Vision-to-Text: Benchmarking Multimodal LLMs on Extremely Low-Resource Languages

Shuoshuo Hou[1,2,3,4], Mieradilijiang Maimaiti[1,2,3,4,*], Zhexin Li[1,2,3,4], Jiaxin Wang[1,2,3,4], Ahmad Hassan[1,2,3,4], Kaishaer Jiapaer[1], Nilufar Abdurakhmonova[5], Roza Urinbayeva[5], Madina Mansurova[6], Shormakova Assem[6], Gulnar Murat[7], Le Wu[8], and Wushouer Silamu[1,2,3,4]

[1]School of Computer Science and Technology, Xinjiang University, Urumqi, China
[2]Xinjiang Laboratory of Multi-Language Information Technology, Xinjiang University, Urumqi, China
[3]Xinjiang Multilingual Information Technology Research Center, Urumqi, China
[4]Joint International Research Laboratory of Silk Road Multilingual Cognitive Computing, Urumqi, China
[5]Department of Computational Linguistics, National University of Uzbekistan, Tashkent, Uzbekistan
[6]Faculty of Information Technology and Artificial Intelligence, Al-Farabi Kazakh National University, Almaty, Kazakhstan
[7]Kapshagay Bidai Onimderi, Kapshagay, Kazakhstan
[8]Integrated Laboratory for Space, Air, and Ground Systems, Changji, China
Email: {yumoya, lizhexin, 107552503749, k20241401814_}@stu.xju.edu.cn,
{miradeljan51, wushour}@xju.edu.cn, {n.abduraxmonova, orinboyeva_r}@nuu.uz,
ahmadhassan2067@gmail.com, madina.mansurova@kaznu.edu.kz,
Shormakova.asem1@kaznu.edu.kz, DrGulnarMurat@hotmail.com, alaia@richnetworks.hk

***Abstract*—Multimodal large language models (MLLMs) have emerged as a dominant paradigm for visually grounded language tasks. However, multimodal machine translation (MMT) remains constrained by the scarcity of multimodal data for low-resource languages (LRLs). Previously proposed approaches have also explored some strategies for machine translation with MLLMs under low-resource scenarios; however, they still face the critical bottleneck of a lack of high-quality multimodal datasets for extremely low-resource languages. To this end, we propose a novel and straightforward method for constructing a high-quality benchmark for MMT in extremely low-resource languages by harnessing LLMs and leveraging a human evaluation strategy. The presented approach for building the multimodal low-resource benchmark consists of three phases. First, we select an optimal visual captioner via metric-driven evaluation to ensure reliable grounding; second, we use a hybrid ensemble of diverse translation models to reduce bias and improve the fluency–faithfulness balance; finally, we apply a triangular quality estimation scheme—reference-free quality, semantic consistency, and visual alignment—to filter candidates. We conduct extensive experiments on multiple corpora using various evaluation metrics. The results show that strong baselines (e.g., Qwen3-VL-8B and LLaVA-v1.6-7B), when fine-tuned on our proposed benchmark, achieve substantial improvements not only on our benchmark but also on well-known multimodal datasets such as Visual Genome and Multi30K, with average gains of +7.67, +8.50, and +9.67 on COMET-Kiwi, CLIP, and BERTScore, respectively. The corresponding code and constructed high-quality multimodal low-resource languages benchmark are publicly available[1].**

***Index Terms*—Multimodal Corpus Construction, Low-Resource Languages, Synthetic Data Generation, Quality Estimation, Visual Grounding**

*Corresponding author.

[1]https://huggingface.co/datasets/yumoya/SilkRoad-VL

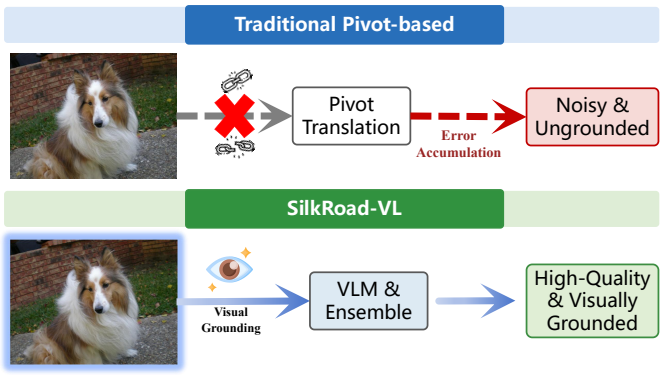

Fig. 1. Paradigm Comparison. Unlike traditional methods (top), prone to ungrounded hallucinations, SilkRoad-VL (bottom) leverages visual grounding to ensure semantic and script fidelity.

I. Introduction

Multimodal large language models (MLLMs) have emerged as powerful foundation models for vision-language applications. In machine translation, they leverage visual context to ground meaning and resolve ambiguities such as polysemy and gender, which is especially beneficial for low-resource languages with scarce parallel data [1], [2]. However, despite recent scaling in MT, many underrepresented languages, especially in Central and South Asia, still suffer from a persistent performance gap [3]. Progress is limited by the scarcity of high-quality, image-aligned multilingual corpora: existing datasets are concentrated in a few high-resource languages and provide limited coverage for diverse scripts and regions in Central and South Asia.

To address this scarcity, large-scale image-caption-translation corpora are needed, yet high-quality resources remain largely human-curated and limited to a few high-resource pairs (e.g., Multi30K) [4]. Consequently, scaling with mined or synthetic data (e.g., CCMatrix [5]) has become necessary. However, such data often produce "translationese" [6] and, under data scarcity, can amplify language confusion and hallucinations [7]. These issues necessitate a scalable construction paradigm in which visual grounding serves as an explicit constraint for filtering artifacts and enhancing the quality of the low-resource MMT data.

