TokenFocus-VQA: Enhancing Text-to-Image Alignment with Position-Aware Focus and Multi-Perspective Aggregations on LVLMs

Anonymous CVPR submission

Paper ID 92

Abstract

While text-to-image (T2I) generation models have achieved remarkable progress in recent years, existing evaluation methodologies for vision-language alignment still struggle with the fine-grained semantic matching. Current approaches based on global similarity metrics often overlook critical token-level correspondences between textual descriptions and visual content. To this end, we present TokenFocus-VQA, a novel evaluation framework that leverages Large Vision-Language Models (LVLMs) through visual question answering (VQA) paradigm with position-specific probability optimization. Our key innovation lies in designing a token-aware loss function that selectively focuses on probability distributions at pre-defined vocabulary positions corresponding to crucial semantic elements, enabling precise measurement of fine-grained semantical alignment. The proposed framework further integrates ensemble learning techniques to aggregate multi-perspective assessments from diverse LVLMs architectures, thereby achieving further performance enhancement. Evaluated on the NTIRE 2025 T2I Quality Assessment Challenge Track 1, our TokenFocus-VQA ranks 2nd place (0.8445, only 0.0001 lower than the 1st method) on public evaluation and 2nd place (0.8426) on the official private test sets, demonstrating superiority in capturing nuanced text-image correspondences compared to conventional evaluation methods.

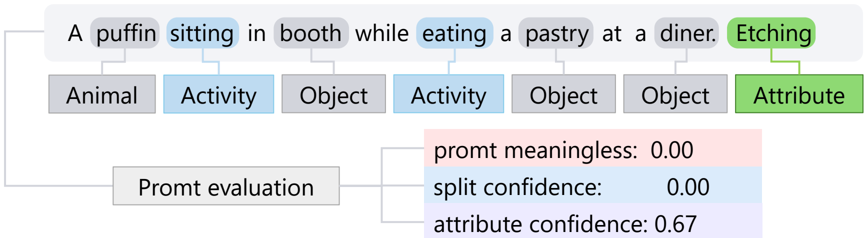
1. Introduction

The remarkable progress in text-to-image (T2I) generation has fundamentally transformed creative workflows, yet simultaneously exposed critical gaps in evaluation methodologies. As the generative models achieve unprecedented photorealism, the research community faces growing challenges in systematically assessing fine-grained alignment between textual descriptions and visual content, which is a capability essential for model refinement and real-world application deployments.

Traditional evaluation paradigms have evolved from holistic quality metrics like FID [16] and IS [40] to specialized benchmarks probing specific capabilities. Frameworks such as T2I-CompBench [19] systematically evaluate colors, shapes, or texture binding through the structured prompts, while REAL [30] evaluates visual authenticity across attributes, relationships, as well as styles. Emerging knowledge-intensive evaluations like T2I-FactualBench [20] further verify scientific and historical accuracy, with Winoground-T2I [50] examining compositional sensitivity through the contrastive examples. As the NTIRE 2025 competition, which is based on the EvalMuse-40k dataset [14], has emerged, element existence verification becomes more focused than ever before, aiming to develop specific models that can evaluate detailed image-text alignment scores more consistent and accurate with human preferences. A detailed data use case is illustrated in the Fig. 1 below.

Current technical approaches for alignment assessment primarily reveal three distinct evolutionary paths. (**I**) Global similarity metrics like CLIP Score [15] and BLIP Score [28] compute image-text embedding correlations but fail to capture the token-level correspondences. (**II**) Cross-modal attention mechanisms in SCAN [24] and ALBEF [27] successfully improve the localization capabilities through feature alignment, yet still struggle with the positional binding verification. (**III**) The recent paradigm shift toward VQA-based evaluation, exemplified by TIFA [18] and contemporary works [25], converts alignment assessment into question-answering (QA) tasks but critically overlooks probability distributions at the semantically crucial token positions. This critical limitation arises from the reliance of current methods on binary (yes/no) classification outputs, which discard crucial confidence information embedded in LVLMs'outputs—especially at the vocabulary positions corresponding to key objects and attributes, as extensively demonstrated in BLIP2 [29] and FGA-BLIP2 [14].

The above observations reveal the fundamental limitations in modern evaluation frameworks: the underutilization of position-specific probability signals, uniform treat-



	SDXL-Turbo	Playground-v2.5	Kandinsky3	IF	PixArt-Sigma
Total Score	3.33	3.33	3.00	3.00	3.00
Puffin (Animal)	0.67	1.00	1.00	0.67	1.00
Sitting (Activity)	1.00	1.00	0.67	0.33	1.00
Booth (Object)	1.00	1.00	1.00	0.67	1.00
Eating (Activity)	0.0	0.00	0.00	0.67	0.00
Pastry (Object)	0.67	0.67	0.67	0.0	0.00
Diner (Object)	1.0	1.00	0.67	0.67	1.00
Etching (Attribute)	0.33	0.67	0.67	1.0	0.67

