MPTc-Bench🎬: Measuring Cross-Market Generative Ability of Vision-Language Models via Movie Poster Transcreation

Youyuan Lin[1] **Yuan Li**[2] **Yahan Yu**[1] **Fei Cheng**[3] **Shinya Nishida**[3] **Chenhui Chu**[3]
[123]Kyoto University
[1]{youyuan, yahan}@nlp.ist.i.kyoto-u.ac.jp
[2]{li.yuan.67n}@st.kyoto-u.ac.jp
[3]{feicheng, shinyanishida, chu}@i.kyoto-u.ac.jp

Abstract

Generative vision-language models (VLMs) can edit and synthesize images, yet their ability to adapt visual assets across markets remains under-evaluated. We study cross-market image transcreation via movie posters, where localization must preserve a movie's identity while matching market-specific design preferences and multilingual typography. We introduce the Movie Poster Transcreation Benchmark (MPTc-Bench), a cross-market benchmark of 582 aligned poster examples spanning 34 target markets, and define two task variants: **Surface** (text-centric localization) and **Deep** (preference-level style adaptation). We propose a two-stage planner-editor pipeline in which a VLM planner specifies executable edits and an image editor renders them. We evaluate in a triplet setup (source, human target-market poster, model output) using information-preservation checks, LLM-as-a-judge ratings for aesthetics and target-market fit, and objective similarity signals. Across multiple planners and editors, experiments reveal substantial gaps between model outputs and human target-market posters, highlighting open challenges for market-aware generation. MPTc-Bench enables controlled, quantitative progress on cross-market image editing beyond understanding-centric benchmarks.[1]

Figure 1: Surface vs. Deep transcreation in MPTc-Bench. Blue regions indicate the source market; Orange regions indicate the target market. **Top:** Surface examples primarily involve translated typography with minor layout adjustments. **Bottom:** Deep examples exhibit significant shifts in composition and artistic style. (Original copyrights retained by owners.)

1 Introduction

In translation studies, *textual transcreation* refers to creative adaptation beyond literal translation to elicit an audience-aligned effect (Carreira, 2024; Kaindl, 2024; Seo, 2018). Recent work adapts this notion to image editing and formulates *image transcreation* as translating an image to better match a target audience while preserving meaning (Khanuja et al., 2024). Meanwhile, generative vision–language models (VLMs) have made text-to-image generation and instruction-based editing increasingly practical (Ramesh et al., 2021, 2022; Saharia et al., 2022; Rombach et al., 2022; Brooks et al., 2023). This progress raises a core evaluation question for real deployments: can a model transcreate a visual asset for a different market while preserving the underlying product identity?

In this paper, we study movie poster transcreation, which is distinct from conventional image transcreation tasks in two key ways. First, movie posters are inherently multimodal: cross-market adaptation often requires transcreating *text* (title, headlines, slogans) via market-aware copywriting, while staying faithful to the movie and the poster's visual hierarchy (Carreira, 2024; Kaindl, 2024; Seo, 2018). Second, movie posters are semantically anchored to a fixed narrative and cast, limiting concept-level visual substitutions; unlike general image transcreation where culturally grounded visual elements can be swapped to increase relevance

[1]Code and dataset: minamotoorin.github.io/mptc-bench.

Findings of the Association for Computational Linguistics: ACL 2026, pages 37897–37913
July 2-7, 2026

(Khanuja et al., 2024), movie poster transcreation often manifests as *implicit market preferences*—latent tendencies in design rhetoric and style (e.g., typography hierarchy, layout density, color treatment, and soft-sell versus hard-sell appeals). Such tendencies are consistent with systematic differences observed in localized advertising appeals between Japanese and American print ads (Okazaki and Mueller, 2008; Okazaki et al., 2010).

