Towards Automatic Evaluation for Image Transcreation

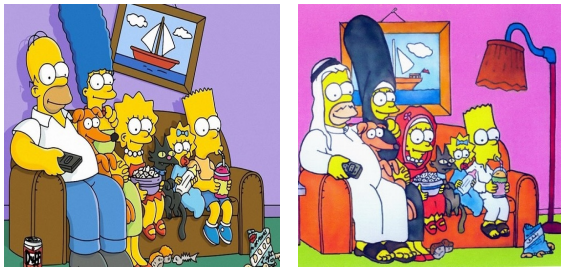
Anonymous NAACL submission

Abstract

Beyond conventional paradigms of translating speech and text, recently, there has been interest in automated transcreation of *images* to facilitate localization of visual content across different cultures. Attempts to define this as a formal Machine Learning (ML) problem have been impeded by the lack of automatic evaluation mechanisms, with previous work relying solely on human evaluation. In this paper, we seek to close this gap by proposing a suite of automatic evaluation metrics inspired by machine translation (MT) metrics, categorized into: **a)** *Object-based*, **b)** *Embedding-based*, and **c)** *VLM-based*. Drawing on theories from translation studies and real-world transcreation practices, we identify three critical dimensions of transcreation: *intent*, *cultural relevance*, and *creativity*, and design our metrics to evaluate systems along these axes. Our results show that proprietary VLMs best identify *intent* and *cultural relevance*, while vision-encoder representations are adept at encoding *creativity*. Meta-evaluation across 7 countries shows our metrics agree strongly with human ratings, with average segment-level correlations ranging from 0.55-0.87. Finally, through a discussion of the merits and demerits of each metric, we offer a robust framework for automated image transcreation evaluation, grounded in both theoretical foundations and practical application.

1 Introduction

While MT has made commendable progress in the text and speech modalities (Kocmi et al., 2024; Ahmad et al., 2024), there has recently been interest in the ML community to approach translation in the *visual* modality. This new task, called **image transcreation**, seeks to adapt (or localize) images originating from a *source culture* for a given *target culture* in an attempt to enhance familiarity and relatability for the target audience (Khanuja et al., 2024). We provide a popular real-world example of image transcreation in Figure 1, where content was recreated to match cultural norms of a target audience. While this problem is well-explored in Translation Studies (Section 2) and is carried out widely in industrial settings by human translation experts (Section 3), it has not yet taken off in ML, despite the availability of powerful Vision-Language Models (VLMs). A central reason for this is the lack of automated evaluation mechanisms, with prior works relying solely on expensive human evaluation. Khanuja et al. (2024), for instance, report spending about 4000 USD for annotating a test set of 3500 images across 7 countries, with only 1 annotator per image. While such a contribution is indeed beneficial, as ML researchers, we naturally need automated metrics to train, validate and benchmark image transcreation systems at scale.



(a) Source: United States (b) Target: Middle East

Figure 1: Image Transcreation in the real world: *The Simpsons* (1989) was remade into *Al-Shamshoon* (2005) for a Middle-Eastern audience.

And yet, the very nature of the task makes it difficult to propose a singular automated metric. Unlike traditional Natural Language Processing (NLP) problems like summarization or translation that have a relatively small space of valid outputs, there can be multiple valid ways of transcreating images. The suitability of an adaptation depends on factors like the *intent* of the user, the *creativity* of the transcreator and the degree of *cultural relevance* for the target audience. This subjectivity makes it hard to define a singular metric like BLEU (Papineni et al., 2002), ROUGE (Lin, 2004),

or even neural metrics like COMET (Rei et al., 2020) that optimize for similarity of model outputs against a fixed set of reference(s).

In this paper, we make progress towards building and evaluating automated metrics for image transcreation. First, given that this is a novel task in ML, we turn to Translation Studies for a better understanding of its scope and variables (Section 2). We identify key dimensions critical to define a transcreation process: *intent*, *creativity* and *cultural-relevance*, and look at real-world applications of transcreation through the lens of these dimensions in Section 3. Grounded in the conceptual and practical insights from previous sections, and drawing inspiration from MT metrics for text, we propose three metrics that account for the varied dimensions of this subjective task – **a)** *Object-based*; **b)** *Embedding-based* and **c)** *VLM-based* metrics, each of which leverages powerful open and closed VLMs and Large Language Models (LLMs). After conducting meta-evaluation along all three dimensions of human ratings, we report average segment-level (i.e. instance-level) correlations of up to 0.87 on image transcreation test-sets spanning 7 countries. Thus, our primary contributions are:

1. We propose three image transcreation metrics that are **a)** *Object-based*; **b)** *Embedding-based* and **c)** *VLM-based* that achieve strong segment-level correlation, and to the best of our knowledge, are the first set of automatic evaluation metrics for this task.