	PixArt-Alpha	SSD1B	SDXL	SD-v1.5	SD-v1.2
Total Score	3.00	3.00	3.00	2.67	3.00
Puffin (Animal)	0.67	1.00	0.67	1.00	0.33
Sitting (Activity)	0.67	1.00	0.67	1.00	0.67
Booth (Object)	0.33	1.00	0.67	0.33	0.67
Eating (Activity)	0.0	0.00	0.33	0.33	0.67
Pastry (Object)	0.33	0.33	0.33	0.67	0.33
Diner (Object)	0.33	1.00	0.67	0.67	0.33
Etching (Attribute)	0.67	0.67	0.67	0.67	0.67

Figure 1. Actual use case demonstration of the EvalMuse-40K in the NTIRE 2025 competition. Different types of elements are marked with special colors (*i.e.*, ■ for object elements, ■ for action elements, and ■ for item attributes). The total score is classified into 1-5, and the element-level score is 0 and 1. The values shown in the tables above are the averaged results of three or six annotators.

ment of all vocabulary items during the loss calculation, and inherent biases in single-model assessments. Our heuristic analysis of the VSE++ [13] and CLIP [38] architectures further demonstrates how standard similarity metrics dilute focus on critical semantic elements during global aggregations. This technical landscape motivates TokenFocus-VQA, a framework that reimagines VQA-based evaluation through targeted probability optimization on LVLMs. Overall, the main contributions of this paper are three-fold:

- We first introduce the **T**oken-**F**ocus supervised and **P**osition-**S**pecific loss function to promote LVLMs fine-tuning, thereby leading to significant improvements in fine-grained image-text matching.
- We then propose a newly optimized ensemble framework to perform multi-perspective aggregations, which integrates Bagging, Stacking and Blending, to overcome the limitations of single LVLMs in image-text evaluation.
- Extensive experimental results demonstrate the effectiveness of our proposed TokenFocus-VQA, exhibiting impressive superiority on the EvalMuse-40K dataset and the NTIRE 2025 compitition test bed.

2. Related Works

2.1. Image-Text Alignment

With the exponential growth of multimedia content on the Internet, cross-modal image-text matching has emerged as a fundamental task in information retrieval [45, 48], social media analysis [31, 49], and intelligent recommendation systems [12, 37, 39]. Prior to the deep learning era, image feature extraction predominantly relied on the hand-crafted descriptors such as SIFT and SURF [6, 35]. While these methods demonstrated certain effectiveness, they suffered from limited generalization capabilities and poor adaptability to complex scenarios. The rapid advancement of deep learning has revolutionized feature extraction paradigms for both visual and textual modalities. Pioneering works like VSE++ [13] have established baseline frameworks by optimizing cosine similarity loss between cross-modal feature representations. Subsequent methods introduced finer-grained alignment mechanisms, exemplified by SCAN [24] with its stacked cross-attention modules. The introduction of dual-stream architectures reached a milestone with ViLBERT [36], which extended the BERT [22] pretraining paradigm to the multimodal domain through the masked multimodal data modeling.

The paradigm shift towards large-scale pre-training has yielded groundbreaking models like CLIP [38], which leveraged contrastive learning on 400M image-text pairs to achieve SOTA zero-shot cross-modal retrieval capabilities. Building upon visual transformer architectures, ViLT [23] pioneered a unified transformer framework that directly processes image patches and text tokens, enabling efficient cross-modal fusion. To reconcile architectural flexibility with performance, VLMO [5] proposed a mixture-of-modality-experts approach supporting both unimodal and multimodal tasks through task-specific expert modules. To tackle the data quality challenges, the BLIP series [28] introduced novel architectures combining understanding and generation capabilities. The Q-Former module of BLIP-2 [29] has achieved state-of-the-art visual reasoning performance through efficient cross-modal interaction learning, while maintaining computational efficiency by freezing pre-trained vision-language backbones [29].

2.2. Large Vision-Language Models

In recent years, Large Vision-Language Models (LVLMs) have made significant progress in the field of multimodal understanding by integrating large-scale pretrained language models with specific vision encoders. For example, LLaVA series [2, 33, 34] models achieve precise image-text matching by directly connecting the CLIP vision encoder with the backend language model LLaMA [42] through end-to-end visual instruction tuning [32]. InternVL [10], by constructing a vision encoder with 6 billion parameters (ViT-6B) aligned with the language model, has achieved parameter balance between the vision and language branches for the first time, thereby overcoming the modality gap in cross-modal feature fusion [11]. Meanwhile, GPT-4V [1] and Gemini [41], through large-scale parameter size and multimodal instruction tuning, can support complex visual reasoning tasks [1, 41]. In terms of fine-grained and dynamic modeling, LLaVA-NeXT [26] expands capacity to a scale of 34 billion parameters, supports input with 4 $\times$ pixel resolution, and achieves general understanding across images and videos through multitask joint training. InternVL-2.5 [9] proposes a dynamic resolution adaptation strategy, supporting multi-scale image input resolution from 224 pixel to 1024 pixel, and achieves semantic consistency across resolutions through a feature pyramid network [9]. Qwen-VL [3] introduces a textual encoding strategy for the bounding boxes, enabling spatial position awareness through extensive text labeling [4].

