></th><th>SFVE (ours)</th><th>SBERT</th></tr><tr><td>LXMERT skiing</td><td>94.4</td><td>95.1</td></tr><tr><td>mPLUG-Owl These people are at a ski resort, as they are skiing down snow</td><td>93.6</td><td>87.0</td></tr></table>		SFVE (ours)	SBERT	LXMERT skiing	94.4	95.1	mPLUG-Owl These people are at a ski resort, as they are skiing down snow	93.6	87.0
	SFVE (ours)	SBERT																																									
LXMERT pan	79.4	83.6																																									
mPLUG-Owl The dish is being cooked in a large black wok.	80.3	71.2																																									
	SFVE (ours)	SBERT																																									
LXMERT 2	100.0	100.0																																									
mPLUG-Owl There are 2 people with skateboards in the image.	91.1	81.8																																									
	SFVE (ours)	SBERT																																									
LXMERT none	13.9	54.9																																									
mPLUG-Owl The women in the image are wearing traditional chinese clothing.	47.7	57.5																																									
	SFVE (ours)	SBERT																																									
LXMERT skiing	94.4	95.1																																									
mPLUG-Owl These people are at a ski resort, as they are skiing down snow	93.6	87.0																																									
(a)		(b)		(c)		(d)																																					
	<p>Question: What is the man in green's job?</p> <p>Answer: street cleaner</p> <table><tr><th></th><th>SFVE (ours)</th><th>BGE</th></tr><tr><td>LXMERT police</td><td>63.0</td><td>89.7</td></tr><tr><td>mPLUG-Owl The man in green is a garbage collector, as he is driving a truck.</td><td>74.5</td><td>87.7</td></tr></table>		SFVE (ours)	BGE	LXMERT police	63.0	89.7	mPLUG-Owl The man in green is a garbage collector, as he is driving a truck.	74.5	87.7		<p>Question: What brand is a competitor of the brand on the billboard?</p> <p>Answer: coke</p> <table><tr><th></th><th>SFVE (ours)</th><th>BGE</th></tr><tr><td>LXMERT pepsi</td><td>74.9</td><td>96.9</td></tr><tr><td>mPLUG-Owl A competitor of the brand on the billboard is pepsi.</td><td>73.2</td><td>91.9</td></tr></table>		SFVE (ours)	BGE	LXMERT pepsi	74.9	96.9	mPLUG-Owl A competitor of the brand on the billboard is pepsi.	73.2	91.9		<p>Question: What is the term for the brown marks on the bottom portion of the fire hydrant?</p> <p>Answer: rust</p> <table><tr><th></th><th>SFVE (ours)</th><th>BGE</th></tr><tr><td>LXMERT paint</td><td>75.0</td><td>95.4</td></tr><tr><td>mPLUG-Owl The brown marks on the bottom portion of the fire hydrant are called rust.</td><td>87.7</td><td>90.8</td></tr></table>		SFVE (ours)	BGE	LXMERT paint	75.0	95.4	mPLUG-Owl The brown marks on the bottom portion of the fire hydrant are called rust.	87.7	90.8		<p>Question: Why are the lights red with no traffic?</p> <p>Answer: late night</p> <table><tr><th></th><th>SFVE (ours)</th><th>BGE</th></tr><tr><td>LXMERT stop</td><td>47.3</td><td>91.5</td></tr><tr><td>mPLUG-Owl Because the street is empty and there are no vehicles present.</td><td>51.8</td><td>84.6</td></tr></table>		SFVE (ours)	BGE	LXMERT stop	47.3	91.5	mPLUG-Owl Because the street is empty and there are no vehicles present.	51.8	84.6
	SFVE (ours)	BGE																																									
LXMERT police	63.0	89.7																																									
mPLUG-Owl The man in green is a garbage collector, as he is driving a truck.	74.5	87.7																																									
	SFVE (ours)	BGE																																									
LXMERT pepsi	74.9	96.9																																									
mPLUG-Owl A competitor of the brand on the billboard is pepsi.	73.2	91.9																																									
	SFVE (ours)	BGE																																									
LXMERT paint	75.0	95.4																																									
mPLUG-Owl The brown marks on the bottom portion of the fire hydrant are called rust.	87.7	90.8																																									
	SFVE (ours)	BGE																																									
LXMERT stop	47.3	91.5																																									
mPLUG-Owl Because the street is empty and there are no vehicles present.	51.8	84.6																																									
(e)		(f)		(g)		(h)																																					

Figure 4: Cases for analysis. The samples come from the open-ended part of A-OKVQA [48] validation set. The first row comes from results of SFVE-large and SBERT, and the second comes from SFVE-large and BGE.