2. We assess the capabilities of powerful open and closed VLMs like Gemini, GPT-4o and Molmo to recognise, compare and reason over culturally specific objects in images, contributing to growing research in this area;

3. Finally, we widen the scope of image transcreation for ML researchers by drawing from Translation Studies literature and elucidating complex real-world examples of this task. We hope this helps advance ML research on this challenging but important problem.

2 Transcreation in Translation Studies

Transcreation has been widely studied by linguists and philosophers in translation studies. We trace its development in this linguistics-adjacent field below, to ground our evaluation in this research.

Definitions of Transcreation: Transcreation, a concept introduced by the translation industry in the last century, has been developed to help clients of translation service providers tailor marketing campaigns for foreign markets (Ray and Kelly, 2010), and is generally commonly associated with creative industries like advertising and marketing. With its growing popularity, various definitions were proposed to underscore its difference from translation. Today, transcreation is generally described as a blend of linguistic translation, cultural adaptation, and creative reinterpretation of text elements (Díaz-Millón and Olvera-Lobo, 2023). Given its "cultural" and "creative" aspects, it is also seen as a hybrid practice between "translating" and "copywriting" (Benetello, 2018). As Ray and Kelly (2010) emphasize, transcreation focuses on effectively conveying the original message in a way that resonates with the target audience, even if this requires substantial deviations from the original text or imagery. The source text can be dethroned to apply focus on the target culture.

Parameterizing Transcreation: Broadly, as Venuti (2017) discusses, translation often navigates between the dichotomy of *foreignization*, which emphasizes preserving the source culture, and *domestication*, which adapts the content to fit the target culture. Various terms have been used to describe the ends of this spectrum, such as "literal" vs. "free", "formal" vs. "dynamic", and "semantic" vs. "communicative". Yet, no translation decision is ever solely one or the other (Venuti, 2017) and translators continuously navigate this spectrum, weighing several critical factors.

In transcreation, this dichotomy is even more pronounced, and there is a strong emphasis on making the content feel relevant and engaging in a new cultural context, especially in areas like marketing and advertising. This often involves adjusting or replacing Culturally Specific Items (CSIs), where cultural references from the source material may not translate well into the target culture (Aixelá, 1996). The goal is to make the adapted content not only *culturally relevant* but also effective in reaching and persuading the audience.

A key factor in guiding the transcreation beyond cultural relevance, is the *intent* behind the process. Skopos theory (Reiss et al., 2014) emphasizes that all translation activities are intrinsically purpose-driven. In the context of transcreation, the intent can be as broad as the overall goals of a marketing campaign or as specific as the instructions in a project brief (Díaz-Millón and Olvera-Lobo, 2023).

Figure 2: The Coca-Cola homepage depicts different cultural concepts in different countries (TranslatePlus, 2023). The top row shows Marvel-themed ads where visuals are consistent but the text is translated for accessibility, while the bottom one depicts the Lunar New Year to celebrate the Year of the Rabbit in Korea.

Whether the aim is to inform, persuade, or entertain, the *intent* determines how much of the original material is adapted to suit the target audience's needs (Ray and Kelly, 2010).

Finally, transcreation requires a high degree of *creativity*. Creativity is the core component that drives the transcreation process, extending beyond accuracy of faithfulness to the source, to involve a comprehensive reimagining of content (Martín, 2024). Creativity becomes the mechanism through which intent and cultural relevance are aligned. As Nord (2009) points out, this creative process often demands unique, non-replicable solutions tailored to the specific needs of each project.

3 Visual Transcreation in the Wild

While text translation operates in a relatively constrained and predictable output space, image transcreation allows for a broader range of creative expression and the desired output depends on the user's objectives and application context. From the references above, we identify three components critical to the transcreation process: **a)** *intent* of the user, application, or task; **b)** *creativity* involved in adapting the source content; and **c)** *cultural relevance* to the target audience. In this section, we examine how visual content is transcreated in the real world for varied applications, and present findings through the lens of these three components.

3.1 Advertisements

Intent: *to emotionally relate to the target audience, while preserving the core brand identity*

Given the strong focus on being relatable to the target audience to maximize product sales, we often see ad transcreations span a broad spectrum of ***creativity***. A well-known example is Coca-Cola, which embodies its slogan, "Think Global, Act Local", by maintaining brand consistency while seamlessly integrating cultural elements into its ads. For example, they released a Lunar New Year ad in Korea to celebrate the year of the rabbit (Figure 2). In a particularly ambitious example, they transcreated an entire video ad, using different actors speaking in different languages, but with the script being exactly the same otherwise (Coca-Cola, 2023). Beyond region-specific adaptations, some ads also focus on universal themes while still targeting specific audiences, such as their Marvel-themed campaign (Figure 2). When released in different regions, these are minimally transcreated, often only translating the text while keeping the visuals unchanged. In some cases, subtle visual differences can emerge across regions, even when ads present similar concepts. For instance, in Figure 3, both Slack ads promote hybrid work, but the Japanese and U.S. versions modify the names and avatars to better reflect the demographics of each country.