Compared to traditional multimodal models, LVLMs demonstrates significant advantages in image-text alignment tasks. By employing end-to-end semantic fusion architecture and dynamic computation optimization, LVLMs has the capability to overcome the reliance of traditional models on fixed resolution input and manual feature engineering, achieving SOTA fine-grained semantic alignment across languages and scales [47]. LVLMs can support multimodal autonomous reasoning, adaptive token compression, as well as zero-shot transfer learning, significantly enhancing alignment accuracy and robustness in complex scenarios such as occlusion, abstract metaphors, and multi-object interactions. In addition, through a multi-stage reasoning mechanism, LVLMs improve the efficiency of high-resolution image processing, providing more efficient solutions for practical applications such as cross-modal retrieval and multilingual matching [21, 43].

3. Methodology

In this section, we introduce the T2I alignment enhanced evaluation method termed TokenFocus-VQA based on LVLMs for both holistic and fine-grained level matching. Only by deploying VQA and applying token-level supervised loss calculation during supervised fine-tuning (SFT), accurate image-text matching evaluation and recognition can be achieved at various granularities.

The overall framework of our proposed TokenFocus-VQA system, as illustrated in Fig. 2, builds upon the established paradigm of VQA while introducing several critical innovations. First of all, the image and structured query are encoded into visual and textual tokens, which are then generatively understood and predicted by pre-trained LVLMs. Our key innovation emerges in the answer generation phase. Contrary to standard VQA approaches that consider complete answer sequences, we implement **Token-**

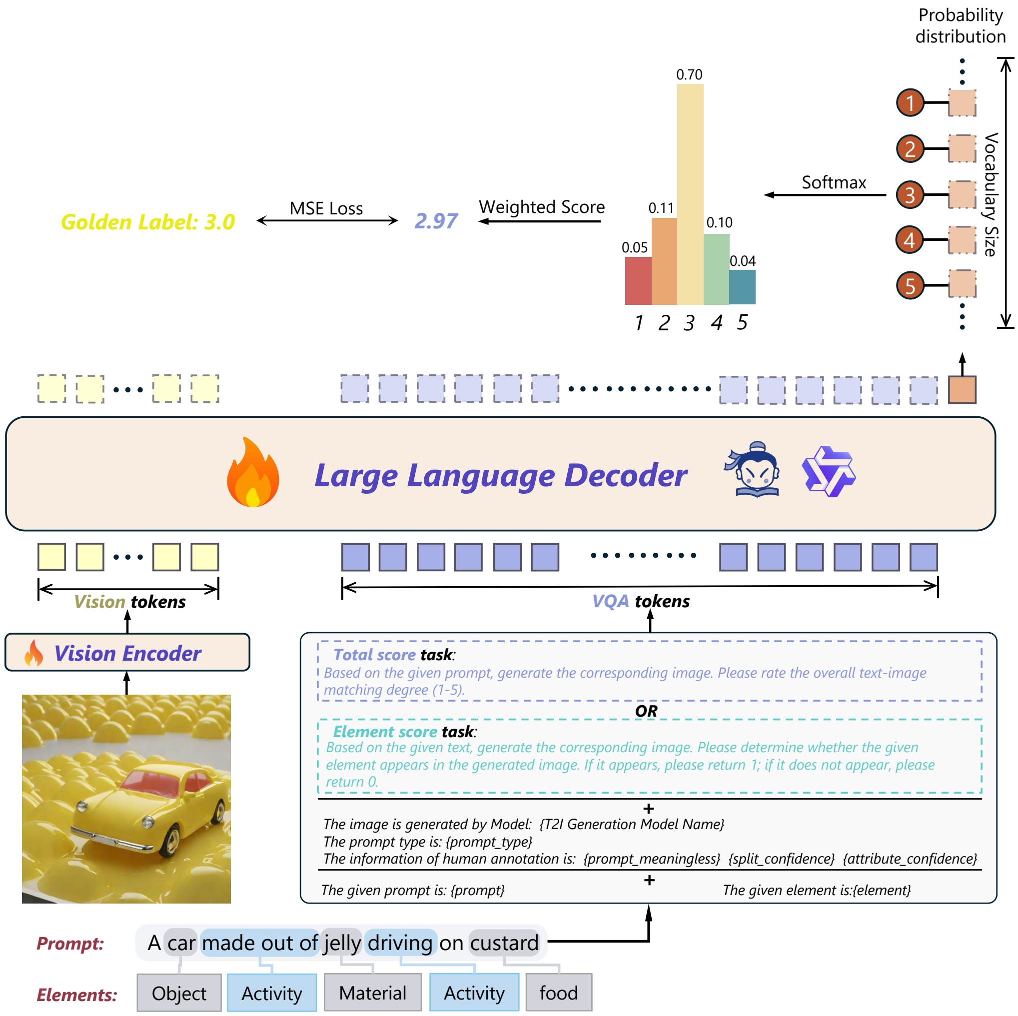


Figure 2. The overall framework of our proposed TokenFocus-VQA, which is proposed for LVLMs-based T2I alignment accessment at both the holistic and fine-grained levels. The visual encoding process begins with transforming input images into the visual tokens via a vision encoder. For distinct scoring tasks (*i.e.*, *Total Score* & *Element Score*), we construct task-specific input prompts augmented with the structured meta-data. These multimodal tokens are then jointly processed in the large language decoder (*i.e.*, InternLM[7] and Qwen2.5[46]) for the generative score prediction. The framework is ultimately refined through our proposed Position-Aware Token-Focused Optimization method for more performance gains.

