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

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

[1] School of Computer Science and Technology, Xinjiang University;
[2] Xinjiang Laboratory of Multi-Language Information Technology; [3] Xinjiang Multilingual Information Technology Research Center;
[4] Joint International Research Laboratory of Silk Road Multilingual Cognitive Computing, Urumqi, China
[5] Department of Computational and Applied·Linguistics, National University of Uzbekistan, Tashkent, Uzbekistan;
[6] Faculty of Information Technology and Artificial Intelligence, Al-Farabi Kazakh National University, Almaty, Kazakhstan;
[7] Integrated Laboratory for Space, Air, and Ground Systems, Changji, China

Overview & Motivation

Research Gap

- Existing multimodal MT resources mainly cover high-resource languages.
- Extremely low-resource languages still lack visually grounded parallel corpora.

Challenges

- Translationese artifacts
- Semantic drift
- Weak visual grounding

Our Goal

We introduce SilkRoad-VL, a visually grounded and metric-driven pipeline for building a high-quality benchmark for six extremely low-resource languages: **Uyghur (ug)**, **Kazakh (kk)**, **Kyrgyz (ky)**, **Tajik (tg)**, **Uzbek (uz)**, and **Urdu (ur)**.

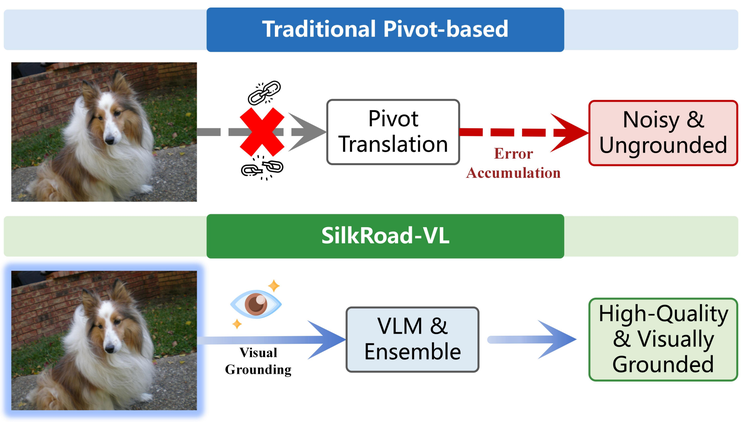

Fig. 1. Paradigm comparison between traditional pivot-based generation and SilkRoad-VL.

