Proceedings of the 5th International Conference on Interdisciplinary Humanities and Communication Studies
DOI: 10.54254/2753-7064/2026.38412

Visual Context and AI Translation Accuracy in Chinese Museum Object Descriptions: An Exploratory Paired Comparison of Text Only and Text Plus Image Conditions

Jinqi Kang

School of Foreign Languages, Nankai University, Tianjin, China
kang.jinqi@mail.nankai.edu.cn

Abstract. A photograph can make an object feature easier to inspect without resolving how that feature should be translated. This study examines that problem in 12 Chinese museum descriptions through a paired comparison of 24 retained English outputs from the National Museum of China. An adapted Multidimensional Quality Metrics scheme separates confirmed accuracy and terminology errors from unresolved judgements, recorded as reservations. The text-only condition contained 11 confirmed errors and 7 reservations; the text-plus-image condition contained 9 errors and 13 reservations. Accuracy errors numbered 9 in each condition, and terminology errors 2 and 0. Five samples had fewer confirmed errors with an image, five had the same count, and two had more. Case analysis shows contrasting changes: some outputs retain quantities and local forms more fully, while others omit details or substitute an animal category. Technical identification still depends on terminology and curatorial evidence. Checking was assisted by artificial intelligence (AI), with condition labels visible and no second independent evaluator. Because the comparison uses single outputs under uncontrolled generation conditions and leaves some judgements unresolved, it cannot establish an overall quality ranking or attribute the differences to image input. Its contribution is an account of specific translation decisions that require review.

Keywords: multimodal translation, museum translation, AI-assisted translation, visual context, translation quality

1. Introduction

Museum photographs can clarify visible form without establishing the correct technical term. A pendant may appear flat rather than ring-shaped, while identifying a vessel as a steamer requires evidence about function. This distinction matters when vision-language models translate Chinese collection descriptions containing specialised vocabulary and relations among object parts.

Visual context can help when textual information is restricted [1], but performance gains need not demonstrate image use [2]. On-demand visual reasoning likewise treats images as a resource to consult selectively [3]. The present comparison examines observable translation differences without assuming that image input caused them.

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-7064/2026.38412

The study selects a concretizing image-text relation: a source statement names a feature whose form, position, or arrangement is potentially visible in the photograph. Informed by Martinec and Salway [4], this is an operational sampling rule, not a new canonical category.

Two questions guide the analysis: (1) How do confirmed accuracy and terminology errors differ between text-only and text-plus-image outputs? (2) Which descriptions of shape, ornament, structure, and spatial relations are preserved, lost, or replaced? Unresolved judgements are recorded separately as reservations.

2. Literature review

Multimodal translation involves resources with different meaning potentials [5], requiring precise definitions [6]. Museum translation can redistribute meaning across verbal, visual, and spatial modes [7]. Research on museum computer vision also identifies limits to automated interpretation [8]; recognising a feature does not necessarily establish its specialised English name.

Image-based translation has shown lower error counts in some caption tasks [9], while surveys call for challenging data and targeted evaluation [10]. Work on Chinese museum communication considers broader AI-supported multimodal mediation [11]. These findings motivate item-level assessment of museum descriptions, but do not establish a general benefit from adding photographs.

The adapted Multidimensional Quality Metrics (MQM) scheme [12] uses mutually exclusive categories. Accuracy covers mistranslation, omission, or unsupported addition concerning shape, number, location, ornament, or function. Terminology covers incorrect identification of an object or component supported by museum evidence. A terminology error is not counted again as accuracy.

Major errors alter core class, decisive function, or a key quantity relation; minor errors affect local properties while identity remains recoverable. A confirmed error requires a direct conflict, definite omission, or supported terminology judgement. Reasonable alternative readings remain reservations. Fluency and stylistic preference are excluded.