Focus, a strategic emphasis on the **first generated token** under strictly controlled output formatting. We then integrate the loss calculation method of numerical regression model for the label prediction dimension of language model. Additional multi-model integration (including serial stacking, parallel bagging, as well as hybrid blending), targeted learning rates of language and vision modalities, and other methods are also applied to enhance performance.

3.1. Position-Aware Token-Focused Optimization

Language models (LMs) generate probabilistic outputs via softmax, which inherently conflicts with deterministic regression tasks (*i.e.*, numerical scoring) requiring focus on specific tokens. Standard cross-entropy supervision forces probability mass allocation across all tokens, diluting learning signals and slowing convergence. To address this, we propose **token-focus** supervision, re-weighting the loss to concentrate on task-critical tokens. This filters extraneous noise and transforms probabilistic training into value-driven optimization, directly aligning LM generation with continuous regression metrics. Specifically, we only focus on the first generated token and obtain the **position-aware** probability of the label corresponding to the score ([0, 1] or [1, 2, 3, 4, 5]) in its predicted distributions. After normalization, we then multiply the probability of the corresponding label by the score weight to obtain the LVLMs-based regression

Extra Information	Detailed Descriptions
T2I Model Name	The specific **model** used to perform image generation.
Prompt Type	The **real user prompts**, which are extensively collected from DiffusionDB [44], as well as the **synthetic prompts**.
Prompt Evaluation	The data includes **manually annotated fields**, which are deployed to evaluate both the **prompt quality**, assessing its semantic clarity and generability, and the **division clarity** of fine-grained alignment targets, including segmentation and attribute confidence.

Table 1. The detailed descriptions of the external structural information.

or classification results, and deploy MSE (Mean-Square Error) and other methods to calculate the loss accordingly.

Let the language model vocabulary be V, the target score set be $S = \{s_1, s_2, \dots, s_k\}$ ($[1, 2, \dots, 5]$ for element score tasks, [1, 2] for total score task, k for score label nums). Given an input image-text pair X, the model's original probability distribution for the first generated (t = 1) token can be formulated as:

$$p_{t=1}(w|X) = \text{softmax}(z_w), \quad \forall w \in V, \tag{1}$$

where z_w represents the output value of token w from the last output linear layer. After filtering irrelevant tokens, we can get the conditional probability distribution after normalization (*i.e.*, Softmax function):

$$P(s_i) = \frac{exp(p_{t=1}(s_i|X))}{\sum_{j=1}^{k} exp(p_{t=1}(s_j|X))}, \quad s_i \in S. \tag{2}$$

This operation projects the original probability space into the target score space to eliminate potential noise interference. Then the discrete-to-continuous conversion of predicted value $\hat{y}$ is achieved through the expected value mapping, *i.e.*,

$$\hat{y} = \mathbb{E}_{s \sim P}[s] = \sum_{i=1}^{k} s_i P(s_i). \tag{3}$$

The MSE is deployed to directly optimize the gaps between the predicted value and true value, the loss of any task($\mathcal{L}_{\text{task}}$) can be calculated as:

$$\mathcal{L}_{\text{task}} = (\hat{y}_n - y_n)^2, \tag{4}$$

where $\hat{y}_n$ and y_n refers to the final predictions and ground-truth, respectively.

3.2. External Structural Information Integration

Considering that prompt engineering can effectively enhance the performance of Large Language Models (LLMs) on specific tasks, and inspired by the common practice of leveraging additional features to improve recognition accuracy in machine learning, we propose a structured prompt construction method specifically designed for image-text

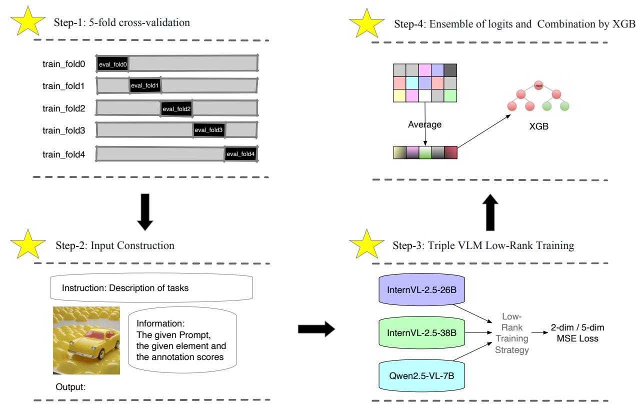


Figure 3. The overall illustration of our ensemble training and inference workflow.

pairs in VQA tasks. It systematically incorporates textualized external information as engineered prompt features, as illustrated in Tab. 1. Our feature augmentation is motivated by two considerations, which are detailed as below:

- Given the potentially significant disparities in generative capabilities across different model architectures, we further integrate detailed model-specific information into the prompts to compensate for model discrepancies.
- The quality of the generated prompt itself will directly affect the effects of subsequent generation and interfere with the model's understanding of complex or ambiguous language. The spatial description of fine-grained elements that need to be judged will also affect the modeling ability of complex scenes. The accuracy of attributes directly guides the upper limit of model detail evaluation.