METHOD

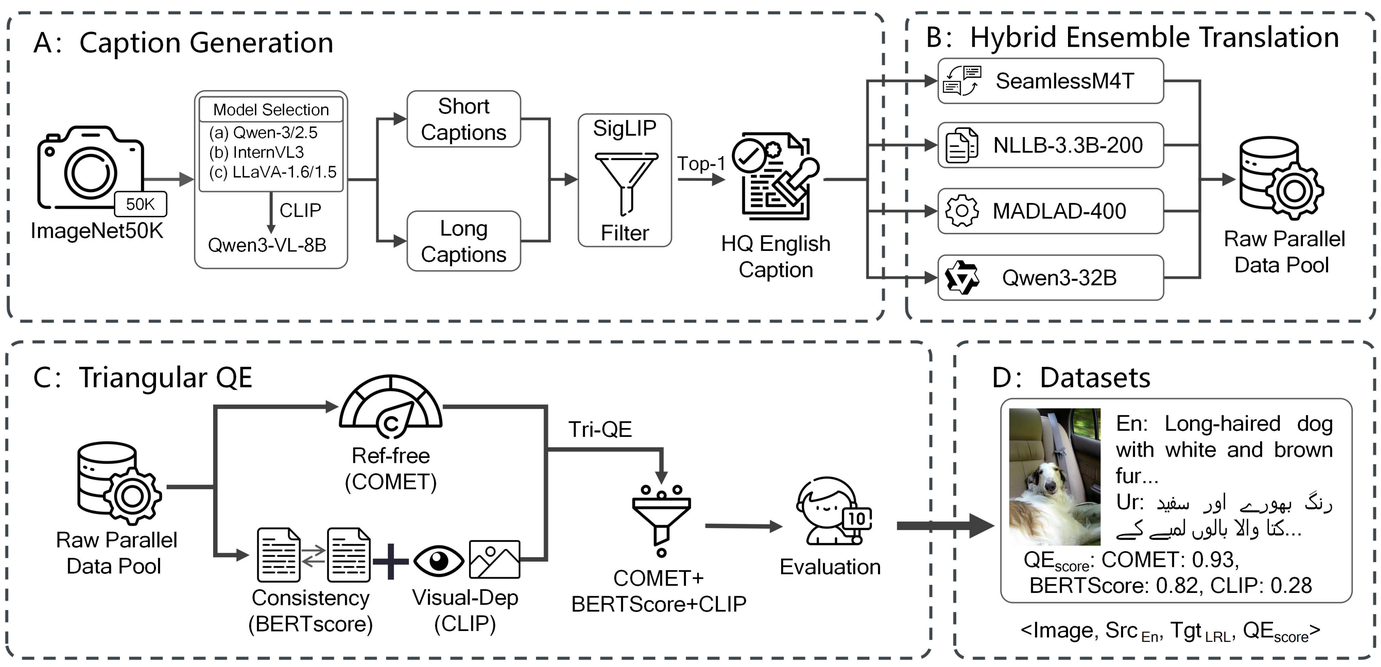

Fig. 2. Overview of the SilkRoad-VL construction pipeline.

Metric-Driven Caption Generation

- **VLM Selection:** We compare candidate VLMs and select Qwen3-VL-8B as the optimal caption backbone.
- **SigLIP Filtering:** We generate short and long captions, then retain visually faithful descriptions through SigLIP-based filtering.

Hybrid Ensemble Translation

- **Diverse Models:** We generate translation candidates using NLLB-200, MADLAD-400, SeamlessM4T-v2, and Qwen3-32B.
- **Complementary Strengths:** The ensemble balances lexical fidelity, multilingual fluency, and multimodal alignment while reducing single-model bias.

Tri-QE Filtering

- **Multi-Dimensional QE:** We jointly evaluate reference-free quality, semantic consistency, and visual grounding using COMET-Kiwi, BERTScore, and CLIPScore.
- **Strict Filtering:** Only candidates that satisfy all three constraints are retained, reducing noisy, low-fidelity, and hallucinated samples.

RESULTS

Main Results on SilkRoad-VL:

- Ours outperforms all strong baselines on all six languages.
- The highest score reaches 62.47 in Kazakh, with 62.33 in Uzbek and 61.24 in Kyrgyz.
- Four languages exceed 60 points, indicating stable cross-lingual gains.
- Tajik shows the largest improvement, rising from 46.61 to 58.25 (+11.64).
- These results confirm the effectiveness of SilkRoad-VL as a high-quality supervision source.

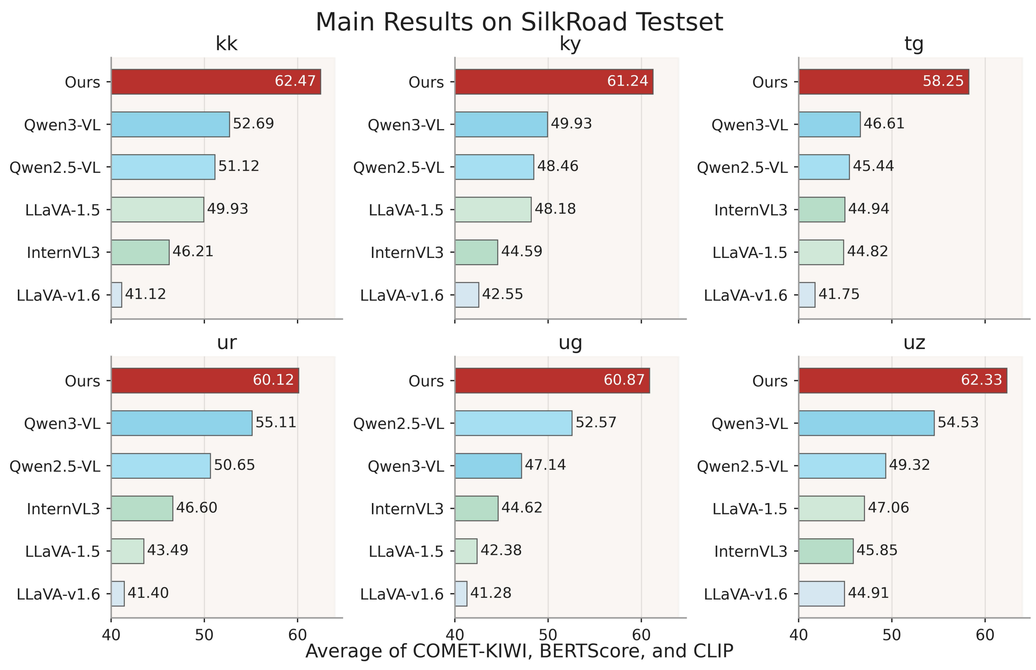

Fig. 3. Main results on the SilkRoad-VL test set across six languages.