3. Method

3.1. Materials and sampling

The dataset comprises twelve National Museum of China descriptions: nine bronzes and three ceramics. The project record gives 19 August 2026 as the access date; Appendix A lists the source pages. Excerpts contain 50–67 Chinese characters, including punctuation and numerals, and may combine non-contiguous passages about the selected features.

Purposive selection required a corresponding photograph and a potentially visible shape, ornament, connection, quantity, or location. Selection preceded review of the English outputs. Official collection information and comparable museum terminology supported checking; no existing English description served as the sole acceptable translation.

3.2. Input conditions and retained generation record

Condition A contained the Chinese source text only, while Condition B contained the same source text together with the corresponding object image. Both conditions used the same instruction: "Please translate the following Chinese museum description into English for publication on an international museum website. Preserve all factual and cultural information and do not add unsupported information."

DOI: 10.54254/2753-7064/2026.38412

The project record identifies ChatGPT with GPT-5 and dates generation to 19 August 2026. Samples were processed separately, always A before B, with no recorded follow-up, regeneration, or post-editing. Complete logs, exact image files, and stochastic settings were not independently verified. Separate samples do not establish separate sessions for A and B; carry-over and ordinary generation variability remain unresolved. Only one output per condition was retained, and B had no additional image-use instruction.

3.3. Coding and analysis

Each independently judgeable source proposition was paired with its English wording and a category, severity, and justification, or a reason for reservation. A single quantity relation counted once; independently stated shape and location could count separately. Paraphrase alone was not penalised.

Checking was AI-assisted and unblinded, without a second independent human evaluator. The 24 retained outputs support descriptive counts and close reading, not an overall error-rate estimate. Multiple errors within an output are not independent observations.

4. Results

Table 1 reports 11 confirmed errors in A and 9 in B. Accuracy errors remain 9 in each; terminology errors are 2 and 0. Category and severity describe the same inventory and must not be added together.

Table 1. Aggregate error distribution

Measure	A text only	B text plus image	B minus A
Accuracy	9	9	0
Terminology	2	0	-2
Major errors	2	1	-1
Minor errors	9	8	-1
Confirmed errors	11	9	-2
Reservations	7	13	No inference

Reservations number 7 in A and 13 in B. They are neither confirmed errors nor confirmed correct items. Without an inventory of all information units, the counts cannot yield an overall error rate or percentage quality improvement.

B has fewer confirmed errors in five samples, unchanged counts in five, and more in two (Table 2). Lower counts may coexist with unresolved wording, as in S11; unchanged counts may reflect shared retention or shared omissions.

Table 2. Sample-level confirmed errors and reservations

ID	A errors	B errors	Reservations A / B	Direction
S01	2	0	1 / 2	Fewer
S02	0	0	0 / 0	Unchanged

DOI: 10.54254/2753-7064/2026.38412

Table 2. (continued)

S03	3	3	1 / 1	Unchanged
S04	0	0	1 / 1	Unchanged
S05	3	0	1 / 1	Fewer
S06	1	0	0 / 0	Fewer
S07	0	3	0 / 4	More
S08	0	3	1 / 0	More
S09	1	0	1 / 1	Fewer
S10	0	0	0 / 0	Unchanged
S11	1	0	1 / 3	Fewer
S12	0	0	0 / 0	Unchanged

5. Case analysis

5.1. Information preserved more fully in the image condition

In S05, A's "rope-patterned loops" shifts component shape to decoration; "Thick bands with raised diamonds" omits central position; and "field-shaped panels" leaves the four-part arrangement unspecified. B retains "rope-twisted rings", "a raised diamond at the center", and "a four-compartment grid". Handle location and number remain reserved in both. No extra error is assigned merely because A lacks B's verb "suspends".

In S06, A's "Four animal-head ring handles" reduces four pairs to four handles; B preserves a pair on each side. Number and distribution constitute one major accuracy error. Both omit the historical comparison identifying the vessel as the largest known of its kind from the Shang and Zhou periods; that omission falls outside the visually relevant inventory.