3.3. Data Sub-packaging and Model Ensemble

We introduce a novel hierarchical ensemble architecture, as displayed in Fig. 3, that systematically integrates ensemble skills to establish multi-strategy consensus formation, effectively addressing heterogeneous representation learning bottlenecks and distributional bias inherent in singular LVLM for cross-modal evaluation tasks. Our training procedure consists of three stages, *i.e.*, (**I**) We partition the dataset into five folds, allocating 80% for training and 20% for validation within each fold. Our guiding principle is to eliminate duplicate prompts, while allowing image genera-

Configurations	Value
Optimizer	AdamW
LoRA rank	64
LoRA alpha	128
Base learning rate	1e-4
Vision learning rate	1e-5
Weight decay	0.05
Optimizer momentum	$\beta_1, \beta_2 = 0.9, 0.95$
Global batch size	64
Learning rate schedule	cosine decay
Warmup steps	200
Seed	1234
Epoch	3

Table 2. The training settings of our proposed TokenFocus-VQA. To faithfully ensure consistency, the same training setup are deployed for our LVLMs of varying architectures and scales.

tion models to overlap across folds, aligning with the testing data distribution. (**II**) We modify the configuration of the model input, utilizing the prompt construction method mentioned in Sec. 3.2 to integrating certain statistical features directly as textual inputs. (**III**) For the testing procedure, we utilize various models to predict the test data by deploying checkpoints derived from the training phase across different folds. Subsequently, we employ XGBoost [8] to integrate the predicted scores with selected statistical features, jointly consolidating them into the final predictions.

The labels in the dataset represent averages from multiple annotators, and utilizing them directly without bucketing has shown superior results for the overall score task. For the element score task, the variation between employing MSE and cross-entropy is minimal. Our empirical analysis reveals that while full-parameter fine-tuning achieves modest gains (+0.5 pp) in localized 5-fold validation, but exhibits critical generalization deficits (-1.2% SRCC) on the leaderboard, suggesting inherent limitations in data-constrained scenarios. This observation motivates our adoption of LoRA [17] adaptation.

4. Experiments

4.1. Implementation Details

We split the overall data into 5 non-overlapping folds, selecting four folds for training each time and the rest for cross-validation, making sure there is no data overlap between prompts when splitting. We deploy Qwen2.5-VL [4] and InternVL-2.5 [9] as the baseline models to perform extensive model training.

Training Settings: Different learning rates are set for the vision encoder and LM decoder layer to balance the emphasis on visual understanding and task instruction following. The LoRA [17] is applied for efficient fine-tuning while conducting extensive training experiments on different models of varying sizes. The specific training parameters are shown in Tab. 2. All the experiments are conducted on the *NTIRE 2025 Text to Image Generation Model Quality Assessment Challenge* Track 1 dataset using a machine with 8 × NVIDIA A100 GPUs, with respective training durations of 7 hours (Total Scoring) and 25 hours (Element Scoring) under a standardized 5-fold cross-validation protocol with independent optimization across splits.

Method	Visual Enc.	Text Enc.	PLCC (↑)	SRCC (↑)
Qwen2.5-VL-7B [4] (VQA)	660M	7B	0.6796	0.6783
InternVL-2.5-4B [9] (VQA)	300M	4B	0.6922	0.7054
CLIP-Score [15]	88M	63M	0.3023	0.2975
BLIPv2Score [15]	300M	2.7B	0.3621	0.3381
FGA-BLIP2[14]	300M	2.7B	0.7754	0.7741
Qwen2.5-VL-7B [4]	660M	7B	0.7962	0.8020
Qwen2.5-VL-32B [4]	660M	32B	0.7977	0.8007
InternVL-2.5-4B [9]	300M	3B	0.7988	0.8025
InternVL-2.5-8B [9]	300M	7B	0.8003	0.8046
InternVL-2.5-26B [9]	6B	20B	0.8096	**0.8141**
InternVL-2.5-38B [9]	6B	32B	**0.8098**	0.8133

Table 3. The experimental results of different LVLMs using one fold data without extra structural information. VQA stands for applying VQA method on LVLMs and Enc. refers to Encoder, which is deployed to compare the size of Vision encoder and Text encoder. InternVL-2.5 [9] series features multiple vision encoder variants (*i.e.*, 300M and 6B parameters) for scalable deployment, while the Qwen2.5-VL [4] series maintains architectural uniformity with a fixed 660M visual encoder across all configurations.