These constraints make movie poster transcreation a representative task for *cross-market* generation and editing: the system must map implicit preferences into design choices while staying anchored to the movie. Movie poster data is abundant and naturally paired across markets for the same title, making it suitable for quantitative evaluation: IMDb provides rich movie metadata, and Douban.com and Eiga.com host regional movie poster variants.[2]

We therefore introduce the **Movie Poster Transcreation Benchmark (MPTc-Bench)**, where the extra "c" emphasizes transcreation (vs. translation). MPTc-Bench is a cross-market movie poster dataset of aligned source–target pairs for the same movie, linked with movie metadata for grounded generation and evaluation (§3). Our released lists contain 582 examples spanning 34 target markets, selected from 4,569 candidate cross-market pairs. Pairs range from text-centric localization to broader style adaptation reflecting implicit market preferences. To separate text-dominated localization from preference-level visual adaptation, MPTc-Bench defines **Surface** and **Deep** task variants using perceptual-hash distance between the source movie poster (SRC) and a human target-market movie poster (GT) as a proxy for required change (Zauner, 2010) (§4). Figure 1 illustrates representative Surface and Deep examples, and Figure 2 summarizes the benchmark pipeline.

Using MPTc-Bench, we evaluate current models on cross-market transcreation. Movie posters are complex multimodal artifacts: they combine dense typography with non-trivial visual layout, and they must remain anchored to the movie identity. We therefore adopt a two-stage *planner-editor* framework (§4): a VLM planner explicitly enumerates transcreation edits (including preference-level style changes), while an image editor stage focuses on faithful execution. This decomposition enables stage-wise evaluation (e.g., planned text localization versus rendered text fidelity; §7) and controlled comparisons across planners and editors.

Evaluating movie poster transcreation is challenging: there are no established metrics for whether an output matches target-market preferences. We therefore contribute an evaluation suite tailored to this setting. In a triplet setup (SRC, GT, model output), we combine (i) basic-information checks (title fidelity and genre/year quizzes), (ii) judge-based assessments of aesthetic and adaptation quality from a target-audience perspective, and (iii) reference-based similarity metrics (CLIP embeddings (Radford et al., 2021) and FID (Heusel et al., 2018)) to capture preservation–adaptation trade-offs.

Experiments across multiple planners and editors reveal substantial gaps to human target-market posters, showing that robust cross-market transcreation remains challenging for current systems. In particular, models struggle to reliably render localized text within coherent layouts, which strongly affects perceived target-market fit.

Our contributions are:

- MPTc-Bench: an aligned cross-market movie poster dataset and released evaluation lists (Surface/Deep) covering 582 examples across 34 target markets.
- A planner-editor framework for transcreation that enables controlled comparisons across planners and image editors.
- An evaluation suite that measures basic information preservation, target-market fit, and objective similarity, including an LLM-as-a-judge with market-preference validation.
- Empirical findings showing substantial gaps to human target-market posters and highlighting challenges in market-aware planning and faithful typography/layout execution.

2 Related Work

Transcreation and marketing localization. In translation studies and the language industry, *transcreation* refers to creative adaptation that aims to elicit an audience-aligned effect, often going beyond literal translation in marketing materials (Carreira, 2024; Kaindl, 2024; Seo, 2018). Recent work extends this notion to *image transcreation*, where an image is adapted to improve cultural relevance

[2]IMDb: https://www.imdb.com/; Douban: https://www.douban.com/; Eiga: https://eiga.com/.