In S09, A calls the upper zeng components "cauldrons"; B uses "steamers". The museum account describes food above water heated below, supporting one major terminology error in A. Shape alone cannot establish this function.

In S01, A's "hanging rings" conflicts with flat pendant components identified in the report's photographic check; B uses "pendant tabs". A also omits the animal-mask decoration on the feet. These are one minor terminology and one minor accuracy error. "Tile patterns" remains reserved in both, as does B's "Repeated-ring patterns". The later photograph check does not verify the exact generation image.

In S11, A's "long neck" omits straightness; B's "long upright neck" remains reserved because orientation need not express linear shape. Thus, zero confirmed errors in B does not establish correctness. Object and foot labels also remain unresolved; the comparable museum object cited in Appendix A prevents dismissing A's "bell-shaped zun" outright.

5.2. Shared retention and shared omissions

S02, S10, and S12 have no confirmed errors or reservations in the assessed inventory. S03 instead shares three omissions: the rounded inward transition, near-flat base, and straight inner legs; rim wording remains reserved. S04 preserves the assessed structure but leaves the qiequ motif reserved.

DOI: 10.54254/2753-7064/2026.38412

Both S04 outputs omit a historical classification outside the inventory. These outcomes concern selected features, not whole-text quality.

5.3. New deviations in the image condition

In S07, B substitutes "ram-like head" for the source's explicit horse comparison, producing its only major error. It also omits the flattened body and front-narrow, rear-broad profile. Four points remain reserved: leg number, paired horns, single-hoof distinction, and the scope of "four bird-like talons". Curled horns might suggest a ram, but the wording cannot be causally attributed to image input.

In S08, B replaces a character-shaped eye outline with "round protruding eyes", omits the slightly projecting chest, and fails to express gripping in "four-toed feet" and "stable stance". These are three minor accuracy errors; "stable stance" is not counted again. A preserves the assessed features but has a reservation about the added relation "above a broad drooping tail".

6. Discussion

Six new accuracy errors in S07 and S08 offset six absent from B elsewhere, leaving equal accuracy totals despite different information losses. The terminology difference likewise requires case-specific evidence: a photograph supports inspection of pendant shape, whereas identifying a steamer requires curatorial information. Neither pattern proves that the model attended to the image.

Reservations preserve uncertainty that a forced error-or-correct decision would conceal. S11's labels, S03's rim, and S07's horns and toes require further evidence or clearer explicitness criteria. B's higher reservation count does not establish that images caused uncertainty. Independent qualified evaluators should revisit these units using source statements, photographs, and institutional terminology.

The cases suggest checking English claims against the Chinese text first, consulting photographs for visible properties, and museum documentation for technical identification. Class, function, number, and explicit source-image conflicts deserve attention before stylistic polishing. This proposed review sequence remains untested. Following on-demand visual reasoning [3], future experiments could compare selective image consultation with routine image input.

7. Conclusion

Across twelve paired descriptions, image-condition outputs retain some local forms and quantities more fully but also introduce omissions and an animal-category substitution. The 11-versus-9 confirmed-error difference supports case-level review, not an overall quality ranking. Visual inspection and technical identification require different evidence, and reservations mark unresolved decisions.

The purposive, bronze-heavy sample, single output per condition, fixed A-before-B order, incomplete generation records, and AI-assisted unblinded coding without independent replication prevent causal attribution to image input. Future work should archive all inputs and outputs, counterbalance conditions, repeat generation, and use independent qualified evaluators.