Recent scalable pipelines increasingly replace costly human annotation with synthetic generation. A common strategy is to generate English captions using strong VLMs such as Qwen3-VL [8] or LLaVA-style models [9], followed by translation with multilingual NMT systems, including NLLB [3], MADLAD-400 [10], or LLM translators [11]. Although many pipelines apply vision–language alignment scorers (e.g., CLIP [12] or SigLIP [13]) to filter noise, they often treat translation and grounding as separate steps. Despite improved coverage, these pipelines remain vulnerable to model-specific biases and imperfect grounding, leaving quality control and robustness in low-resource settings unresolved.

To address this gap, as illustrated in Figure 1, we propose SilkRoad-VL, a fully automated pipeline for the construction of a low-resource multimodal corpus. The pipeline follows a three-stage *metric-driven* and *ensemble-based* workflow. First, it improves source fidelity by selecting the best caption generator from several state-of-the-art VLMs. Second, it introduces a *Hybrid Ensemble Generation* module that combines four complementary architectures (e.g., NLLB, SeamlessM4T, and Qwen) to balance lexical precision and fluency. Third, it applies a *triangular quality estimation (Tri-QE)* protocol to filter candidates from three complementary perspectives: quality, consistency, and visual grounding.

Our main contributions are as follows:

- We establish a pipeline that integrates metric-driven VLM selection and a four-model ensemble, enhancing the robustness of translation and accuracy of the script.
- We propose the Tri-QE protocol, which jointly evaluates reference-free quality, semantic consistency, and visual grounding to filter hallucinated and low-fidelity samples.
- We release SilkRoad-VL and validate its effectiveness via instruction tuning, demonstrating significant gains over baselines for six languages.

II. Related Work

A. Multimodal Resources

High-quality, image-aligned parallel corpora are fundamental to multimodal research. Multi30K [4] remains the most widely used benchmark, and subsequent efforts have extended coverage to languages such as Japanese [14] and Hindi [15]. However, existing resources remain concentrated in high-resource European and East Asian languages.

Low-resource languages in Central and South Asia (e.g., Uyghur, Kazakh, and Urdu) remain severely underrepresented. As a result, research on these languages often relies on noisy text-only corpora such as CCMatrix [5], which suffer from misalignment and lack visual grounding. Our work addresses this gap by constructing a large-scale, visually grounded parallel corpus for six under-resourced languages in this region.

B. Synthetic Data Paradigms

Given the prohibitive cost of manual annotation, synthetic corpus construction has become a standard paradigm. Traditional approaches utilize pivot-based architectures, such as NLLB [3] or Flores-101 [16]. While providing broad coverage, these models often yield "translationese"—outputs that are grammatically correct but lack lexical diversity and natural flow [17].

LLMs have also shown strong few-shot multilingual ability. In the visual domain, works such as LLaVA [18] and ShareGPT4V [19] show that high-quality synthetic captions can improve VLM alignment, although these efforts mainly focus on English. Meanwhile, multilingual LLMs such as ALMA [20] and Qwen [21] can outperform supervised baselines, but in low-resource settings they remain prone to hallucinations and script inconsistencies. Unlike prior work relying on a single model family, we adopt a hybrid ensemble strategy that combines the lexical precision of supervised systems with the fluency of LLMs.

C. Quality Estimation and Filtering

Ensuring the quality of synthetic data is paramount. Traditional filtering relies on heuristic rules or round-trip consistency metrics [22], which often correlate poorly with human judgment. Recently, reference-free Quality Estimation (QE) metrics like COMET-Kiwi [23] have become the gold standard for assessing textual adequacy.

However, text-only metrics are insufficient in multimodal settings because they ignore visual grounding. Although recent studies use CLIP [24] to measure image-text alignment [12], such metrics are often applied independently. Our work instead introduces a unified Triangular Quality Estimation (Tri-QE) protocol that jointly evaluates reference-free quality (COMET), semantic consistency (BERTScore), and visual grounding (CLIP).

III. Method

We propose a fully automated pipeline to construct SilkRoad-VL, prioritizing semantic fidelity, linguistic diversity, and visual grounding. As illustrated in Figure 2 and formalized in Algorithm 1, the workflow consists of three sequential stages: metric-driven visual captioning, hybrid ensemble generation, and a triangular quality estimation (Tri-QE) filtering protocol.

A. Visual Captioning via VLM

High-quality English captions serve as the semantic anchor of our multimodal parallel corpus. To ensure fidelity and diversity, we adopt a metric-driven VLM selection stage followed by SigLIP-based filtering.