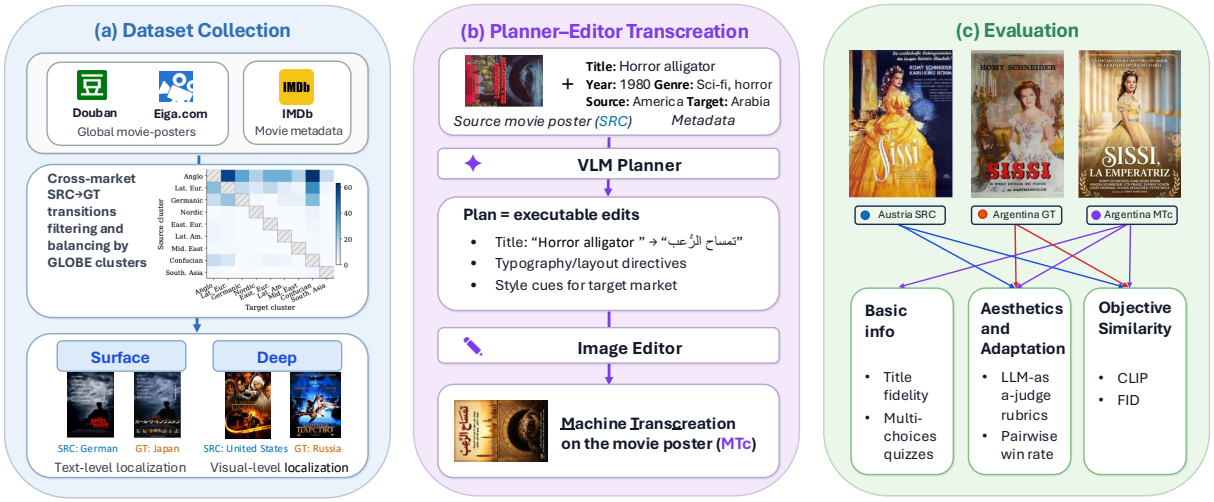

Figure 2: Overview of MPTc-Bench.

for a target market while preserving meaning, and shows that the task remains challenging for current generative systems (Khanuja et al., 2024). HanfuBench studies cross-temporal cultural understanding and image transcreation in a heritage domain, using Hanfu (traditional Chinese garments) as the anchor (Zhou et al., 2025). To scale evaluation beyond expensive human ratings, prior work proposes automatic metrics that jointly consider cultural relevance, semantic equivalence, and visual similarity (Khanuja et al., 2025). Our MPTc-Bench differs by focusing on typography-heavy movie posters with market-specific design preferences and evaluating outputs with GT as reference.

Multilingual and multicultural VLM evaluation. Prior work on multicultural evaluation of VLMs mostly focuses on *understanding*: recognition, QA, and reasoning across languages and domains. General-purpose benchmarks such as MMBench (Liu et al., 2024) and MMMU (Yue et al., 2024) stress broad multimodal capability. Cross-lingual variants extend VQA to multiple languages (e.g., xGQA (Pfeiffer et al., 2022)), while recent efforts probe cultural grounding and multicultural concepts beyond language translation (Liu et al., 2021; Bhatia et al., 2024). These benchmarks provide valuable signals about cultural *understanding*, but they do not directly test whether a system can produce market-appropriate visual artifacts. Our benchmark targets this complementary axis by evaluating cross-market *editing* under explicit semantic constraints and target-market design conventions.

Text-to-image generation and editing. Modern text-to-image models (Ramesh et al., 2021, 2022; Saharia et al., 2022; Rombach et al., 2022) enable high-fidelity synthesis, and instruction-guided editing methods have made controlled visual transformation more accessible (Brooks et al., 2023). Diffusion-based editing has been studied via text-driven editing and conditioning mechanisms (Avrahami et al., 2022; Kawar et al., 2023; Zhang et al., 2023). Parallel to method development, several datasets and benchmarks evaluate instruction-based editing and inpainting (Wang et al., 2023; Zhang et al., 2024). MPTc-Bench differs in that it tests *localized creative adaptation* (e.g., typography and layout conventions) using real cross-market movie poster pairs, making the task closer to practical asset localization than generic instruction following.

3 Dataset Collection

The dataset collection method is summarized in Figure 2 (a).

3.1 Data Sources and Alignment

We construct MPTc-Bench from a merged pool of movie posters and metadata aggregated from Douban, Eiga, and IMDb. Concretely, we combine movie poster images from two large regional movie databases with abundant movie poster variants: Douban and Eiga, and align them to IMDb for canonical identifiers and rich movie metadata. We keep only movies with IMDb alignment and use the IMDb ID as the canonical key.

3.2 Cross-Market Pairing

Aligned pairs. For each movie with posters from at least two markets, we select one representative

Stage	Unit	Count	#Markets
Merged pool	Movies	38,177	34
Cross-market candidate	Pairs	4,569	34
MPTc-Bench (Surface)	Pairs	187	29
MPTc-Bench (Deep)	Pairs	395	33

Table 1: Dataset construction statistics. Movie poster counts reflect a cap of 32 movie posters per movie during market annotation for efficiency.

poster per market using a deterministic rule (preferring locally available images when multiple records exist) and generate cross-market pairs $(\mathrm{SRC}, \mathrm{GT})$. This produces 4,569 candidate cross-market pairs from a merged pool of 38,177 movies (Table 1). To make the direction meaningful, we restrict the source market to be one of the movie's production regions, and treat the remaining markets as targets.