References

[1] Caglayan, O., Madhyastha, P., Specia, L., & Barrault, L. (2019). Probing the need for visual context in multimodal machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for*

DOI: 10.54254/2753-7064/2026.38412

Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4159–4170). https://doi.org/10.18653/v1/N19-1422

[2] Wu, Z., Kong, L., Bi, W., Li, X., & Kao, B. (2021). Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 6153–6166). https://doi.org/10.18653/v1/2021.acl-long.480

[3] Liu, D., Kong, F., Sun, X., Patil, D., Vajpayee, A., Liu, Z., Bhat, V., & Sadoughi, N. (2025). Detect, disambiguate, and translate: On-demand visual reasoning for multimodal machine translation with large vision-language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 1559–1570). https://doi.org/10.18653/v1/2025.naacl-long.74

[4] Martinec, R., & Salway, A. (2005). A system for image-text relations in new (and old) media. *Visual Communication, 4*, 337–371. https://doi.org/10.1177/1470357205055928

[5] Adami, E. (2023). A social semiotic multimodal approach to translation. In *The Routledge Handbook of Translation Theory and Concepts*. Routledge, London.

[6] Valdeón, R. A. (2024). The translation of multimodal texts: Challenges and theoretical approaches. *Perspectives: Studies in Translation Theory and Practice, 32*, 1-13. https://doi.org/10.1080/0907676X.2024.2290928

[7] Liao, M.-H. (2023). Translation as a practice of resemiotization: A case study of the Opium War Museum. *Translation Studies, 16*, 48-63. https://doi.org/10.1080/14781700.2022.2103024

[8] Villaespesa, E., & Murphy, O. (2021). This is not an apple! Benefits and challenges of applying computer vision to museum collections. *Museum Management and Curatorship, 36*, 362-383. https://doi.org/10.1080/09647775.2021.1873827

[9] Calixto, I., & Liu, Q. (2019). An error analysis for image-based multi-modal neural machine translation. *Machine Translation, 33*, 155-177. https://doi.org/10.1007/s10590-019-09226-9

[10] Sulubacak, U., Caglayan, O., Grönroos, S.-A., Rouhe, A., Elliott, D., Specia, L., & Tiedemann, J. (2020). Multimodal machine translation through visuals and speech. *Machine Translation, 34*, 97-147. https://doi.org/10.1007/s10590-020-09250-0

[11] Yang, C., & Cui, F. (2023). Multimodal mediation in translation and communication of Chinese museum culture in the era of artificial intelligence. *Corpus-Based Studies Across Humanities, 1*, 51-77. https://doi.org/10.1515/csh-2023-0020

[12] Lommel, A. R., Uszkoreit, H., & Burchardt, A. (2014). Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica, 12*, 455-463. https://doi.org/10.5565/rev/tradumatica.77

DOI: 10.54254/2753-7064/2026.38412

Appendix A. Source pages

These source pages do not archive the exact generation images. Inputs, outputs, and coding justifications are recorded in the Final Report. The S11 comparison is https://www.metmuseum.org/art/collection/search/50717 (checked 28 August 2026).

Table A1. National museum of China source pages

ID	Source URL
S01	https://www.chnmuseum.cn/zp/zpml/csp/202209/t20220901_257142.shtml
S02	https://www.chnmuseum.cn/zp/zpml/csp/202112/t20211221_253289.shtml
S03	https://www.chnmuseum.cn/zp/zpml/csp/202209/t20220901_257145.shtml
S04	https://www.chnmuseum.cn/zp/zpml/kgfjp/202209/t20220901_257148.shtml
S05	https://www.chnmuseum.cn/zp/zpml/csp/202209/t20220905_257250.shtml
S06	https://www.chnmuseum.cn/zp/zpml/csp/202008/t20200826_247376.shtml
S07	https://www.chnmuseum.cn/zp/zpml/kgfjp/202107/t20210728_250875.shtml
S08	https://www.chnmuseum.cn/zp/zpml/kgfjp/202107/t20210728_250872.shtml
S09	https://www.chnmuseum.cn/zp/zpml/kgfjp/202107/t20210728_250870.shtml
S10	https://www.chnmuseum.cn/zp/zpml/csp/202203/t20220311_254257.shtml
S11	https://www.chnmuseum.cn/zp/zpml/csp/202203/t20220309_254248.shtml
S12	https://www.chnmuseum.cn/zp/zpml/csp/202203/t20220315_254330.shtml