Model	Evaluation Metric			
	VQA Score	BGE	SBERT	SFVE (ours)
LXMERT [†] [51]	19.5	83.3	75.6	43.6
LXMERT	37.3	94.9	83.6	67.9
VisualBERT [31]	37.6	94.8	83.4	66.3
LLaVA-7b [35]	3.6	89.1	82.5	72.3
BLIP2-opt-2.7b [30]	15.5	94.1	83.1	70.2
InstructBLIP-Vicuna [18]	21.4	94.8	86.4	74.3
mPLUG-Owl [60]	0.0	91.0	82.7	69.1
OFA-large [55]	39.5	95.3	86.5	78.0
Qwen-VL-chat [8]	54.9	96.1	89.7	83.5

Table 3: Practical application of utilizing our proposed evaluator for assessing the responses from MLLMs. The VQA dataset for response generation is the open-ended validation set of A-OKVQA[48]. Models in the upper part of the table are smaller than 0.5B. VisualBERT and LXMERT are fine-tuned on VQA v2[22]. LXMERT[†] means the LXMERT that is not sufficiently trained, which ends training at the half of the first epoch to provide comparison. SFVE (ours) uses the RoBERTa-large evaluator trained with our proposed pretraining tasks. Refer to Appendix for the calculation of VQA Score. Note that the scores are for comparison within an evaluator itself, and it is meaningless to compare scores across evaluators, as evaluators are not aligned.

verify that existing well-performing methods, BGE and SBERT, fail to perform consistent evaluation and bias towards short responses while penalizing longer ones. For example, in (a) of Figure 4, LXMERT response *pan* receives a much higher score than the mPLUG-Owl response which is a descriptive sentence containing the correct answer. In (d), the descriptive text and the single word response receive similar scores under our SFVE, but SBERT considers the short answer of LXMERT is much better than the descriptive sentence of mPLUG-Owl. Similar phenomena exist in BGE as well. As in (e), mPLUG-Owl response describing *garbage collector* is a

better response than LXMERT output *police*, yet the latter receives even higher scores.

In addition, not only does the length impede a fair evaluation, but the incompetence in fine-grained semantics discrimination also causes absurd results. Like in (c) and (g), where LXMERT answers are less correct but they receive competitive or even higher scores than reasonable responses from BGE and SBERT. Such error, we speculate, is caused by focusing more on the overall meaning of text, as the questions are the same within a pair, while neglecting fine-grained difference.

Due to the phenomena above, it is clear the superficial superiority of LXMERT over mPLUG-Owl is merely a mistake by incompetent VQA evaluators, which also demonstrates the importance of fairness and consistency in VQA evaluation. Therefore, we consider the proposed pretraining tasks and SFVE effective, not only on our proposed AVE dataset, but also in practical application where previous methods fail to perform fair and insightful evaluation.

6 Conclusion

This paper proposes a practical task of utilizing semantic correctness to evaluate unconstrained open-ended VQA responses, facilitating the assessment of MLLMs' multimodal comprehension abilities by VQA data. We propose three key properties for assessing VQA evaluators, i.e., Alignment, Consistency and Generalization. In addition, this paper proposes a new dataset assessing VQA evaluators (AVE) to comprehensively analyze multiple aspects of evaluators. Based on contrastive learning with meticulously designed pretraining tasks, this paper provides a Semantically Flexible VQA Evaluator (SFVE) that performs significantly better than existing evaluators on VQA evaluation and the training scheme generalizes to both the encoder-only and decoder-only models.