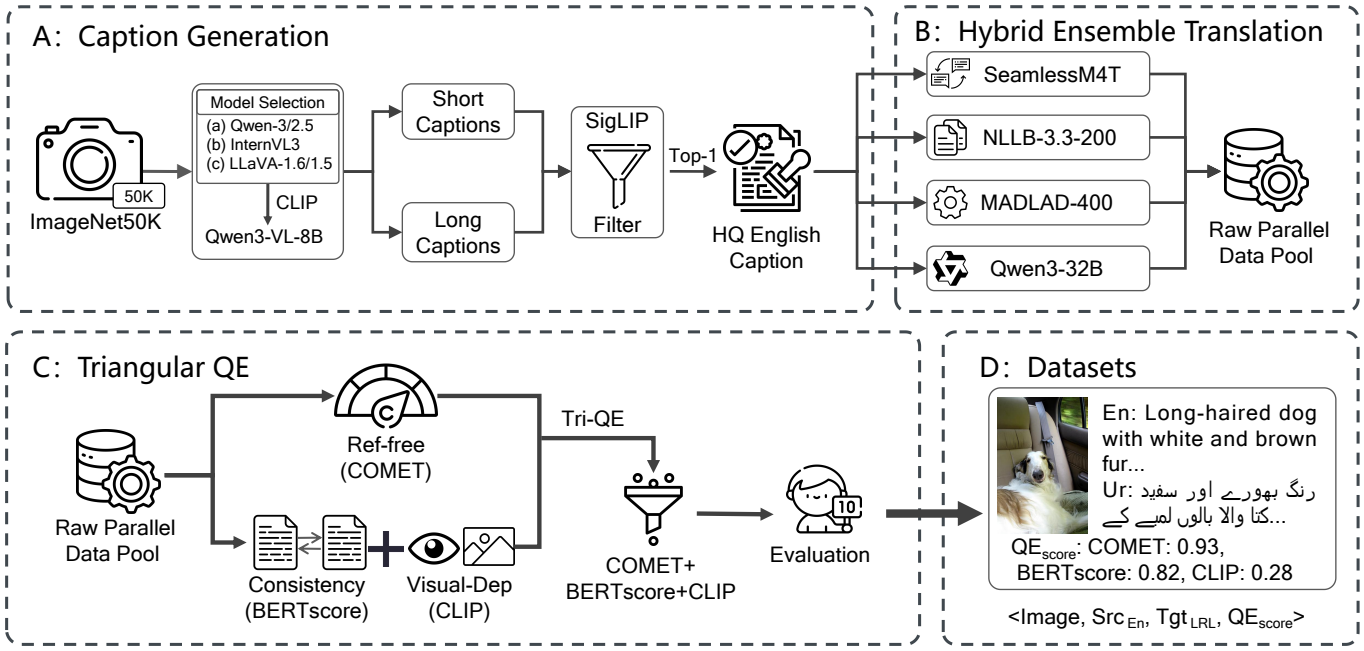

Fig. 2. Overview of the SilkRoad-VL construction pipeline. The process begins with metric-driven VLM selection for source captioning, followed by a hybrid ensemble generation strategy, and concludes with a rigorous Tri-QE filtering protocol to ensure high-quality, visually grounded data.

1) Optimal VLM Selection: We first identified the optimal generator from a candidate set of five state-of-the-art Vision-Language Models (VLMs): $\mathcal{M} = \{M_{\text{qwen3}}, M_{\text{qwen2.5}}, M_{\text{intern3}}, M_{\text{llava1.6}}, M_{\text{llava1.5}}\}$. Rather than relying on subjective manual evaluation, we formulated the selection as a constrained optimization problem. For a validation set $\mathcal{D}_{\text{val}}$, we generate captions $C_{m,v}$ using model m given image v. We define an indicator function $\mathbb{I}_{\text{len}}(C)$ that strictly enforces general length compliance. The optimal model M^* is selected by maximizing the average visual alignment score subject to these constraints:

$$M^* = \underset{m \in \mathcal{M}}{\text{argmax}} \frac{1}{|\mathcal{D}_{\text{val}}|} \sum_{v \in \mathcal{D}_{\text{val}}} \mathcal{S}_{\text{clip}}(v, C_{m,v}) \cdot \mathbb{I}_{\text{len}}(C_{m,v}), \quad (1)$$

where $\mathcal{S}_{\text{clip}}(\cdot)$ denotes CLIPScore. As shown in Section IV-C, Qwen3-VL-8B achieves the best trade-off between visual fidelity and length controllability, and is therefore used as the backbone generator.

2) Dual-Granularity Generation: To increase linguistic diversity, we use two prompts to generate *short captions* capturing global scene semantics and *long captions* describing fine-grained attributes.

3) SigLIP Filtering: We then filter the generated captions with SigLIP to suppress hallucinations and enforce structural consistency. Let $\mathcal{R}_{\text{short}} = [8, 20]$ and $\mathcal{R}_{\text{long}} = [28, 45]$. For a caption c of type $k \in \{\text{short}, \text{long}\}$ and image I, a sample is kept only if its SigLIP score exceeds $\tau_{\text{sig}} = 0.9$ and its length falls within the target range:

Algorithm 1 SilkRoad-VL Construction Pipeline with Tri-QE Filtering

Input: Image set $\mathcal{I}$, VLM $\mathcal{M}_{vlm}$, Ensemble Models $\mathcal{M}_{gen}$, Thresholds $\tau_{\{c,b,v,sig\}}$
Output: Multimodal Parallel Corpus $\mathcal{D} \leftarrow \emptyset$

for all image $I \in \mathcal{I}$ **do**
  Step 1: Generate & Filter Captions:
  $C_{\text{raw}} \leftarrow \mathcal{M}_{vlm}(I)$
  $C_{\text{valid}} \leftarrow \{c \in C_{\text{raw}} \mid \text{SigLIP}(I, c) > \tau_{\text{sig}}\}$
  for all $C_{en} \in C_{\text{valid}}$ **and** $l \in L_{tgt}$ **do**
    Step 2: Ensemble Generation:
    $\mathcal{T} \leftarrow \{m(C_{en}) \mid m \in \mathcal{M}_{gen}^{\text{Pivot}}\} \cup \{\mathcal{L}(C_{en})\}$
    Step 3: Tri-QE:
    $T_{bt} \leftarrow \text{BackTranslate}(T)$
    Calculate scores: $S_{\text{comet}} \leftarrow \text{Kiwi}(T)$, $S_{\text{bert}} \leftarrow \text{BS}(C_{en}, T_{bt})$, $S_{\text{clip}} \leftarrow \text{CLIP}(I, T_{bt})$
    $T^* \leftarrow \arg\max_{T \in \mathcal{T}} \{S_{\text{comet}}(T) \mid S_{\text{comet}} > \tau_c \wedge S_{\text{bert}} > \tau_b \wedge S_{\text{clip}} > \tau_v\}$
    if $T^* \neq$ None **then**
      $\mathcal{D} \leftarrow \mathcal{D} \cup \{(I, C_{en}, T^*, l)\}$
    end if
  end for
end for
return $\mathcal{D}$

$$\text{Keep}(c) \iff \mathcal{S}_{\text{siglip}}(I, c) > \tau_{\text{sig}} \quad \wedge \quad |c| \in \mathcal{R}_k. \quad (2)$$

This strict filtering ensures that our source dataset consists exclusively of visually faithful and structurally precise captions.