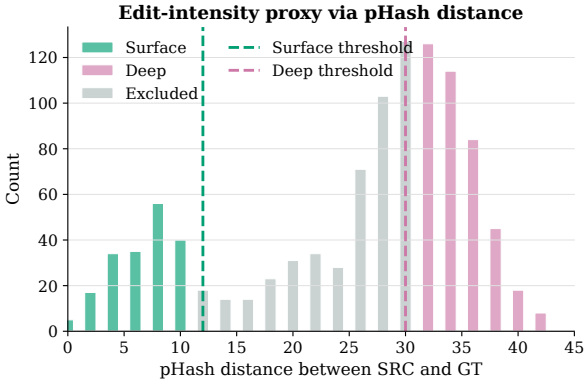

Figure 3: pHash distance distributions. We stratify by thresholds 12 (Surface) and 30 (Deep), discarding intermediate-distance pairs.

Edit intensity proxy. To stratify examples by required adaptation strength, we compute the perceptual-hash (pHash) (Zauner, 2010) distance between SRC and GT. Let $\mathbf{h}(I) = (h_1(I), \ldots, h_{64}(I)) \in \{0,1\}^{64}$ denote the 64-bit pHash of image I; we use the Hamming distance:

$$d_{\mathrm{pHash}}(I_a, I_b) = \sum_{i=1}^{64} [h_i(I_a) \neq h_i(I_b)]\,. \quad (1)$$

where $[\cdot]$ is 1 if the predicate is true and 0 otherwise. Low distances typically indicate near-identical imagery, while high distances correspond to substantial visual redesign. We define **Surface** candidates as pairs with distance < 12 and **Deep** candidates as pairs with distance > 30, discarding the intermediate band to avoid ambiguous cases. Figure 3 shows the resulting separation (see also Figure 9 in Appendix B). We report a threshold-sensitivity check in Appendix B.2.

Balancing and market diversity. To reduce skew toward markets with abundant movie poster coverage, we cap sampling to at most $k{=}4$ examples for each directed (source, target) market pair. We additionally require the source and target markets to fall into different GLOBE cultural clusters (House et al., 2004), encouraging cross-cluster transitions rather than within-cluster variations. See Appendix B (Figures 8, 10) for detailed statistics.

4 Movie Poster Transcreation

4.1 Problem Setting

We study *image transcreation* for global movie marketing: adapting a source-market movie poster to a target market. In translation studies, transcreation is commonly framed as going beyond literal translation to achieve audience-aligned communication in marketing materials (Carreira, 2024; Seo, 2018). In our setting, a system receives a source movie poster SRC, source/target market and movie metadata (e.g., genre, year), and outputs a machine-transcreated target-market movie poster MTc.

The task is constrained by two requirements: (i) Textual information preservation: preserve recognizable identity and key information; (ii) Target-market fit: match target-market conventions such as script, typography hierarchy, layout density, and common genre tropes. We use aligned human movie posters (GT) as references for evaluation.

4.2 Task Variants

We define two task variants that differ in the allowed edit scope:

- **Surface transcreation.** Text-only localization: translate and re-typeset title/tagline/credits in the target language while preserving the original composition and imagery.
- **Deep transcreation.** Holistic localization: allow both textual and visual changes (e.g., background, color, layout, and localized motifs) to better match target-market preferences while keeping the movie recognizable.

4.3 Framework

We adopt a two-stage *planner-editor* machine transcreation framework, as shown in Figure 2 (b). Movie poster transcreation couples cross-market

reasoning (what to adapt given the target market and movie metadata) with challenging visual execution (especially typography and layout). We therefore let a VLM planner externalize decisions as an explicit edit plan, while an image editor focuses on faithful rendering. This decoupling improves controllability and enables stage-wise diagnostics via controlled comparisons across planners and editors. Each transcreation job provides the source movie poster (SRC), source/target markets and languages, and movie metadata (e.g., genre and year) for grounding. The planner first generates a structured edit plan; for **Surface** it focuses on translating and re-typesetting textual elements with explicit translations (e.g., provide target-language title/tagline strings and specify their placement and hierarchy), while for **Deep** it additionally proposes preference-level visual edits (e.g., adjust color palette, layout density, and background treatment) while keeping the movie recognizable. We then compose a unified editing instruction by combining a shared prompt prefix, the planner plan, and a task-specific template, and pass it together with SRC to the image editor to produce the machine-transcreated movie poster MTc. Prompts and decoding settings are provided in Appendix A.1.