References

- [1] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Inigo Lopez-Gazpio, Montse Maritxalar, Rada Mihalcea, et al. 2015. Semeval-2015 task 2: Semantic textual similarity, english, spanish and pilot on interpretability. In *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*. 252–263.
- [2] Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. Semeval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*. 81–91.
- [3] Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez Agirre, Rada Mihalcea, German Rigau Claramunt, and Janyce Wiebe. 2016. Semeval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *SemEval-2016. 10th International Workshop on Semantic Evaluation; 2016 Jun 16-17; San Diego, CA. Stroudsburg (PA): ACL; 2016. p. 497-511. ACL (Association for Computational Linguistics)*.
- [4] Eneko Agirre, Daniel Cer, Mona Diab, and Aitor Gonzalez-Agirre. 2012. SemEval-2012 Task 6: A Pilot on Semantic Textual Similarity.* SEM 2012: The First Joint Conference on Lexical and Computational Semantics—. In *Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, Montréal, QC, Canada. 7–8.
- [5] Eneko Agirre, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, and Weiwei Guo. 2013. * SEM 2013 shared task: Semantic textual similarity. In *Second joint conference on lexical and computational semantics (* SEM), volume 1: proceedings of the Main conference and the shared task: semantic textual similarity*. 32–43.
- [6] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.
- [7] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen Technical Report. *arXiv preprint arXiv:2309.16609* (2023).
- [8] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. *arXiv preprint arXiv:2308.12966* (2023).
- [9] Baichuan. 2023. Baichuan 2: Open Large-scale Language Models. *arXiv preprint arXiv:2309.10305* (2023). <https://arxiv.org/abs/2309.10305>
- [10] Satandeep Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. 65–72.
- [11] Manik Bhandari, Pranav Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. *arXiv preprint arXiv:2010.07100* (2020).
- [12] Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055* (2017).
- [13] Fu Chaoyou, Chen Peixian, Shen Yunhang, Qin Yulei, Zhang Mengdan, Lin Xu, Yang Jinrui, Zheng Xiawu, Li Ke, Sun Xing, Wu Yunsheng, and Ji Rongrong. 2023. MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. *arXiv preprint arXiv:2306.13394* (2023).
- [14] Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2023. BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. *arXiv:2309.07597* [cs.CL]
- [15] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.
- [16] Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. *arXiv preprint arXiv:1803.05449* (2018).
- [17] Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. *arXiv preprint arXiv:1705.02364* (2017).
- [18] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. *arXiv preprint arXiv:2305.06500* (2023).
- [19] Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics* 9 (2021), 391–409.
- [20] W Nelson Francis and Henry Kucera. 1979. Brown corpus manual. *Letters to the Editor* 5, 2 (1979), 7.
- [21] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821* (2021).
- [22] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6904–6913.
- [23] Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. *arXiv preprint arXiv:1804.11283* (2018).
- [24] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizviz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3608–3617.
- [25] Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 6700–6709.
- [26] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825* (2023).
- [27] Kushal Kafle and Christopher Kanan. 2017. An analysis of visual question answering algorithms. In *Proceedings of the IEEE international conference on computer vision*. 1965–1973.
- [28] Klaus Krippendorff. 2011. Computing Krippendorff’s alpha-reliability.
- [29] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision* 123 (2017), 32–73.
- [30] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv preprint arXiv:2301.12597* (2023).
- [31] 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* (2019).
- [32] Xianming Li and Jing Li. 2023. AnglE-optimized Text Embeddings. *arXiv preprint arXiv:2309.12871* (2023).
- [33] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.
- [34] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744* (2023).
- [35] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems* 36 (2024).
- [36] Shuo Liu, Kaining Ying, Hao Zhang, Yue Yang, Yuqi Lin, Tianle Zhang, Chuanhao Li, Yu Qiao, Ping Luo, Wenqi Shao, et al. 2024. Convbench: A multi-turn conversation evaluation benchmark with hierarchical capability for large vision-language models. *arXiv preprint arXiv:2403.20194* (2024).
- [37] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).
- [38] Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. (2017).
- [39] Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. *Advances in neural information processing systems* 27 (2014).
- [40] Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *9th Language Resources and Evaluation Conference*. Reykjavik.
- [41] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*. 3195–3204.
- [42] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography* 3, 4 (1990), 235–244.
- [43] Muennighoff Niklas, Su Hongjin, Wang Liang, Yang Nan, Wei Furu, Yu Tao, Singh Amanpreet, and Kiela Douwe. 2024. Generative Representational Instruction Tuning. *arXiv preprint arXiv:2402.09906* (2024).
- [44] Hyeonwoo Noh, Paul Hongsuck Seo, and Bohyung Han. 2016. Image question answering using convolutional neural network with dynamic parameter prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 30–38.
- [45] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.