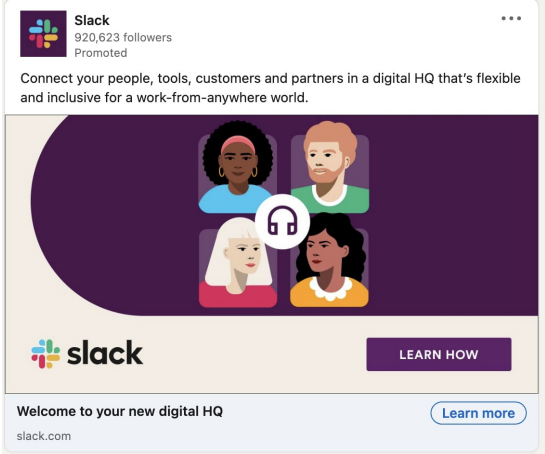
Overall, we observe a strong focus on ***cultural relevance***, where culture is not strictly defined by geographic boundaries but can be connected by shared preferences. Transcreators take the liberty to exercise creative freedom in adapting ads to culturally-specific contexts.

3.2 Comics

Intent: *to adapt the story and characters in a way that remains faithful to the original while making them accessible to a new audience*

The transcreation of Japanese manga for Western audiences shows how preferences for these adaptations have evolved over time. In early days, U.S. publishers believed that altering key elements like reading direction would make it more accessible. In *Akira*, for instance, pages were flipped for left-to-right reading, but this caused significant cultural errors. Japanese cars were incorrectly depicted as driving on the right, and characters were shown left-handed, even in situations where this affected key plot points, such as losing their dominant hand in a fight (Gianordoli and Ito, 2023). Over time,

(a) Japan advertisement



(b) US advertisement

Figure 3: Both Slack ads convey how the product can be used to facilitate hybrid work, but the names and avatars of users are localized to reflect the demographics of each country. Figures sourced from Olaussen.

it became clear that ***fidelity is strongly preferred over creativity*** in the transcreation of comics, particularly by manga fans[1,2].

However, there are still instances where adaptation is unavoidable, often driven by the need to align with social norms, legal requirements, cultural differences, or when concepts are completely absent in the target culture. For example, when Disney comics like *Tarzan* and *Donald Duck* were adapted for Kuwait and the UAE, changes were made to clothing to conform to local cultural expectations (Figure 4). Similarly, in Marvel's *The Punisher*, a swastika tattoo on a villain was replaced with a parallelogram when adapted for Germany, adhering to the country's strict legal restrictions on the depiction of Nazi symbols (Zanettin et al., 2014). In the US TV adaptation of the manga, *Sailor Moon*, the character of *Artemis*, a cat, was creatively adapted to match the stereotypes and perceptions of American culture (Figure 5) – showing transcreation also needs to factor in cultural ideas and beliefs. Finally, we note that Japanese manga often uses onomatopoeia to represent sounds, which is less common in Western comics. For instance, in Tsuge Yoshiharu's *The Swamp*, the sound of a rifle crack was changed from "zdom" to "kbooom, a term more familiar for the target audience (Gianordoli and Ito, 2023).

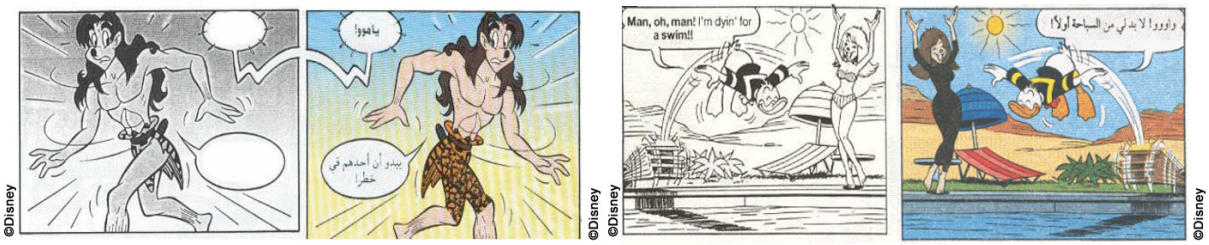
Overall, we note that while readers typically prefer transcreations that remain true to the original, adaptations are sometimes necessary to meet social norms, legalities, or cultural differences. Transcreators walk a fine line between staying faithful and making those critical adjustments.