5 Evaluation

Evaluating movie poster transcreation is challenging because quality is multi-faceted and the "right" target-market adaptation is not uniquely defined. First, an output must preserve movie identity while localizing dense typography and layout conventions; failures in multilingual text rendering can dominate perceived usability. Second, Deep transcreation may require stylistic departures, making reference-based similarity an incomplete proxy for target-market fit.

To address these challenges, we adopt a triplet protocol (Figure 2 (c)): for each transcreation job, we evaluate a model output MTc against the source movie poster SRC and an aligned human target-market movie poster GT, forming (SRC, GT, MTc). The reference GT is used only for evaluation and is never provided during generation. We then combine complementary metrics that directly target the challenges above: (i) **basic information preservation** to verify identity and quantify typography execution, and (ii) **subjective and objective signals** by pairing judge-based aesthetics/adaptation scores with CLIP/FID similarity measures to describe target-market fit and preservation–adaptation trade-offs, especially for Deep transcreation where similarity alone is insufficient. Table 2 summarizes all metrics.

Metric	Definition
Basic Info.	
Title chrF	chrF between titles extracted from GT and MTc movie posters.
Plan chrF	chrF between the planner-proposed title translation and the title extracted from GT.
Genre hit	Multiple-choice genre classification accuracy on MTc.
Year hit	Multiple-choice release-decade classification accuracy on MTc.
Aes. and Adap.	
Aesthetic score	Judge score (1–5) for visual naturalness/polish of MTc.
Adaptation score	Judge score (1–5) for target-market fit of MTc.
Win rate	Judge preference rate of MTc versus GT (pairwise).
Obj. Similarity	
CLIP sim.	CLIP cosine similarity between GT and MTc (Radford et al., 2021).
FID_{GT}	FID between GT and MTc distributions (Heusel et al., 2018).
FID_{SRC}	FID between SRC and MTc distributions (Heusel et al., 2018).

Table 2: Evaluation metrics in MPTc-Bench. Basic and aesthetics metrics are averaged over examples; FID is computed at the dataset level.

5.1 Basic Information Preservation

We estimate rendered title fidelity by extracting titles from GT and MTc movie posters via a vision-based OCR instruction and computing chrF similarity (Popović, 2015). To isolate planning quality from rendering, we also compute planned-title chrF by comparing the planner-proposed title translation against the GT title extracted from GT. We additionally evaluate coarse semantic correctness with two multiple-choice quizzes on MTc (genre and release decade), and report hit rates.

5.2 Aesthetics and Adaptation

We use a judge LLM (GPT-5.1, deterministic decoding) (OpenAI, 2025) instructed to role-play as a target-market audience member and marketing consultant. The judge assigns 1–5 scores for (i) visual aesthetics and (ii) target-market adaptation, and we additionally compute a pairwise preference rate between GT and MTc with deterministic order randomization to mitigate position bias. Prompt

templates and key decoding parameters are provided in Appendix A.

Judge validation. We validate that the judge is sensitive to target-market preferences using a market-preference control task. Given a source movie poster SRC and a role-play target market, we pair two aligned human posters from different markets: the target-market poster GT_{tgt} (for the role-play market) and a distractor-market poster GT_{other}. We hide market labels, randomize the A/B order, and report the target-market preference rate (fraction choosing GT_{tgt}); the prompt is in Appendix A.2. Across 1,178 pairs, the judge selects GT_{tgt} in 96.8% of cases (Figure 4), confirming strong market sensitivity.