B. Hybrid Ensemble Generation

To mitigate translationese and the limited coverage in single-model generation, we adopt a *Hybrid Ensemble Generation* strategy. Given a source caption x, we construct a diverse candidate set $\mathcal{T}(x)$ using four complementary models for lexical fidelity, multimodal alignment, and fluency:

$$\mathcal{T}(x) = \{\mathcal{M}_{\text{nllb}}(x), \mathcal{M}_{\text{madlad}}(x), \mathcal{M}_{\text{seamless}}(x), \mathcal{M}_{\text{llm}}(x|\mathcal{P})\}, \tag{3}$$

where $\mathcal{M}_k$ denotes model k and $\mathcal{P}$ is the script-constrained prompt for the LLM.

Precision-Oriented Architectures: We employ NLLB-200 (3.3B) and MADLAD-400 (7B) as text translation backbones, which provide strong grammaticality and lexical fidelity, especially for morphologically rich languages.

Multimodal-Native Alignment: We incorporate SeamlessM4T-v2 (Large), whose multimodal pretraining helps reduce semantic ambiguity and improves alignment with visually grounded concepts.

Fluency-Oriented LLM: We leverage Qwen3-32B-Instruct to improve naturalness. To reduce script hallucinations in low-resource settings, we apply a *Script-Constrained Prompting* mechanism $\mathcal{P}$ that enforces the official target script and forbids transliteration. The LLM output is defined as $T_{\text{LLM}} = \mathcal{L}(x, \mathcal{P})$.

C. Tri-QE Filtering Protocol

To rigorously filter the synthetic candidate pool, we introduce the *Triangular Quality Estimation (Tri-QE)* protocol, which evaluates each candidate translation T from three complementary perspectives. We measure *reference-free quality* (S_{comet}) using COMET-Kiwi as the primary proxy for fluency and adequacy, *semantic consistency* (S_{bert}) using BERTScore between the English source S and its back-translation S' (generated by NLLB-200), and *visual grounding* (S_{clip}) using CLIPScore between the original image and the back-translated text S'.

Constrained Maximization: For each source instance in language l, we construct a language-specific candidate set $\mathcal{C}_l$ and select the final translation via constrained maximization. We first define a validity indicator $\mathbb{I}_{\text{valid}}(T)$:

$$\mathbb{I}_{\text{valid}}(T) = \mathbb{I}(S_{\text{comet}} > \tau_c) \cdot \mathbb{I}(S_{\text{bert}} > \tau_b) \cdot \mathbb{I}(S_{\text{clip}} > \tau_v), \tag{4}$$

where τ_c, τ_b, τ_v are empirical thresholds. Based on a sensitivity analysis balancing data quality and corpus scale, we set $\tau_c = 78.0$, $\tau_b = 0.90$, and $\tau_v = 0.27$.

The final selection is given by:

$$T^* = \underset{T \in \mathcal{C}_l}{\operatorname{argmax}} \left(S_{\text{comet}}(T) \cdot \mathbb{I}_{\text{valid}}(T) \right). \tag{5}$$

If no candidate satisfies the constraints, the sample is discarded. This filtering removes hallucinated and low-fidelity translations, yielding a cleaner corpus for downstream fine-tuning.

IV. Experiments

A. Setup

Datasets We constructed SilkRoad-VL, a large-scale multimodal corpus for six low-resource languages, namely Uyghur (ug), Kazakh (kk), Kyrgyz (ky), Tajik (tg), Uzbek (uz), and Urdu (ur). As summarized in Table I, our dataset comprises 82.7k training, 0.4k validation, and 1.8k test samples, significantly exceeding the scale of existing benchmarks. To assess generalization, we strictly follow standard protocols on Visual Genome[2] (VG) and Multi30k[3], utilizing their established partitions for evaluation.

TABLE I
STATISTICS OF SILKROAD-VL AND BENCHMARKS

Dataset	Train	Dev	Test
Visual Genome	29k	1k	1.6k
Multi30k	29k	–	1k
SilkRoad-VL	82.7k	0.4k	1.8k

To ensure a holistic assessment, we employ three orthogonal reference-free metrics: COMET-Kiwi[4] [23] for textual adequacy, CLIPScore[5] [12] for visual grounding assessment, and BERTScore[6] [25] for semantic consistency via back-translation.

Corpus Statistics Table II summarizes the fundamental statistics of the dataset. After applying the rigorous Tri-QE filtering protocol, the final corpus contains 84,902 high-quality image-text pairs across six languages.

Dual-Granularity Distribution: A key structural innovation of SilkRoad-VL is the balanced inclusion of both concise global summaries and detailed attribute descriptions. As visualized in Figure 3, the sentence length distribution exhibits a distinct bimodal pattern. The first peak ($\sim$16 tokens) corresponds to Short Captions providing high-level context, while the second peak ($\sim$35 tokens) captures fine-grained visual details. This balanced distribution ensures the dataset supports diverse multimodal tasks, addressing the limitations of existing corpora that favor simplistic descriptions.

Baselines To comprehensively evaluate the quality and effectiveness of SilkRoad-VL, we compare our approach against a diverse set of state-of-the-art Multimodal Large Language Models (MLLMs). We select representative open-source models across different architectures:

- Qwen Series: We include Qwen3-VL-8B [8] and Qwen2.5-VL-7B [26], which currently represent the strong recent baselines in multilingual visual understanding and generation.
- LLaVA Series: We employ LLaVA-v1.6-7B [27] and the classic LLaVA-1.5-7B [9], serving as robust baselines for instruction-tuned visual capabilities.