3.3 Children's Media

Intent: *to aid learning and cognitive development through familiar objects, concepts and experiences*

Several studies highlight how incorporating familiar imagery and references improves cognitive engagement for children by making abstract concepts easier to grasp (Unicef et al., 2018; Purcell-Gates et al., 2011). We see several examples of this, for example in Pixar and Disney movies, like in *Zootopia* (2016) where the character of the newscaster was adapted to match a country's national animal (Figure 6). Another prominent example is *Dora the Explorer*, where Dora switches between Hindi and English in its adaptation for India, references local animals like the tiger and peacock, and also talks about issues affecting girl children like early marriage, and unequal distribution of household chores (Khalid, 2012). This approach not only promotes bilingual learning and concept-learning, but also empowers children to challenge social norms. *Sesame Square*, Sesame Street's adaptation in Nigeria, similarly had Muppets who spoke with a Nigerian English accent and tackled weighty issues such as HIV/AIDS discrimination, hygienic practices and ways to cope with death

[1] https://tinyurl.com/58vzy6xu
[2] https://tinyurl.com/5c6wua42



(a) Brad of the Jungle (Kuwaiti version)

(b) Donald Duck (UAE adaptation)

Figure 4: Examples of transcreation to adhere to cultural norms. *Left*: Tarzan covers his thighs with culturally appropriate clothing; *Right*: woman is completely dressed in black.

Figure 5: *Transcreation reflecting cultural differences* – in the U.S. adaptation of *Sailor Moon*, *Artemis* is portrayed as cunning and mischievous ("catty"), while in Japan, cats are seen as part of the cute *kawaii* culture.

Figure 6: The newscaster in the Pixar movie "Zootopia" was transcreated based on the country: a *moose* for the US/Canada/France, a *panda* for China, a *koala* for New Zealand/Australia, and a *jaguar* for Brazil.

(Sangillo, 2019). Beyond movies and TV, educational apps like Toca Boca also incorporate region-specific foods (like sushi in Japan), festivals (like the Lunar New Year in China), and clothing styles, to ensure that children are interacting with culturally familiar elements while learning new concepts.

4 Towards Automatic Evaluation of Image Transcreation

Khanuja et al. (2024) conduct a multi-faceted human evaluation to assess image transcreation systems. Despite being a great first step, a key limitation of the human evaluation questions they design is the lack of grounding in related work from adjacent fields, as well as their specificity to the test set they create. From the theories we study in Section 2, and the applications we explore in Section 3, we identify three factors critical to a transcreation process, namely *intent*; *creativity* and *cultural relevance*. Hence, we must design metrics that can capture all three dimensions.

But how do we design such metrics? Similar to text-level translation that aims to transfer source-side semantics to a target language, during image transcreation, the goal is to recreate visual ideas for a target culture. Tracing the history of MT evaluation, we observe that it first started with human annotators evaluating the adequacy, fluency, and overall quality of the translation. Given high costs and time latency associated with this process, automatic metrics began to develop, which can be broadly categorized into lexical-based, embedding-based, and more recently, LLM-based metrics (Fu et al., 2023; Liu et al., 2023). Lexical-based metrics measure similarity between the hypothesis and reference by a mathematical formula, and are word-based or character-based depending on the match-

ing units (BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005)). Embedding-based metrics use dense vector representations of hypotheses and references to measure their similarity (Zhang* et al., 2020; Yuan et al., 2021; Lo, 2019), and metrics like COMET (Rei et al., 2020) also train regression models on human judgments of translation quality. Finally, LLM-based metrics have recently emerged, leveraging LLMs to evaluate translations by using their contextual understanding and reasoning capabilities. For example, GPTScore (Fu et al., 2023) showed how GPT could be prompted to score hypotheses. Further, G-Eval (Liu et al., 2023) leveraged GPT-4 for evaluating generation based on open-ended reasoning, while Gemba-MQM (Kocmi and Federmann, 2023) used the same model to detect translation error spans and achieved SOTA system-level correlation in the WMT'23 Metrics Shared Task.

Drawing from the above lineage, a parallel evolution of metrics for image transcreation is listed below:

4.1 Object (lexical) - based metrics

Inspired by BLEU (Papineni et al., 2002), we propose an object-based metric that evaluates how well the model replaces culturally-salient objects with valid replacements, given the intent and target demographic. Concretely, this would involve:

- **Step 1: Object Identification**: Identify all objects in both the source and target (model output) images.
- **Step 2: Obtain Valid Translations**: Conditioned on the *intent* of transcreation, identify the culturally-specific items (CSIs) (Aixelá, 1996) that would be unfamiliar to the target culture. Subsequently, determine a valid set of replacements for each CSI based on the constraints of this intent.
- **Step 3: Metric Calculation**: Calculate the proportion of CSIs that have been correctly replaced. This is determined by checking whether the target CSI is a valid replacement of the source CSI according to *Step 2*.