Zero-shot Transfer to External Benchmarks:

- Multi30K: Tajik improves from 26.57 to 40.20, with Kazakh reaching 50.36.
- Visual Genome: Urdu improves from 37.97 to 48.31, and Tajik reaches 37.05.

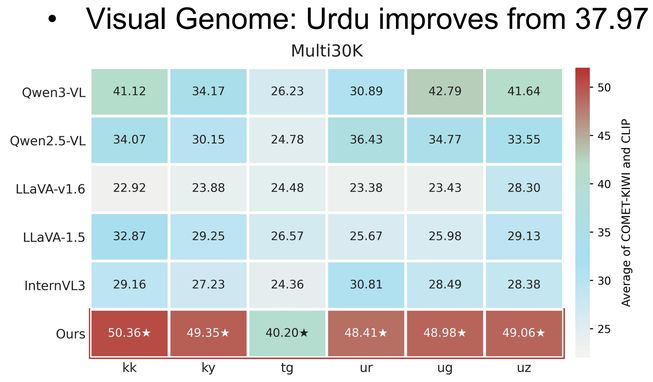

Fig. 4. Zero-shot transfer performance on Multi30K across six languages.

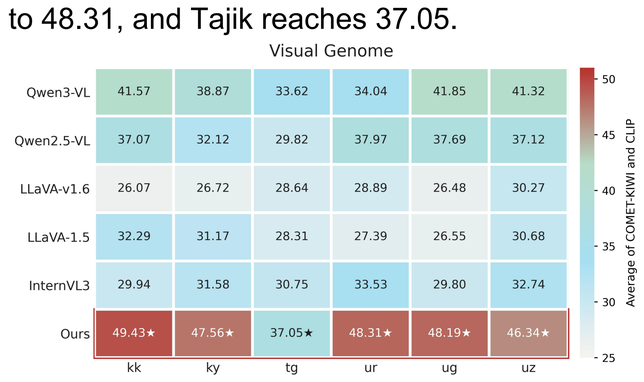

Fig. 5. Zero-shot transfer performance on Visual Genome across six languages.

Human Evaluation: Our outputs achieve 4.46 in Fluency, 4.31 in Adequacy, and 4.71 in Visual Relevance, confirming strong linguistic quality and visual grounding.

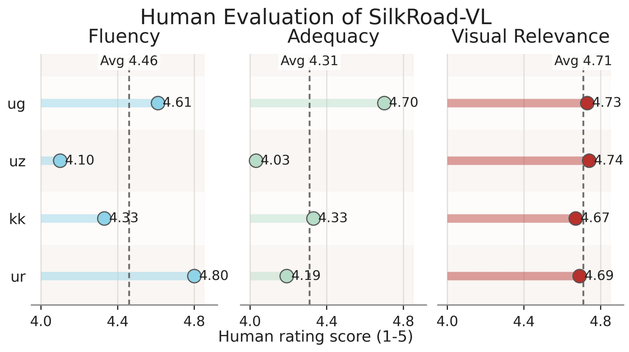

Fig. 6. Human evaluation on fluency, adequacy, and visual relevance.

Corpus Scale: Final Corpus Scale: SilkRoad-VL contains 84.9k pairs, split into 82.7k train, 0.4k dev, and 1.8k test samples.

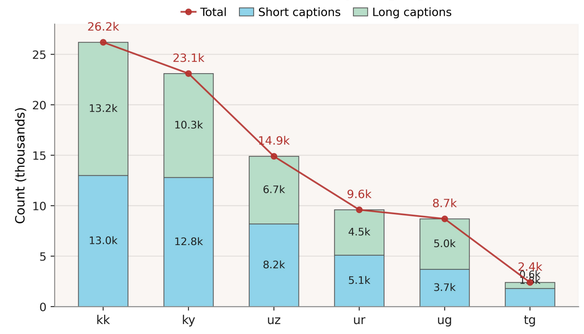

Fig. 7. Language-wise distribution of the final SilkRoad-VL corpus.

CONCLUSION

- We construct SilkRoad-VL, a high-quality multimodal benchmark for six extremely low-resource languages.
- Metric-driven captioning, hybrid ensemble generation, and Tri-QE substantially improve data quality.
- Fine-tuned models generalize well to Visual Genome and Multi30K.
- Future work will extend the pipeline to more low-resource languages and additional modalities.

Acknowledgement: Supported by NSFC (Grant Nos. 62406316, 201704041014, and 201404041254) and the Xinjiang "Tianchi Talent" Recruitment and Introduction Program.
*__Corresponding author:__ miradeljan51@xju.edu.cn