[2]https://visualgenome.org/
[3]https://github.com/multi30k/dataset
[4]https://github.com/Unbabel/COMET
[5]https://github.com/jmhessel/clipscore
[6]https://github.com/Tiiiger/bert_score

TABLE II
STATISTICS OF SILKROAD-VL DATASET

Language	Total	Short		Long	
		Count	Length	Count	Length
ug	8.7k	3.7k	15.7	5.0k	34.7
kk	26.1k	13.0k	14.8	13.2k	31.6
ky	23.2k	12.8k	15.4	10.3k	33.0
tg	2.4k	1.8k	17.0	0.6k	40.0
uz	14.9k	8.2k	15.2	6.7k	33.0
ur	9.6k	5.1k	23.3	4.5k	56.6
Total / Avg.	**84.9k**	**44.6k**	**16.9**	**40.3k**	**38.2**

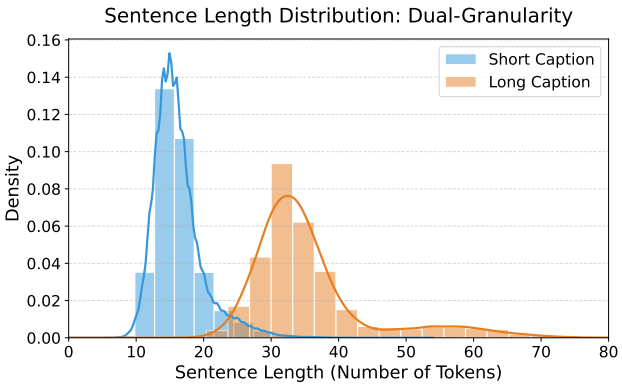

Fig. 3. Sentence length distribution of the SilkRoad-VL dataset. The distinct bimodal pattern validates the effectiveness of our Dual-Granularity strategy, with the first peak (∼16 tokens) capturing concise global context and the second peak (∼35 tokens) providing detailed attribute descriptions.

- InternVL Series: We also compare against InternVL3-8B [28], a strong competitor known for its powerful visual encoder.

Implementation Details We present the comprehensive hyperparameter settings for our implementation in Table III. The pipeline was built using PyTorch [29] and HuggingFace Transformers [30], executed on a cluster of 8× NVIDIA H100 (80GB) GPUs. To balance computational efficiency with model expressivity, we adopted the Weight-Decomposed Low-Rank Adaptation (DoRA) [31] strategy rather than full-parameter fine-tuning. DoRA enhances the standard LoRA approach by decomposing pre-trained weights into magnitude and direction components, ensuring more stable updates during the adaptation of the LLM backbone's linear projection layers (`q_proj`, `k_proj`, `v_proj`, `o_proj`). The training process utilized the AdamW optimizer with a cosine learning rate schedule and gradient accumulation to optimize performance within the specified compute budget.

B. Main Results

As detailed in Table IV, fine-tuning on SilkRoad-VL yields consistent and substantial gains. In this and subsequent tables, "Ours" refers to the model initialized with the Qwen3-VL backbone. Our approach significantly outperforms all strong baselines across six languages, achieving a peak score of 62.47 in Kazakh. Notably, in the low-resource language of Tajik, our method surpasses the best zero-shot baseline (46.61) by a remarkable margin of 11.6 points. These universal improvements confirm that our dataset provides a high-quality supervision signal, effectively bridging the visual-linguistic alignment gap.

Comparison on Multi30k. As shown in Table V, our models demonstrate robust generalization capabilities, consistently outperforming baselines across all languages. Our approach achieves remarkable adaptability, reaching peak scores of 50.36 in Kazakh and 48.98 in Uyghur. Furthermore, in the lowest-resource setting of Tajik, our method establishes a new state-of-the-art at 40.20, exceeding the strongest baseline (26.57) by a significant margin of 13.6 points, confirming the efficacy of our alignment strategy.

Comparison on Visual Genome. As summarized in Table VI, our strategy demonstrates exceptional zero-shot generalization on the Visual Genome benchmark. We consistently outperform strong baselines across all six languages. Notably, in Urdu, our method achieves a striking improvement of +10.3 points over

TABLE III
IMPLEMENTATION HYPERPARAMETERS

Stage	Parameter	Value
Visual Captioning	Temperature	0.8
	Repetition Penalty	1.1
	Short Prompt Limit	$\leq$ 20 words
	Long Prompt Limit	$\leq$ 45 words
Ensemble Generation	Beam Size (Pivot)	5
	Top-p Sampling (LLM)	0.9
	LLM Quantization	4-bit
Tri-QE Filtering	COMET Thresh. (τ_c)	78.0
	BERTScore Thresh. (τ_b)	0.90
	CLIPScore Thresh. (τ_v)	0.27
Training (DoRA)	Rank (r)	64
	Alpha (α)	128
	Learning Rate	$2e^{-5}$
	Batch Size	128
	Epochs	3

TABLE IV
COMPARISON WITH STRONG BASELINES ON THE SILKROAD TESTSET USING THE AVERAGE OF COMET-KIWI, BERTSCORE, AND CLIP. UNDERLINED VALUES INDICATE THE BEST BASELINE PERFORMANCE.

Model	kk	ky	tg	ur	ug	uz
Qwen3-VL	<u>52.69</u>	<u>49.93</u>	<u>46.61</u>	<u>55.11</u>	47.14	<u>54.53</u>
Qwen2.5-VL	51.12	48.46	45.44	50.65	<u>52.57</u>	49.32
LLaVA-v1.6	41.12	42.55	41.75	41.40	41.28	44.91
LLaVA-1.5	49.93	48.18	44.82	43.49	42.38	47.06
InternVL3	46.21	44.59	44.94	46.60	44.62	45.85
Ours	**62.47**	**61.24**	**58.25**	**60.12**	**60.87**	**62.33**

TABLE V
COMPARISON WITH STRONG BASELINES ON THE MULTI30K TESTSET USING THE AVERAGE OF COMET-KIWI AND CLIP. UNDERLINED VALUES INDICATE THE BEST BASELINE PERFORMANCE.