- [46] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).
- [47] Mengye Ren, Ryan Kiros, and Richard Zemel. 2015. Exploring models and data for image question answering. *Advances in neural information processing systems* 28 (2015).
- [48] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VIII*. Springer, 146–162.
- [49] Wenqi Shao, Yutao Hu, Peng Gao, Meng Lei, Kaipeng Zhang, Fanqing Meng, Peng Xu, Siyuan Huang, Hongsheng Li, Yu Qiao, et al. 2023. Tiny lvm-ehub: Early multimodal experiments with bard. *arXiv preprint arXiv:2308.03729* (2023).
- [50] Chuangqi Tan, Furu Wei, Wenhui Wang, Weifeng Lv, and Ming Zhou. 2018. Multi-way attention networks for modeling sentence pairs. In *IJCAI* 4411–4417.
- [51] Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490* (2019).
- [52] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shriti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).
- [53] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence* 40, 10 (2017), 2413–2427.
- [54] Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony Dick. 2015. Explicit knowledge-based reasoning for visual question answering. *arXiv preprint arXiv:1511.02570* (2015).
- [55] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*. PMLR, 23318–23340.
- [56] Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426* (2017).
- [57] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. <https://www.aclweb.org/anthology/2020.emnlp-demos.6>
- [58] Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, and Anton Van Den Hengel. 2017. Image captioning and visual question answering based on attributes and external knowledge. *IEEE transactions on pattern analysis and machine intelligence* 40, 6 (2017), 1367–1381.
- [59] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 21–29.
- [60] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178* (2023).
- [61] Wenpeng Yin, Hinrich Schütze, Bing Xiang, and Bowen Zhou. 2016. Abcmn: Attention-based convolutional neural network for modeling sentence pairs. *Transactions of the Association for computational linguistics* 4 (2016), 259–272.
- [62] Kaining Ying, Fanqing Meng, Jin Wang, Zhiqian Li, Han Lin, Yue Yang, Hao Zhang, Wenbo Zhang, Yuqi Lin, Shuo Liu, et al. 2024. Mmt-bench: A comprehensive multimodal benchmark for evaluating large vision-language models towards multitask agi. *arXiv preprint arXiv:2404.16006* (2024).
- [63] Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490* (2023).
- [64] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).

A Traditional VQA Evaluation Metrics

Traditional VQA evaluation metrics contain Exact Match [39] and VQA Score [6]. They apply for different settings in VQA datasets. For datasets where each sample contains only one correct answer, like DAQUAR [39], TDIUC [27], GQA [25], the Exact Match is used. If each sample contains ten candidate answers, like VQA v2 [22], OKVQA [41], VizWiz [24], VQA Score is commonly used.

Exact Match. Exact Match calculates by judging whether the response is identical to the annotated ground-truth answer, and if matches, the score is 1, otherwise 0.

VQA Score. VQA Score evaluates how many times the response appear in the ten candidate answers, and is computed as follows:

$$accuracy = \min\left(\frac{\# \text{ correct hits}}{3}, 1\right)$$

As there are ten candidate answers, # correct hits represents numbers of matched answers, which means as long as there are three or more candidates are the same with the predicted answer, the answer will be considered fully correct, and gets a score of 1.

B Prompt for Decoder Models

The following is the similarity calculation prompt provided to the LLMs and ChatGPT in the experiments:

Sentence similarity evaluation here refers to the task of measuring the semantic similarity score between two sentences. For example, "what a good day" and "how nice the weather is" are almost the same, your output shall be {"score":0.91}. Now please evaluate the similarity score between the following two sentences: sentence1: sample["sentence1"]. sentence2: sample["sentence2"]. The score shall range continuously from 0-1. DO NOT output anything else but the .json-style dictionary, like {"score":x}, where x is your predicted score.

The sample["sentence1"] and sample["sentence2"] indicate a pair of texts for similarity calculation. The question and answer are concatenated with the prompt "Question: {question} Answer: {answer}" before similarity calculation, importing contextual information, just the same as other models in experiments.

ChatGPT prompt for converting each question-answer pair into a description:

Concatenate the question with the answer and form assertions. For example, Question:What kind of dog is in the photo? Answer:golden retriever. Assertion: The dog in the photo is a golden retriever. Infer for the following: Question: {question} Answer: {answer}. Please think of three different forms of naturally-sounded assertions for this question-answer pair with small disturbance but do not output them. Choose the two assertions that are closest in meaning to the original question-answer for output. Output shall be in .json style so that I can directly save them in a .txt and open by json. Do not output anything else including explanation, reasoning or instructions.

C Annotation Score Distribution

The score distribution of the dataset Assessing VQA Evaluators (AVE) is shown in Figure 5. The annotation covers all scores.

D Selection of Annotating Aspect

For the evaluation of the quality of responses, we considered adopting multiple aspects for analyzing. The text summary task [11, 19, 23] adopts four aspects, i.e., *relevance*, *consistency*, *fluency* and *coherence* for analyzing the generated summary based on the reference text. As VQA responses are generally short, the *fluency* and *coherence* that measure the fluency of the text is less necessary. The aspect of *consistency* measures whether the generated summary contains hallucination that generates untrue information. In the

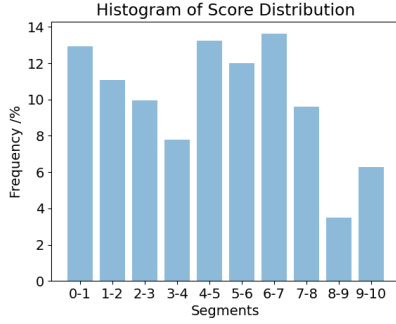


Figure 5: The annotated scores distribution of AVE.

scene of VQA response evaluation, such measure corresponds to the overlap in semantics from the response towards the ground-truth answer, which is similar to *relevance* that measures how well the generated text captures the key points. Therefore, we decide to use semantic similarity as the only score of annotation.