Fold	SRCC (↑)	PLCC (↑)	ACC (↑)
= 1	**0.8371**	**0.8313**	82.35%
= 2	0.8213	0.8184	82.06%
= 3	0.8175	0.8144	**82.40%**
= 4	0.8272	0.8226	82.32%
= 5	0.8163	0.8122	81.72%

Table 4. The performance comparisons of 5-fold cross validation using InternVL-2.5-26B [9].

Evaluation Settings: For the overall alignment scores, we report the Spearman Rank Correlation Coefficient (SRCC) and Pearson Linear Correlation Coefficient (PLCC) to measure the correlation between model predictions and human annotations. We further conduct fine-grained elements evaluation by reporting the accuracy (ACC) of the output predictions. To further ensure equitable weighting of both holistic alignment measurements and granular element matching in the comprehensive evaluation framework, we formulate the composite evaluation metric through the following mathematically formalized weighted integration:

$$O = 0.25 \times S + 0.25 \times P + 0.5 \times A, \quad (5)$$

where S, P, and A represent SRCC, PLCC, and ACC, respectively. O denotes the final composite evaluation metric.

CVPR
#92

ID	Method	Evaluation Bed	SRCC	PLCC	ACC	Overvall
0	Qwen2.5-VL-7B	Cross Validation Leaderboard	0.8256 –	0.8205 –	0.8252 –	0.8241 –
1	InternVL-2.5-26B	Cross Validation Leaderboard	0.8258 0.7839	0.8198 0.8125	0.8217 0.8509	0.8223 0.8245
2	InternVL-2.5-38B	Cross Validation Leaderboard	0.8273 –	0.8232 –	0.8226 –	0.8239 –
3	Qwen2.5-VL-7B + InternVL-2.5	Cross Validation Leaderboard	– 0.8002 (+0.0163)	– 0.8321 (+0.0196)	– 0.8619 (+0.0110)	– 0.8390 (+0.0145)
4	Qwen2.5-VL-7B + InternVL-2.5 + SF	Cross Validation Leaderboard	– 0.8002 (+0.0163)	– 0.8321 (+0.0196)	– 0.8691 (+0.0182)	– 0.8426 (+0.0181)

Table 5. The performance comparisons of different methods in terms of SRCC, PLCC, ACC, as well as Overvall metrics on both 5-fold Cross Validation and Leaderboard evaluation beds. InternVL-2.5: InternVL-2.5-26B & InternVL-2.5-38B, SF: Statistic Features. *Green* refers to the baseline results for longitudinal comparison, representing the non-ensemble learning-enhanced approach. *Red* denotes the performance enhancement.

Fold	T-Samples	E-Samples
= 1	26,191	6,526
= 2	26,099	6,618
= 3	26,164	6,553
= 4	26,184	6,533
= 5	26,245	6,472

Table 6. The detailed informaction on data fold splitting. T and E denote Training and Evaluation, respectively. The overall data is devided according to the unique prompt ID, maintaining a $4:1$ ratio (2,393 : 598) of unique prompts between training and evaluation sets in each fold.

4.2. Overall Performance Comparisons

We establish comprehensive comparative baselines utilizing FGA-BLIP2 [14], CLIP-Score [15], and BLIPv2Score [15]. As for our experimental protocol initiates with preliminary validation on a held-out validation fold to benchmark performance variations across model architectures and scales. We evaluate two top-performing open-source models (*i.e.*, Qwen2.5-VL and InternVL-2.5) across parameter scales ranging from 4B to 38B. Besides, We conducted controlled experiments employing VQA-typical implementations on LVLMs, to verify the methodological superiority of our proposed approach.

As illustrated in Tab. 3 above, our TokenFocus-VQA method demonstrates statistically significant superiority over both conventional VQA approaches and the SOTA FGA-BLIP2 [15]. InternVL-2.5 consistently outperforms its counterpart in cross-modal alignment accuracy at comparable parameter scales. Remarkably, increasing LVLMs parameters through decoder expansion demonstrates negligible performance impact (*e.g.*, Qwen2.5-VL-32B shows minimal PLCC improvement with SRCC degradation, a pattern replicated in InternVL-2.5 variants). This phenomenon underscores the decisive role of vision encoder capacity – scaling InternVL-2.5's vision encoder from 300M to 6B parameters yields $\approx 1\%$ absolute performance improvement, revealing the vision-centric scaling laws in multimodal systems on image-text alignment evaluation task.

4.3. Performance of Different Folds

The 5-fold cross-validation protocol, with stratified partitioning by prompt IDs (each corresponding to multiple annotated samples from different text-to-image models), is detailed in Tab. 6 above. Building upon the superior performance of InternVL-2.5 [9] with larger vision encoders established in prior experiments, we adopt it as our baseline for subsequent investigations. Tab. 4 presents the model performance on total and element-level scoring tasks across different folds when trained with our **E**xternal **S**tructural **I**nformation **I**ntegration **P**rompting **S**trategy in Sec. 3.2. This training achieves statistically significant improvements over the baseline in Sec. 4.2 on both holistic alignment (SRCC/PLCC) and granular element matching (ACC), resolving the inherent problem of insufficient information in conventional VQA prompting. Such multifaceted advancement positions our approach as a solid solution for reliable cross-modal evaluation.