Model	kk	ky	tg	ur	ug	uz
Qwen3-VL	41.12	34.17	26.23	30.89	42.79	41.64
Qwen2.5-VL	34.07	30.15	24.78	36.43	34.77	33.55
LLaVA-v1.6	22.92	23.88	24.48	23.38	23.43	28.30
LLaVA-1.5	32.87	29.25	26.57	25.67	25.98	29.13
InternVL3	29.16	27.23	24.36	30.81	28.49	28.38
Ours	**50.36**	**49.35**	**40.20**	**48.41**	**48.98**	**49.06**

TABLE VI
COMPARISON WITH STRONG BASELINES ON THE VISUAL GENOME TESTSET USING THE AVERAGE OF COMET-KIWI AND CLIP. UNDERLINED VALUES INDICATE THE BEST BASELINE PERFORMANCE.

Model	kk	ky	tg	ur	ug	uz
Qwen3-VL	41.57	38.87	33.62	34.04	41.85	41.32
Qwen2.5-VL	37.07	32.12	29.82	37.97	37.69	37.12
LLaVA-v1.6	26.07	26.72	28.64	28.89	26.48	30.27
LLaVA-1.5	32.29	31.17	28.31	27.39	26.55	30.68
InternVL3	29.94	31.58	30.75	33.53	29.80	32.74
Ours	**49.43**	**47.56**	**37.05**	**48.31**	**48.19**	**46.34**

the best baseline (37.97 → 48.31). Even in the challenging Tajik subset, where prior models struggle, our approach establishes a new state-of-the-art at 37.05, significantly surpassing the strongest competitor Qwen3-VL (33.62). This confirms that our data provides robust visual-linguistic alignment that transfers effectively to unseen domains.

Multi-dimensional Quality Analysis. To provide a holistic view of the dataset's quality, Figure 4 visualizes the performance across three key dimensions, revealing a remarkably consistent "high-quality profile" across all six languages. Specifically, BERTScore (orange) consistently exceeds 93.5, indicating strict preservation of core meaning, while stable COMET scores around 82.0 (blue) demonstrate the ensemble strategy's robustness in handling linguistic nuances. Furthermore, CLIP scores (green, right axis) hold steady at 0.30, confirming that our Tri-QE filtering effectively maintains visual alignment without compromising text fluency.

Human Evaluation. We further verify the quality of the constructed corpora (i.e., SilkRoad-VL) by inviting 6 native speakers to evaluate 250 sampled items for each of the 4 representative languages (Uyghur, Uzbek, Kazakh, and Urdu). As shown in Table VII, the sampled short captions assessed on a 1–5 Likert scale achieved generally favorable ratings, particularly an impressive average of 4.71 for Visual Relevance. These results empirically validate that our Tri-QE pipeline effectively ensures linguistic correctness and precise visual grounding. We intend to extend this rigorous assessment to Kyrgyz and Tajik in future work.

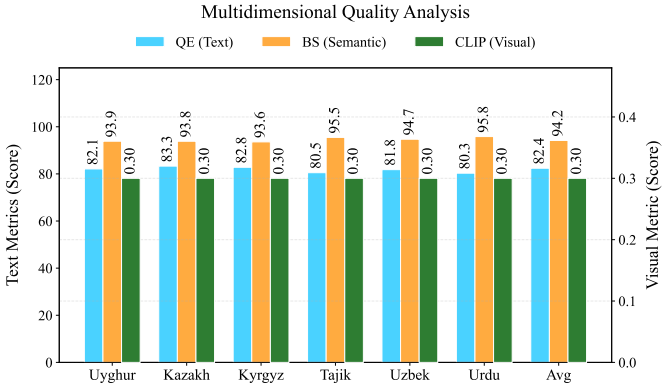

Fig. 4. Visual-semantic analysis of the filtered SilkRoad-VL dataset. The left axis represents linguistic quality (QE) and semantic consistency (BS), while the right axis represents visual grounding (CLIP). The data confirms high consistency across all languages.

TABLE VII
HUMAN EVALUATION OF SILKROAD-VL

Language	Fluency	Adequacy	Visual Relevance
ug	4.61	4.70	4.73
uz	4.10	4.03	4.74
kk	4.33	4.33	4.67
ur	4.80	4.19	4.69
Average	**4.46**	**4.31**	**4.71**

C. Ablation Studies

VLM Backbone Selection. As illustrated in Figure 6, we compare five state-of-the-art VLMs. The dual-axis evaluation reveals a clear divergence between generation length and visual grounding. Specifically, LLaVA-v1.6 exhibits signs of *generative degeneration*, where a sharp increase in average length (86.0 tokens) fails to yield proportional gains in CLIP score, suggesting the presence of hallucination loops. In contrast, Qwen3-VL-8B [8] demonstrates superior *semantic density*, achieving the best performance among the evaluated VLMs (36.48) with a concise token count (34.4). This favorable balance between informativeness and groundedness justifies Qwen3-VL as our optimal backbone generator.

Effectiveness of Hybrid Ensemble Strategy. Figure 5 compares our hybrid strategy with four individual baselines. While single-model generators exhibit clear trade-offs—such as Qwen's decline in semantic consistency (BERTScore ∼93.8)—our ensemble effectively mitigates these biases. By combining the complementary strengths of diverse architectures, the Hybrid Ensemble consistently surpasses all single-source counterparts across every metric. It attains the best performance with 94.48 BERTScore, 82.95 COMET, and 29.71 CLIP, confirming that our ensemble-then-filter paradigm provides a stronger and more reliable supervision signal than any individual teacher.