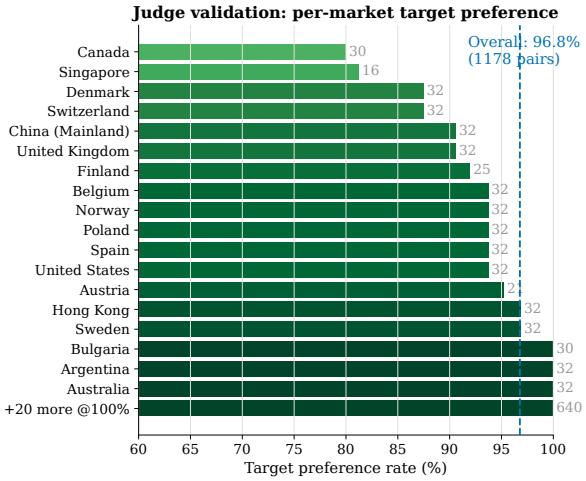

Figure 4: Validating judge sensitivity. Bars show how often the LLM judge favors target-market movie posters. The blue dashed line marks the overall rate; bar-end numbers give samples per market.

5.3 Objective Similarity

We compute CLIP similarity between GT and MTc (Radford et al., 2021). We also report Fréchet Inception Distance (FID) between (GT, MTc) and between (SRC, MTc) (Heusel et al., 2018). These metrics are auxiliary: they are insensitive to text readability and do not define a universal objective, especially for Deep transcreation where stronger adaptation can legitimately depart from SRC.

6 Experimental Setup

Planners. We evaluate two planners: GPT-4.1 (OpenAI, 2025) and Qwen3-VL-8B (Qwen3) (Qwen Team, 2025b) (deployed locally on two NVIDIA RTX 6000 Ada GPUs). This selection balances a hosted planner with a locally deployable open-source alternative.

Image editors. We compare six image editing systems: Gemini 2.5-flash-Image (Gemini 2.5) and Gemini 3-Pro-Image (Gemini 3) (Google, 2025), FLUX.1-schnell (Flux.1 Edit) (Black Forest Labs et al., 2025), SDXL-Turbo (SDXL Edit) (Podell et al., 2023), Qwen-Image-Edit (Qwen Edit) (Qwen Team, 2025a), and Step1X-Edit-v1p1-diffusers (Step1X) (Liu et al., 2025). These editors span both hosted multimodal APIs and local models, covering distinct paradigms from mask-based inpainting to instruction-following editing. Prompt templates are detailed in Appendix A.1.

Judge model. For LLM-as-a-judge, we use GPT-5.1 (OpenAI, 2025). The hyperparameters are shown in Table 6 of Appendix A.3. We compute CLIP using clip-ViT-B-32.[3] We implement FID using `torch_fidelity` Python package.[4]

7 Results

We organize the results into four research questions (RQs) about cross-market generative ability in MPTc-Bench.

7.1 RQ1: Which Editors Best Support Cross-market Transcreation?

On **Surface** transcreation, the intended optimum is to preserve the SRC imagery and layout while producing readable, well-placed localized typography, so high title fidelity and strong preservation signals are desirable. While Genre/Year accuracy remains high across editors (Table 4), editors differ sharply in title rendering and judged quality (Table 3), with Gemini 3 consistently preferred over GT and diffusion-based editors and other models rarely winning.

On **Deep** transcreation, the optimum shifts toward higher judged adaptation with movie identity preserved, so similarity to SRC becomes a descriptive trade-off rather than a target. Deep results show larger gaps in title fidelity and judged adaptation (Tables 4 and 3); objective similarity can look reasonable even when outputs are not preferred (Table 5), underscoring that CLIP/FID do not capture typography and target-market fit. Similar to Surface, Gemini 3 attains the best judged adaptation and preference, while other editors lag behind. Figure 5 (top-left) illustrates these editor differences.

[3] https://huggingface.co/sentence-transformers/clip-ViT-B-32

[4] https://github.com/toshas/torch-fidelity

Figure 5: Case studies in MPTc-Bench (original copyrights retained by owners). **Top-left** illustrates editor differences in judged aesthetics and preference. **Top-right** shows planner sensitivity on Deep: with the editor fixed, GPT-4.1 planning yields a more effective adaptation than Qwen3. **Bottom-left** highlights a caveat: the plan introduces a Japanese-style room background without movie-specific evidence. **Bottom-right** shows a Surface failure where the plan is plausible but the editor fails to render the localized title.