However, the substantial performance variance across 5-fold validations suggests that limited prompt diversity and inadequate sample cardinality can significantly induce non-negligible inter-partition discrepancies when conducting stratified data splitting.

4.4. Performance of Ensemble Stragety

To synergize cross-validated models capabilities for enhanced holistic and fine-grained cross-modal alignment accuracy, we implement the ensemble protocol detailed in Sec. 3.3, achieving statistically significant gains through differential weighting of vision-language attention patterns across validation folds. The extensive experimental results demonstrate that our hierarchical integration strategy, which combines 5-fold cross-validation partitioning of multiple LVLMs (including Qwen2.5-VL-7B [4] and InternVL-2.5 variants (26B & 38B) [9]) with subsequent stacking using gradient-boosted tree models, achieves significant performance gains across all evaluation metrics: SRCC (+1.63%), PLCC (+1.96%) in total score evaluation, and ACC (+1.10%) in element-level analysis, culminating in the impressive overall performance improvement. Further enhancements are also observed through the incorporation of external structural information as statistic features in prompt engineering, yielding additional gains across SRCC, PLCC, and ACC metrics. This progression validates our proposed ensemble stragety: 1) Base model diversification, 2) Meta-learning through stacked generalization, as well as 3) Structural-integration refinement.

Collectively, these innovations culminate in a highly effective framework that not only addresses the limitations of existing methods but also sets a new benchmark for cross-modal alignment evaluation tasks. Our approach demonstrates the potential of ensemble learning and structural integration to push the boundaries of model performance in fine-grained vision-language matching.

5. Conclusion and Prospect

This work aims to tackle the critical challenge of fine-grained vision-language alignment evaluation in text-to-image generation. By introducing TokenFocus-VQA, we establish a new evaluation framework that combines token-aware probability optimization with multi-model ensemble strategies. The proposed position-specific loss calculation enables precise supervision for localized semantic matching, while the systematic integration of Bagging, Stacking, as well as Blending techniques further enhances the evaluation robustness. The extensive experimental results on the *NTIRE 2025 Text to Image Generation Model Quality Assessment Challenge* demonstrates state-of-the-art performance on public evaluations. Our proposed framework not only advances the methodological foundation for T2I (Text-to-Image) quality assessment but also provides actionable insights for advancing the semantically-aware evaluation systems in multimodal AI research.

In the future developments, we plan to explore the dynamic vocabulary adaptation and more advanced cross-modal interaction components for broader applicability.

References

[1] Gpt-4v(ision) system card. 2023. 3

[2] Anonymous. Llava-onevision: Easy visual task transfer. *Technical Report*, 2024. 3

[3] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 3

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3, 6, 8

[5] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 35:32897–32912, 2022. 3

[6] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *Computer Vision – ECCV 2006*, pages 404–417, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. 3

[7] Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, Xiaoyi Dong, Haodong Duan, Qi Fan, Zhaoye Fei, Yang Gao, Jiaye Ge, Chenya Gu, Yuzhe Gu, Tao Gui, Aijia Guo, Qipeng Guo, Conghui He, Yingfan Hu, Ting Huang, Tao Jiang, Penglong Jiao, Zhenjiang Jin, Zhikai Lei, Jiaxing Li, Jingwen Li, Linyang Li, Shuaibin Li, Wei Li, Yining Li, Hongwei Liu, Jiangning Liu, Jiawei Hong, Kaiwen Liu, Kuikun Liu, Xiaoran Liu, Chengqi Lv, Haijun Lv, Kai Lv, Li Ma, Runyuan Ma, Zerun Ma, Wenchang Ning, Linke Ouyang, Jiantao Qiu, Yuan Qu, Fukai Shang, Yunfan Shao, Demin Song, Zifan Song, Zhihao Sui, Peng Sun, Yu Sun, Huanze Tang, Bin Wang, Guoteng Wang, Jiaqi Wang, Jiayu Wang, Rui Wang, Yudong Wang, Ziyi Wang, Xingjian Wei, Qizhen Weng, Fan Wu, Yingtong Xiong, Chao Xu, Ruiliang Xu, Hang Yan, Yirong Yan, Xiaogui Yang, Haochen Ye, Huaiyuan Ying, Jia Yu, Jing Yu, Yuhang Zang, Chuyu Zhang, Li Zhang, Pan Zhang, Peng Zhang, Ruijie Zhang, Shuo Zhang, Songyang Zhang, Wenjian Zhang, Wenwei Zhang, Xingcheng Zhang, Xinyue Zhang, Hui Zhao, Qian Zhao, Xiaomeng Zhao, Fengzhe Zhou, Zaida Zhou, Jingming Zhuo, Yicheng Zou, Xipeng Qiu, Yu Qiao, and Dahua Lin. Internlm2 technical report, 2024. 4