Effectiveness of Tri-QE Filtering. To assess the reliability

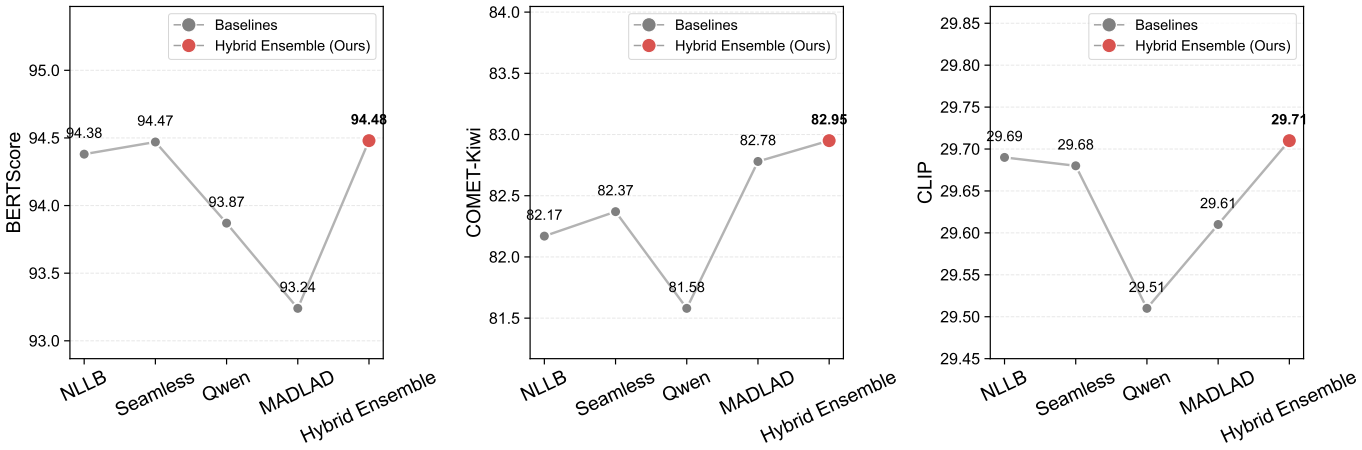

Fig. 5. Performance comparison: Individual Models vs. Hybrid Ensemble. The results demonstrate that our Hybrid Ensemble strategy consistently outperforms all single-model baselines, effectively balancing semantic fidelity (i.e., BERTScore), translation quality (i.e., COMET-Kiwi), and visual-alignment (i.e., CLIP).

TABLE VIII
QUALITY STATISTICS OF RAW POOL VS. FINAL DATASET. IMPROVEMENTS ACROSS QE, CLIP, AND BS VALIDATE THE EFFICACY OF OUR TRI-QE FILTERING

Lang	Raw Pool			Final Dataset		
	QE	CLIP	BS	QE	CLIP	BS
ug	60.55	19.36	88.17	82.13	29.66	93.87
kk	71.83	23.29	91.04	83.25	29.57	93.85
ky	67.49	23.47	91.12	82.80	29.56	93.58
tg	51.70	21.56	90.52	80.49	29.93	95.49
uz	65.08	24.29	91.06	81.81	29.84	94.71
ur	66.40	25.93	94.18	80.31	29.67	95.84
Avg.	63.98	22.98	91.14	**82.35**	**29.71**	**94.20**

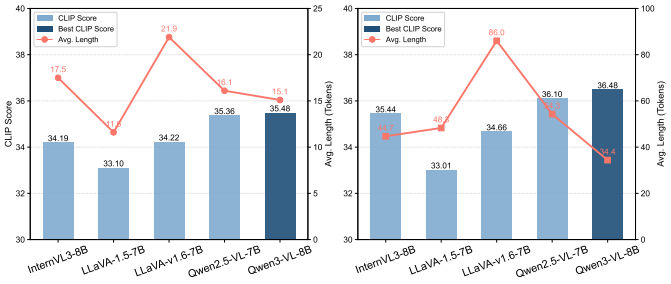

Fig. 6. Ablation on VLM Backbones. The dual-axis plot juxtaposes semantic fidelity (CLIP, bars) against generative verbosity (Length, line). Qwen3-VL achieves the optimal trade-off, securing the highest grounding scores without the pathological length inflation observed in LLaVA-v1.6.

of SilkRoad-VL, we quantitatively compare the final corpus with the raw synthetic pool. As shown in Table VIII, the raw ensemble contains substantial noise, averaging only 63.98 in COMET-Kiwi. In contrast, our Tri-QE protocol filters out low-quality generations, raising the score to 82.35 (+18.37 points). Furthermore, the gains in CLIPScore (22.98 $\rightarrow$29.71) and BERTScore (91.14 $\rightarrow$ 94.20) indicate that the final dataset

Fig. 7. Qualitative comparison. Left: Our fine-tuning corrects Qwen3-VL's hallucinations in Kazakh. Right: It resolves LLaVA-v1.6's repetition loops in Urdu. Red text indicates zero-shot errors.

better preserves both visual grounding and semantic fidelity, reducing hallucinations in synthetic data.

D. Case Study

Figure 7 provides a qualitative assessment of error correction capabilities across diverse architectures. As shown on the left (SilkRoad), the strong baseline Qwen3-VL suffers from severe *object hallucination* in Kazakh, fabricating non-existent attributes by misidentifying a "gray horse" as an object made of "red paper." Similarly, in the Urdu example (Multi30k), LLaVA-v1.6 exhibits characteristic *generation collapse*, degenerating into infinite repetition loops. In stark contrast, our fine-tuned models successfully rectify these pathologies. By aligning visual features with high-quality native text, our approach effectively suppresses hallucinations and restores linguistic fluency, ensuring precise grounding even in complex low-resource scenarios.