For example, let's say we are transcreating a US-based image for Japan, and the source image contains three objects identified as CSIs: broccoli, spoon, and a dollar bill. Let's say that given a generic intent of *relatability to the target audience*, valid replacements are identified as: **a)** broccoli: {bell pepper, radish}; **b)** dollar bill: {yen note}; **c)** spoon: {chopsticks}.

Now, consider the model output contains the following objects: bell pepper (correct replacement for broccoli), spoon (not replaced) and a euro bill (incorrect replacement for dollar bill). Given the model correctly replaces one out of three CSIs, we get a score of 0.33.

$$\text{CSI-Overlap} = \frac{1+0+0}{3} = \frac{1}{3} \approx 0.33$$

4.2 Embedding-based Metrics

In MT, embedding-based metrics compare contextual embeddings of the target and reference texts. For image transcreation, collecting large amounts of parallel data seems infeasible at present. Instead, we can leverage powerful vision-language models (VLMs) to evaluate transcreation outputs across the three axes discussed in Section 3:

a) *intent*: can be measured by the similarity between the transcreated image's visual representation V_{target} and the textual representation of the intended message T_{int}.

b) *creativity*: can be measured by the similarity between visual representations of the source and transcreated images. A high similarity indicates that the image has stayed close to the source, while a lower similarity suggests a greater degree of adaptation.

c) *cultural-relevance*: can be measured by evaluating the change in similarity between the visual embeddings of the source V_{source} and transcreated image V_{target} relative to a cultural reference text T_{culture}, such as "This image is from region `X`." A higher increase in similarity indicates a greater change towards cultural relevance for the target demographic.

4.3 VLM-based Metrics

Finally, similar to LLM-based scores for text, we can prompt VLMs to assign a score on a fixed scale to estimate how well the transcreated image matches the *intent* of the user; how far the model has transcreated the source image for *cultural relevance*; and the level of *creativity*, which measures the visual similarity between the source and transcreated images. Prompt formats can leverage chain-of-thought reasoning Liu et al. (2023), and our implementation of the same is discussed in Section 5.3.

5 Experiments

In this section, we conduct a meta-evaluation of the transcreation metrics proposed above.

5.1 Evaluation Dataset

Country	Images	Country	Images	Country	Images	Country	Images
Brazil	501	India	504	Japan	509	Nigeria	506
Portugal	506	Turkey	509	USA	508	Total	3543

Table 1: Dataset size for the 7 target countries.

Khanuja et al. (2024) create a dataset comprising of commonly occurring concepts across 17 semantic categories (like food, beverages, mammals etc.) that are cross-culturally comprehensive, collected across seven countries. The final dataset includes 585 images, and they apply their proposed pipeline with each country as the target, using images from all other countries as the source. Around 500 images are transcreated for each target country (Table 1). Details about this dataset can be found here.[3]

5.2 Systems

Khanuja et al. (2024) propose three systems comprising of state-of-the-art LLMs, retrievers and image-editing models. Briefly, the first system, **`e2e-instruct`**, is a trivial end-to-end image-editing model prompted to *make the image culturally relevant to X*, where X is the target country. The second pipeline, **`cap-edit`**, first captions the image, then makes the caption 'culturally relevant' using a LLM, and finally edits the original image with the culturally-modified caption. The third system, **`cap-retrieve`** uses the culturally-modified caption from `cap-edit` to retrieve a natural image using the modified caption as the query. In their dataset, all outputs are given a rating on a scale of 1-5 by human annotators, across multiple dimensions including visual change, semantic equivalence, cultural relevance, naturalness and offensiveness. They find that the `cap-retrieve` pipeline performs best, across countries. Questions relevant to our meta-evaluation are in Table 2.

5.3 Correlations with Human Ratings

To evaluate the effectiveness of metrics proposed in Section 4, we calculate segment-level correlations of these metrics with the human ratings from Khanuja et al. (2024) for *intent*, *creativity*, and *cultural-relevance*. Similar to MT that defines segment-level correlation on the order of test instances (sentences), we define it at the level of each image in our test set. We calculate segment-level correlation using the default SciPy (v1.11.4) implementation of Kendall's Tau (Kendall, 1938), and average across the entire test set for a given country. Segments with non-parsable model outputs, or those scored identically by all systems (usually a small minority) are skipped. We detail how the automated metrics are implemented below:

Questions from (Khanuja et al., 2024)	Category
Is the generated image from the same semantic category as the original image?	`intent`
Is there any visual change in the generated image compared to the original image?	`creativity`
Does the image seem like it came from your country / is representative of your culture?	`cultural-relevance`

Table 2: Questions from the evaluation set of Khanuja et al. (2024) used for meta-evaluation in this work.