E Annotation Rules

E.1 Scores Annotation

All of the following in this section is the annotation rules provided to the three annotators during the annotation of scores.

Scoring Format: Discrete integer scoring from 0 to 10 (e.g., 0, 1, 2, 3).

Scoring Rules:

Note that it is not about judging whether the response to the question is correct, but whether the response and the standard answer are semantically the same under the question (For example, even if the standard answer is obviously unreasonable, as long as the answer and response are semantically similar, a high score shall be given).

Semantically similar (the answer is fairly correct in meaning) but different in specific form (for example, the meaning expressed is similar but different in word choice, tense, number), score 6-10 based on the degree of semantic similarity. Examples:

(1) Question: What is the last letter on the license plate? Standard Answer: letter j Response: j Scoring (this is a reasonable range, just mark a specific score when actually annotating): 9-10 Reason (no need to mark, this is to help understand the rules): Under this question, the response and the standard answer are semantically the same, only the form is different. Similarly, synonyms should also be scored highly.

(2) Question: The young man above the swimming pool is wearing what? Standard Answer: swimsuit Response: trunks Scoring: 7-8 Reason: The question asks what is being worn, and swimsuit (swimwear) includes trunks (swim shorts), so the answer is quite correct. This kind of inclusive or included relationship should be scored based on the semantic similarity of the two words. In addition, trunks as swim shorts is a less common meaning and requires more attention to the different meanings of words, not entirely based on experience.

(3) Question: How many trees are there? Standard Answer: 3 Response: three Scoring: 10 Reason: The meanings are exactly the same.

Semantically dissimilar, the answer is incorrect, but the answer is a possible answer for that type of question, score 1-5 based on the degree of semantic similarity. Examples:

(1) Question: What color do you think the trousers the boy is wearing have? Standard Answer: white Response: blue Scoring: 2-3 Reason: The question asks about color, and although the answer blue is different from the standard answer white, both are common answers under the category of color questions. Furthermore, if it's white and black, the difference between the two is greater than between white and blue, so the scoring range should be further reduced to 1-2.

Semantically dissimilar, but the answer and the standard answer mean the same under the question, then score 4-7 based on the correctness. Examples:

(1) Question: What is lit? Standard Answer: cake Response: candle Scoring: 5-7 Reason: Although cake and candle are very different semantically, in this question, they actually mean the same thing. Therefore, one should not only look at the answer but also focus on the question.

For numerical type answers, score 1-8 based on how much the number in the standard answer and the number in the response differ. Examples:

(1) Question: How many trees are there? Standard Answer: 4 Response: 5 Scoring: 3-5 Reason: Although 4 and 5 are different, the difference between them is not particularly large. If the standard answer is still 4, but the response becomes 1, then the scoring should be appropriately lowered to 1-3. If the standard answer is 70, and the response is 75, then it can be considered quite correct, scoring 5-7. If the standard answer is 1 and the response is 0, then score 1-2. Judge the score based on whether the numbers are relatively close to each other.

(2) Question: When did this accident happen? Standard Answer: 1945 Response: 1940 Scoring: 5-7 Reason: The two years are quite close. But if the answer becomes the 1940s (i.e., 1941-1949), it includes 1945, and the range is not particularly large, so it is quite correct, scoring 6-8.

For answers with significantly different meanings, score based on semantic similarity without range restrictions. Examples:

(1) Question: What color bathing suit is the woman wearing? Standard Answer: no woman Response: red Scoring: 0-1 Reason: The meanings are very different.

E.2 Manual Filter

The following is the annotation rules provided to the three annotators during the last stage, manual filter, in the construction of our AVE dataset.

Read the question, answer, response and all augmentation results carefully, and decide whether the augmentation has changed the original meaning. Labels shall be in yes, no, unsure.

F Answer Templates

In the fourth step of constructing the proposed dataset, i.e., description generation, beside collecting ChatGPT-transformed results, we augment each short response with manual written templates: (1) Answer: [response]., (2) The answer to this question is [response]., (3) As shown in the image and question, the answer is [response]., (4) The

answer you are asking for is {response}., (5) As can be deduced from the image, the answer to this question is {response}., (6) The answer to your question appears to be {response}, as shown in the image..

Received 13 April 2024; accepted 21 July 2024