V. CONCLUSION AND FUTURE WORK

In this work, we bridge the digital divide for underrepresented Central and South Asian languages by introducing

SilkRoad-VL, a high-fidelity multimodal corpus constructed via a novel, fully automated framework. Our pipeline synergizes metric-driven VLM captioning with hybrid ensemble generation, employing the *Tri-QE* protocol to rigorously mitigate semantic hallucinations and visual-linguistic misalignment. Extensive evaluations confirm that this meticulously filtered data provides a new benchmark for extremely low-resource multimodal translation, effectively unlocking cross-lingual generation and visual grounding capabilities for languages that lack natural multimodal resources. Future work will focus on scaling our pipeline to cover a broader spectrum of low-resource languages and evolving our system into a unified omni-modal framework by incorporating additional modalities such as audio.

Acknowledgment

We sincerely thank all collaborators and co-authors who worked alongside us throughout this project and provided valuable support at different stages of this work. We also thank the anonymous reviewers for their constructive comments and insightful suggestions, which helped improve the quality and clarity of this paper. This work was supported by the National Natural Science Foundation of China (Grant Nos. 62406316, 201704041014, and 201404041254) and the Xinjiang "Tianchi Talent" Recruitment and Introduction Program (awarded to Mieradilijiang Maimaiti).

References

[1] U. Sulubacak, O. Caglayan, S.-A. Grönroos, A. Rouhe, D. Elliott, L. Specia, and J. Tiedemann, "Multimodal machine translation: A survey and benchmark," *Journal of Artificial Intelligence Research*, vol. 69, pp. 347–414, 2020.

[2] O. Caglayan and L. Madhyastha, Pranava andSpecia, "Probing the roles of visual information in multimodal neural machine translation," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 4140–4150.

[3] M. R. Costa-jussà, J. Cross, O. Celebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard *et al.*, "No language left behind: Scaling human-centered machine translation," *arXiv preprint arXiv:2207.04672*, 2022.

[4] D. Elliott, S. Frank, K. Sima'an, and L. Specia, "Multi30k: Multilingual english-german image descriptions," in *Proceedings of the 5th Workshop on Vision and Language*, 2016, pp. 70–74.

[5] H. Schwenk, G. Wenzek, S. Edunov, E. Grave, and A. Joulin, "Ccmatrix: Mining billions of high-quality parallel sentences on the web," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 2021, pp. 6490–6500.

[6] P. Riley, I. Caswell, M. Freitag, and D. Grangier, "Translationese as a language in "multilingual" NMT," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 7737–7746.

[7] N. M. Guerreiro, D. M. Alves, J. Waldendorf, B. Haddow, A. Birch, P. Colombo, and A. F. T. Martins, "Hallucinations in large multilingual translation models," *Transactions of the Association for Computational Linguistics*, vol. 11, pp. 1500–1517, 2023.

[8] S. Bai *et al.*, "Qwen3-vl technical report," *arXiv preprint arXiv:2511.21631*, 2025.

[9] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[10] S. Kudugunta, I. Caswell, B. Zhang, X. Garcia, C. A. Choquette-Choo, K. Lee, D. Xin, A. Kusupati, R. Stella, A. Bapna, and O. Firat, "Madlad-400: A multilingual and document-level large audited dataset," *arXiv preprint arXiv:2309.04662*, 2023.

[11] A. Yang *et al.*, "Qwen3 technical report," *arXiv preprint arXiv:2505.09388*, 2025.

[12] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi, "Clipscore: A reference-free evaluation metric for image captioning," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 7514–7528.

[13] X. Zhai, B. Mustafa, A. Kolesnikov, and L. Beyer, "Sigmoid loss for language image pre-training," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[14] H. Nakayama, A. Tamura, and T. Ninomiya, "A visually-grounded parallel corpus with phrase-to-region linking," in *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020, pp. 4204–4210.

[15] S. Dutta *et al.*, "Multimodal nmt for low-resource hindi language," *arXiv preprint arXiv:2210.05193*, 2022.

[16] N. Goyal *et al.*, "The flores-101 evaluation benchmark for low-resource and multilingual machine translation," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 522–538, 2022.

[17] A. Toral, "Post-editese: an exacerbated translationese," in *Proceedings of MT Summit XVII*, 2019.

[18] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 26 296–26 306.

[19] L. Chen, J. Li, X. Dong, P. Zhang, C. He, J. Wang, F. Zhao, and D. Lin, "Sharegpt4v: Improving large multi-modal models with better captions," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024, arXiv preprint arXiv:2311.12793.

[20] H. Xu, Y. J. Kim, A. Sharaf, and H. H. Awadallah, "A paradigm shift in machine translation: Boosting translation performance of large language models," *arXiv preprint arXiv:2309.11674*, 2024.

[21] J. Bai, S. Bai, Y. Chu, Z. Cui *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.

[22] R. Sennrich, B. Haddow, and A. Birch, "Improving neural machine translation models with monolingual data," *arXiv preprint arXiv:1511.06709*, 2015.

[23] R. Rei, C. Stewart, A. C. Farinha, and A. Lavie, "Comet-22: Unifying evaluation for machine translation and metric evaluation," in *Proceedings of the Seventh Conference on Machine Translation (WMT)*, 2022, pp. 585–597.

[24] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International Conference on Machine Learning*, 2021, pp. 8748–8763.

[25] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert," in *International Conference on Learning Representations*, 2020.

[26] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Zhou, and J. Zhou, "Qwen2-vl: To see the world more clearly," *arXiv preprint arXiv:2409.12191*, 2024.

[27] H. Liu, C. Li, Y. Li, B. Li, Y. Zhang, S. Shen, and Y. J. Lee, "Llava-next: Stronger llms, faster tokenizers, and more modalities," January 2024.

[28] J. Zhu *et al.*, "Internvl3: Exploring advanced training and test-time strategies for open-source multimodal models," *arXiv preprint arXiv:2504.10479*, 2025.

[29] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.

[30] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, "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, 2020, pp. 38–45.

[31] S.-Y. Liu, C.-Y. Wang, H. Yin, P. Molchanov, Y.-C. F. Wang, K.-T. Cheng, and M.-H. Chen, "DoRA: Weight-decomposed low-rank adaptation," in *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.