Object-based (CSI-Overlap): For the first step of object identification, we prompt Gemini-1.5-Pro (Reid et al., 2024), since current open VLMs have been shown to perform poorly on open-vocabulary, long-tailed object detection (Gupta et al., 2019). For the second and third steps of obtaining valid replacements and calculating the proportion of CSIs correctly replaced, we again leverage Gemini-1.5-Pro, but any LLM can be used for this purpose.

Embedding-based: We use SigLIP (Zhai et al., 2023), a dual-encoder VLM that improves on CLIP (Radford et al., 2021) by using a sigmoid loss instead of softmax, allowing for stronger alignment between images and text without large-scale resources. To estimate *intent* in terms of semantic equivalence, we get text embeddings of the statement "This image belongs to category `X`", where `X` is the semantic category the image belongs to, like 'food', 'beverage', 'mammal' etc. To calculate *cultural-relevance*, we similarly obtain text embeddings for "This image belongs to country `X`", where `X` denotes the target country for transcreation.

VLM-based: Finally, we prompt the closed VLMs Gemini-1.5-Pro, GPT-4o (OpenAI, 2024), and the open-sourced Molmo 7B-D model (Deitke et al., 2024) to obtain VLM-based scores (prompts listed in Appendix A).

[3] `https://machine-transcreation.github.io/image-transcreation/`

Category	Model	intent	creativity	cult-rel
Object-based	Gemini-1.5-Pro	0.29	-0.16	-0.08
Embedding-based	SigLIP	0.58	**0.87**	0.35
VLM-based	Gemini-1.5-Pro	0.81	0.48	0.52
	GPT-4o	**0.86**	0.17	**0.55**
	Molmo	0.44	0.12	-0.16

Table 3: Average segment-level correlation between human ratings and the three automated metrics proposed. 'cult-rel' stands for the `cultural-relevance` category.

5.4 Results

Table 3 presents the average segment-level correlations across seven countries for the three evaluation criteria. For *intent*, proprietary VLMs such as GPT-4o and Gemini-1.5-Pro show the highest performance, likely due to their advanced multimodal reasoning capabilities. We also note that embedding-based metrics outperform open-source VLMs, thus proving to be more cost-effective in budget-lean scenarios. However, as open-source VLMs improve, these dynamics could shift in favor of VLMs. For *creativity*, embedding-based metrics using the SigLIP model excel, achieving a strong correlation of 0.87, far surpassing both open-source and proprietary VLMs. This suggests that embedding models are better at capturing fine-grained visual nuances than VLMs that primarily focus on image-text alignment during training. For *cultural relevance*, the trends mirror those of *intent*, though with lower overall correlations, highlighting the complexity of assessing cultural relevance and the potential for future research in building metrics that capture this dimension.

5.5 Strengths, Limitations, Recommendations

We are inspired by recent research on meta-evaluation in MT that has shown that no single MT metric is effective at capturing all kinds of translation errors and each has its strengths and weaknesses, making a singular overall score from a 'winning' metric uninformative (Moghe et al., 2024). We argue that this problem is exacerbated in transcreation, where the open-ended and application-dependent nature of the problem might make an overall score ambiguous. Thus, we now analyse the strengths and limitations of each metric, and provide recommendations for which metric models researchers might use in different scenarios.

Object-based metrics shares some limitations with BLEU. It focuses on exact object matches and offers no credit for near-correct adaptations. Additionally, like BLEU's surface-level n-gram overlap, this metric doesn't account for the context or positioning of objects, and hence cannot capture awkward or nonsensical edits. Overall, while it captures *intent* and *cultural-relevance* to a certain degree, it fails to account for the extent of *creativity* in the adaptation. It obtains lowest correlation values across all three axes, likely due to its modular approach facilitating error propagation.

Embedding-based metrics that leverage vision-language encoder representations excel at capturing visual similarity, even better than paid, proprietary VLMs. While they aren't as competent in multimodal reasoning, they offer a cost-efficient alternative to evaluate systems across all three dimensions.

VLM-based metrics are highly effective at capturing *intent* and *cultural relevance* due to their advanced multimodal reasoning abilities. However, even proprietary VLMs exhibit only moderate correlations with human judgments on cultural relevance, underscoring the complexity of this task and leaving headroom for improvement.

Overall, our findings suggest that a hybrid approach—leveraging proprietary VLMs for reasoning-heavy tasks like *intent* and *cultural relevance* and embedding models for *creativity*, constitutes the most comprehensive evaluation strategy.