Model	**Deep** *(GT ref.: Aes.=4.19, Adap.=4.06)*						**Surface** *(GT ref.: Aes.=4.49, Adap.=4.17)*					
	Planner: GPT-4.1			Planner: Qwen3			Planner: GPT-4.1			Planner: Qwen3		
	Aes.↑	Adap.↑	Win%↑	Aes.↑	Adap.↑	Win%↑	Aes.↑	Adap.↑	Win%↑	Aes.↑	Adap.↑	Win%↑
Flux Edit	3.18	2.12	1.6	3.13	2.07	1.3	3.03	2.01	0.0	3.01	2.01	1.1
Gemini 2.5	4.28	3.56	41.4	4.11	3.05	25.6	3.95	3.28	11.6	4.16	3.17	9.5
Gemini 3	**4.51**	**4.07**	**67.9**	**4.30**	**3.34**	**40.1**	**4.77**	**4.14**	**54.1**	**4.70**	**3.95**	**43.5**
Qwen Edit	3.76	2.85	10.2	3.68	2.57	4.3	3.39	2.72	8.2	3.42	2.71	5.5
SDXL Edit	3.20	2.14	1.4	3.18	2.14	1.4	3.12	2.07	0.5	3.12	2.12	0.5
Step1X	1.83	1.59	0.8	1.78	1.59	0.5	2.07	1.67	1.1	2.10	1.68	0.0

Table 3: Aesthetics and adaptation scores on MPTc-Bench. Best results are shown in **bold**. Aes. and Adap. range from 1 to 5; Win% is the preference rate of MTc over GT. GT reference scores are shown in each task header.

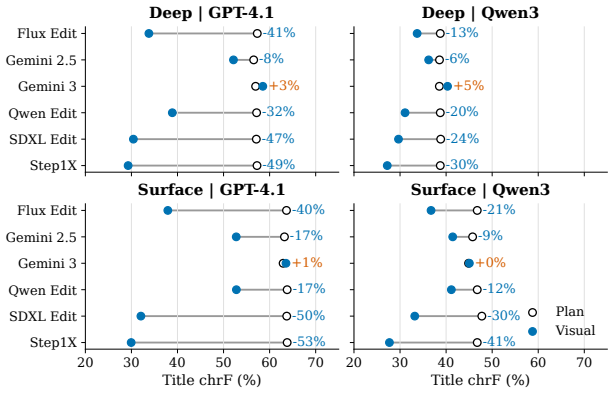

Figure 6: Planner-to-editor title fidelity on MPTc-Bench, visualized as planned-title chrF (Plan) versus rendered-title chrF (Visual). We annotate loss as $(1 - \text{Visual}/\text{Plan}) \times 100\%$.

7.2 RQ2: How Sensitive Is Transcreation To Planning Quality?

Planning quality matters, especially for Deep transcreation. GPT-4.1 produces better plans than Qwen3 in planned-title chrF, with the largest difference on Deep (Table 4). This gain transfers to higher adaptation and preference mainly on Deep (Table 3), while objective similarity changes less (Table 5). Figure 5 (top-right) shows one typical case. We also observe a limitation shared by both planners: over-reliance on target-market visual stereotypes. In such cases, an MTc output can outperform GT in preference while being less grounded in the specific movie content (Figure 5 (bottom-left)).

Task	Model	Planner: GPT-4.1				Planner: Qwen3			
		Genre↑	Year↑	Title↑	Plan↑	Genre↑	Year↑	Title↑	Plan↑
Deep	Flux Edit	78.6	95.1	33.8	**57.3**	79.2	95.4	33.7	38.7
	Gemini 2.5	78.5	94.8	52.2	56.5	79.9	92.8	36.2	38.6
	Gemini 3	**80.2**	93.5	**58.5**	57.0	**80.1**	89.4	**40.3**	38.5
	Qwen Edit	77.4	93.0	38.9	57.2	75.7	91.1	31.1	38.7
	SDXL Edit	72.1	88.8	30.4	57.1	72.8	85.8	29.7	**38.8**
	Step1X	75.5	**96.5**	29.3	**57.3**	77.9	**96.8**	27.2	38.7
Surface	Flux Edit	84.7	98.9	37.9	63.7	85.8	98.4	36.7	46.8
	Gemini 2.5	87.7	99.4	52.8	63.2	86.7	98.7	41.4	45.8
	Gemini 3	87.4	98.7	**63.6**	62.9	87.0	98.8	**45.0**	44.9
	Qwen Edit	**88.5**	**100.0**	52.8	**63.8**	**88.0**	**99.5**	41.1	46.8
	SDXL Edit	82.5	95.6	32.1	63.7	84.2	95.6	33.2	**47.8**
	Step1X	84.7	96.7	29.9	**63.8**	84.2	96.7	27.7	46.8