[8] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016. 6

[9] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 3, 6, 7, 8

[10] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks, 2024. 3

[11] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Pro-*

ceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 24185–24198, 2024. 3

[12] Federico D'Asaro, Sara De Luca, Lorenzo Bongiovanni, Giuseppe Rizzo, Symeon Papadopoulos, Manos Schinas, and Christos Koutlis. Zero-shot content-based crossmodal recommendation system. *Expert Systems with Applications*, 258:125108, 2024. 3

[13] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*, 2017. 2, 3

[14] Shuhao Han, Haotian Fan, Jiachen Fu, Liang Li, Tao Li, Junhui Cui, Yunqiu Wang, Yang Tai, Jingwei Sun, Chunle Guo, and Chongyi Li. Evalmuse-40k: A reliable and fine-grained benchmark with comprehensive human annotations for text-to-image generation model evaluation, 2024. 1, 6, 7

[15] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning, 2022. 1, 6, 7

[16] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 1

[17] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 6

[18] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering, 2023. 1

[19] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36:78723–78747, 2023. 1

[20] Ziwei Huang, Wanggui He, Quanyu Long, Yandi Wang, Haoyuan Li, Zhelun Yu, Fangxun Shu, Long Chan, Hao Jiang, Leilei Gan, et al. T2i-factualbench: Benchmarking the factuality of text-to-image models with knowledge-intensive concepts. *arXiv preprint arXiv:2412.04300*, 2024. 1

[21] Yao Jiang, Xinyu Yan, Ge-Peng Ji, Keren Fu, Meijun Sun, Huan Xiong, Deng-Ping Fan, and Fahad Shahbaz Khan. Effectiveness assessment of recent large vision-language models. *Visual Intelligence*, 2(1):17, 2024. 3

[22] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*. Minneapolis, Minnesota, 2019. 3

[23] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR, 2021. 3

[24] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, pages 201–216, 2018. 1, 3

[25] Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Xide Xia, Pengchuan Zhang, Graham Neubig, and Deva Ramanan. Evaluating and improving compositional text-to-visual generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5290–5301, 2024. 1

[26] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models, 2024. 3

[27] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 1

[28] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 1, 3

[29] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 1, 3

[30] Ran Li, Xiaomeng Jin, et al. Real: Realism evaluation of text-to-image generation models for effective data augmentation. *arXiv preprint arXiv:2502.10663*, 2025. 1

[31] Wenxiong Liao, Bi Zeng, Jianqi Liu, Pengfei Wei, and Jiongkun Fang. Image-text interaction graph neural network for image-text sentiment analysis. *Applied Intelligence*, 52 (10):11184–11198, 2022. 3

[32] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 3

[33] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024. 3

[34] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024. 3

[35] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 3

[36] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 3

[37] Jisu Nam, Heesu Kim, DongJae Lee, Siyoon Jin, Seungryong Kim, and Seunggyu Chang. Dreammatcher: appearance matching self-attention for semantically-consistent text-to-image personalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8100–8110, 2024. 3

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 2, 3

[39] Arnau Ramisa, Rene Vidal, Yashar Deldjoo, Zhankui He, Julian McAuley, Anton Korikov, Scott Sanner, Mahesh Sathiamoorthy, Atoosa Kasrizadeh, Silvia Milano, et al. Multimodal generative models in recommendation system. *arXiv preprint arXiv:2409.10993*, 2024. 3

[40] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans, 2016. 1

[41] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 3

[42] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023. 3

[43] Taowen Wang, Zheng Fang, Haochen Xue, Chong Zhang, Mingyu Jin, Wujiang Xu, Dong Shu, Shanchieh Yang, Zhenting Wang, and Dongfang Liu. Large vision-language model security: A survey. In *Frontiers in Cyber Security*, pages 3–22, Singapore, 2024. Springer Nature Singapore. 3

[44] Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models. *arXiv preprint arXiv:2210.14896*, 2022. 5

[45] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. Multi-modality cross attention network for image and sentence matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10941–10950, 2020. 3

[46] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, et al. Qwen2.5 technical report, 2025. 4

[47] Mang Ye, Xuankun Rong, Wenke Huang, Bo Du, Nenghai Yu, and Dacheng Tao. A survey of safety on large vision-language models: Attacks, defenses and evaluations. 2025. 3

[48] Ying Zhang and Huchuan Lu. Deep cross-modal projection learning for image-text matching. In *Proceedings of the European conference on computer vision (ECCV)*, pages 686–701, 2018. 3

[49] Tong Zhu, Leida Li, Jufeng Yang, Sicheng Zhao, Hantao Liu, and Jiansheng Qian. Multimodal sentiment analysis with image-text interaction network. *IEEE transactions on multimedia*, 25:3375–3385, 2022. 3

[50] Xiangru Zhu, Penglei Sun, Chengyu Wang, Jingping Liu, Zhixu Li, Yanghua Xiao, and Jun Huang. A contrastive compositional benchmark for text-to-image synthesis: A study with unified text-to-image fidelity metrics. *arXiv preprint arXiv:2312.02338*, 2023. 1