6 Conclusion

In this paper, we take a first step towards automatically evaluating image transcreation - the task of adapting visual content from a source culture to a target culture. We first review related works in translation studies and identify that the *intent* of the user, *creativity* involved in adapting the source content, and *cultural relevance* for the target audience, are three key factors guiding any transcreation process. Next, we explore how visual content is being transcreated for real-world applications through this lens, and design evaluation metrics that rank systems across these three axes, drawing inspiration from MT metrics. We find that VLM-based metrics excel at multimodal reasoning (*intent* and *cultural relevance*) but visual similarities can be better captured using visual-encoder embeddings (*creativity*). For all metrics, capturing cultural relevance of an image for a target demographic is challenging, leaving headroom for progress. We hope that our study helps ML researchers make progress on this important, real-world task.

7 Limitations

Our work does not come without limitations. First, even though we broaden the scope of image transcreation to several applications and study real-world examples, we do not cover examples from videos and gaming, the latter being one of the largest markets for localization. However, this is a very ambitious scope and we hope to slowly make progress toward the same. Second, capturing both creativity and cultural relevance can be very nuanced. We rely on divergence in visual similarity to capture the former and country or region names to capture the latter, both of which do not comment on finer-grained nuances across systems. However, we do get high correlation with human ratings, especially for the former, making them a good enough substitute for the lack of better metrics. Third, all our prompting for VLM-based metrics is currently done in English, but it would be interesting to see how results might change on choosing languages that are local to a specific region. We leave this exploration to future work.

8 Ethical Statement

In this work, we emphasize that we are not attempting to formalize the creation of cultural deepfakes or undermine the jobs of artists. On the contrary, the goal is to increase accessibility to culture-specific content. We recognize that the demand for human-generated translations and human creativity will always be paramount.

At the same time, we acknowledge that some people prefer to engage directly with the source culture rather than consume localized adaptations. This is particularly true among manga fans, who often take pride in understanding nuanced cultural references that may be lost in translation. In such fan communities, translations and localizations are frequently criticized, as they inevitably result in a distortion or transformation of the original author's intended meaning. While localized adaptations can expand the global audience, we believe there is inherent value in respecting cultures in their unadulterated form and providing opportunities for people to engage directly with source materials. Our aim is to strike a balance between accessibility and preserving the integrity of cultural works.

References

Ibrahim Sa'id Ahmad, Antonios Anastasopoulos, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, William Chen, Qianqian Dong, Marcello Federico, et al. 2024. Findings of the iwslt 2024 evaluation campaign. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 1–11.

Javier Franco Aixelá. 1996. Culture-specific items in translation. *Translation, power, subversion*, 8:52–78.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Claudia Benetello. 2018. When translation is not enough: Transcreation as a convention-defying practice. a practitioner's perspective. *Journal of Specialised Translation*, pages 28–44.

Coca-Cola. 2023. Marketing mind | three nations, one ad. https://www.instagram.com/reel/C9rAGmqBau1/?igsh=d2IwZmVucmcxN3J6. Instagram Video.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*.

Mar Díaz-Millón and María Dolores Olvera-Lobo. 2023. Towards a definition of transcreation: a systematic literature review. *Perspectives*, 31(2):347–364.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2023. Gptscore: Evaluate as you desire. *arXiv preprint arXiv:2302.04166*.

Gabriel Gianordoli and Robert Ito. 2023. How manga was translated for america. *The New York Times*. Accessed: 2024-07-23.

Agrim Gupta, Piotr Dollar, and Ross Girshick. 2019. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364.

Maurice G Kendall. 1938. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

Farisa Khalid. 2012. Video: Meena, india's dora the explorer, helps unicef promote 'child rights'. *Asia Society*.

Simran Khanuja, Sathyanarayanan Ramamoorthy, Yueqi Song, and Graham Neubig. 2024. An image speaks a thousand words, but can everyone listen? on translating images for cultural relevance. *arXiv preprint arXiv:2404.01247*.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondrej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, et al. 2024. Preliminary wmt24 ranking of general mt systems and llms. *arXiv preprint arXiv:2407.19884*.

Tom Kocmi and Christian Federmann. 2023. GEMBA-MQM: Detecting translation quality error spans with GPT-4. In *Proceedings of the Eighth Conference on Machine Translation*, pages 768–775, Singapore. Association for Computational Linguistics.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Chi-kiu Lo. 2019. Yisi-a unified semantic mt quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513.

Marián Martín. 2024. *Promoting and Assessing Creativity for the Training of Transcreators: Some Inspiring Training Resources*, pages 281–307.

Nikita Moghe, Arnisa Fazla, Chantal Amrhein, Tom Kocmi, Mark Steedman, Alexandra Birch, Rico Sennrich, and Liane Guillou. 2024. Machine Translation Meta Evaluation through Translation Accuracy Challenge Sets. *Computational Linguistics*, pages 1–60.

Christiane Nord. 2009. El funcionalismo en la enseñanza de traducción. *Mutatis Mutandis: Revista Latinoamericana de Traducción*, 2:209–243.

Vilde Olaussen. Why Marketing Translation Matters, and How to Get it Right. `https://redokun.com/blog/marketing-translation`. [Online; accessed 18-July-2024].

OpenAI. 2024. Hello gpt-4o. `https://openai.com/index/hello-gpt-4o/`. Accessed: 2024-10-14.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Victoria Purcell-Gates, Gigliana Melzi, Behnosh Najafi, and Marjorie Faulstich Orellana. 2011. Building literacy instruction from children's sociocultural worlds. *Child Development Perspectives*, 5(1):22–27.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Rebecca Ray and Nataly Kelly. 2010. Reaching new markets through transcreation. Common Sense Advisory.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Katharina Reiss, Hans J. Vermeer, Christiane Nord, and Marina Dudenhöfer. 2014. *Towards a General Theory of Translational Action: Skopos Theory Explained*. Routledge, New York.

Gregg Sangillo. 2019. How to get to sesame street… in nigeria. *American University News*.

TranslatePlus. 2023. Transcreation examples that hit the mark. `https://www.translateplus.com/blog/transcreation-examples-that-hit-the-mark/`. [Online; accessed 18-July-2024].

Unicef et al. 2018. Learning through play. *(No Title)*.

Lawrence Venuti. 2017. *The translator's invisibility: A history of translation*. Routledge.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*.

Federico Zanettin et al. 2014. Visual adaptation in translated comics. *Intralinea*, 16:1–34.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Appendix

All the prompts used to calculate the CSI-Overlap metric can be found in Table A below.

Prompts used to calculate CSI Overlap in Section 4.1

Gemini-1.5 Pro (Object Detection)
"You are an expert in detecting objects in an image. Detect unique objects/entities in the image and only list each object ONCE. Give the Wikipedia or Wikidata entity name of the object detected, if possible. The output format is a comma-separated list of entities."

Gemini-1.5-Pro (Getting Replacements)
"Imagine you are an image editor tasked with adapting visuals for a global audience. You are given an image where object detection identified these entities: ', '.join(extracted_objects). Your goal is to identify any entities that represent specific countries, cultures, or regions and are not be universally recognizable or relevant to someone from COUNTRY. For each culture-specific entity you identify, please provide a diverse and exhaustive list of potential replacements that would be more culturally relevant to someone in COUNTRY, and belongs to the CATEGORY category. Please explain your reasoning step by step for identifying each entity. Please format your response as a JSON object where the keys are the culture-specific entities and the values are lists of potential replacements. The reasoning should be a separate key-value pair. The replacement list should be VERY EXHAUSTIVE and such that the replacements are highly relevant to COUNTRY. ONLY output the JSON, DO NOT WRITE ANYTHING ELSE."

GPT4 (Calculate Overlap)
"You are given two lists of objects, List 1 and List 2. Your task is to identify objects that are similar between the two lists, allowing for minor variations in spelling, wording, or punctuation. Do not consider broad categories like 'animal', 'plant', 'flower', 'object', 'food', 'material', or any similar broad terms. Focus only on specific, meaningful matches between the lists, ignoring any broad or generic terms. Please explain your reasoning step by step and provide a final list of matching objects at the end. Provide the output as a JSON object with the following format. DO NOT WRITE ANYTHING ELSE."

Prompts used to for VLM-scores

Gemini-1.5-Pro (Intent)

"Given the input image and its category CATEGORY, determine if the comparison image belongs to the same category as the input. Please explain your reasoning step by step, and provide a final score at the end. The score should be between 1 and 5, where: 1 = Dissimilar Category, 5 = Same Category."
source-image
target-image
"The output should be a JSON object ONLY with the following format: {"reasoning": ..., "score": number}"

Gemini-1.5-Pro (Creativity)

"Given the input image and its category CATEGORY, determine if there are visual changes in the second image as compared to the first. Please explain your reasoning step by step, and provide a final score at the end. The score should be between 1 and 5, where: 1 = No visual change, 5 = High visual changes."
source-image
target-image
"The output should be a JSON object ONLY with the following format: {"reasoning": ..., "score": number}"

Gemini-1.5-Pro (Cultural Relevance)

"You will be given two images. You have to assess how culturally relevant both images are with respect to the culture of COUNTRY. Please explain your reasoning step by step for both images, specifically considering cultural symbols, styles, traditions, or any features that align with the culture of the speaking population of COUNTRY. For each image, the final score should be a number between 1 to 5, where 1 and 5 mean the following: 1 = Not culturally relevant, 5 = Culturally relevant."
source-image
target-image
"The output should be a JSON object ONLY with the following format:{"first_reasoning": ..., "first_score": number, "second_reasoning": ..., "second_score": number}"