Table 4: Basic information preservation on MPTc-Bench. Best results are shown in **bold**. Genre/Year are quiz hit rates (%). Title is chrF between titles extracted from GT and MTc (%). Plan is the planned-title chrF (%). Genre/Year remain high across editors, but rendered title fidelity varies widely, especially on Deep.

Model	**Deep** / GPT-4.1			**Deep** / Qwen3			**Surface** / GPT-4.1			**Surface** / Qwen3		
	CLIP↑	FID_{GT}↓	FID_{SRC}	CLIP↑	FID_{GT}↓	FID_{SRC}	CLIP↑	FID_{GT}↓	FID_{SRC}↓	CLIP↑	FID_{GT}↓	FID_{SRC}↓
Flux Edit	61.6	**105.8**	**90.8**	61.5	**105.1**	90.7	74.0	108.8	104.9	74.3	107.3	103.5
Gemini 2.5	**66.9**	106.3	108.6	**65.1**	108.0	108.3	86.5	79.0	59.8	85.0	**78.5**	51.5
Gemini 3	65.5	109.9	123.5	63.2	113.3	123.6	**87.5**	**76.5**	**53.6**	**86.1**	79.7	57.7
Qwen Edit	64.7	107.2	113.0	62.5	112.8	118.1	84.1	79.9	54.8	83.0	**78.5**	58.7
SDXL Edit	62.8	106.0	96.1	62.8	105.9	94.5	76.9	115.0	114.5	77.4	113.2	112.3
Step1X	59.4	147.4	137.7	59.6	153.2	149.4	76.1	131.4	117.5	74.6	136.2	124.8

Table 5: Objective similarity metrics on MPTc-Bench. Best results are shown in **bold**. CLIP is the mean cosine similarity between GT and MTc (scaled by 100); FID_{GT} and FID_{SRC} are dataset-level distribution distances to GT and SRC, respectively. These are descriptive similarity signals rather than universal "higher/lower is better" objectives: Surface transcreation typically favors similarity to both GT and SRC, while Deep transcreation can require larger departures from SRC.

7.3 RQ3: Where Do Systems Fail? Text Rendering Dominates the Error Type

Table 4 reveals a consistent bottleneck across tasks: while plans often propose reasonable title translations, many editors fail to render them faithfully, yielding large gaps between planned-title and rendered-title chrF. Figure 5 (bottom-right) shows a representative Surface failure where the plan is plausible but the editor fails to render the localized title. This execution gap directly affects human judgment. Editors with lower title fidelity in Table 4 tend to also receive lower adaptation and preference in Table 3, suggesting that readable, well-placed localized text is a prerequisite for perceived cross-market fit. Meanwhile in Table 5, CLIP/FID primarily reflect global appearance and still remain relatively strong even when text is garbled. To isolate the planning-to-rendering gap, Figure 6 summarizes the plan-to-visual loss; Gemini 3 incurs near-zero loss, whereas diffusion-based editors and Step1X often lose 30–50% of title fidelity.

7.4 RQ4: Cross-Cluster Performance Analysis

This analysis focuses on direction effects in cross-cluster transcreation. We define Western as {Anglo, Latin Europe, Germanic Europe, Nordic Europe, Eastern Europe} and Asian as {Confucian Asia, Southern Asia}. We use Gemini 3.1 and Qwen Edit as two representative editors from different (Western and Eastern) model ecosystems.

Performance varies by transfer direction (Western→Asian vs. Asian→Western). If one direction were universally easier, both representative models would favor the same direction. Instead, Deep+Surface combined summaries show opposite tendencies: Gemini 3.1 is stronger on Asian→Western than Western→Asian (0.630 vs. 0.585), while Qwen Edit is stronger on Western→Asian than Asian→Western (0.066 vs. 0.036). This indicates that direction effects are real and model-dependent.

Another observation is that scalar scores and

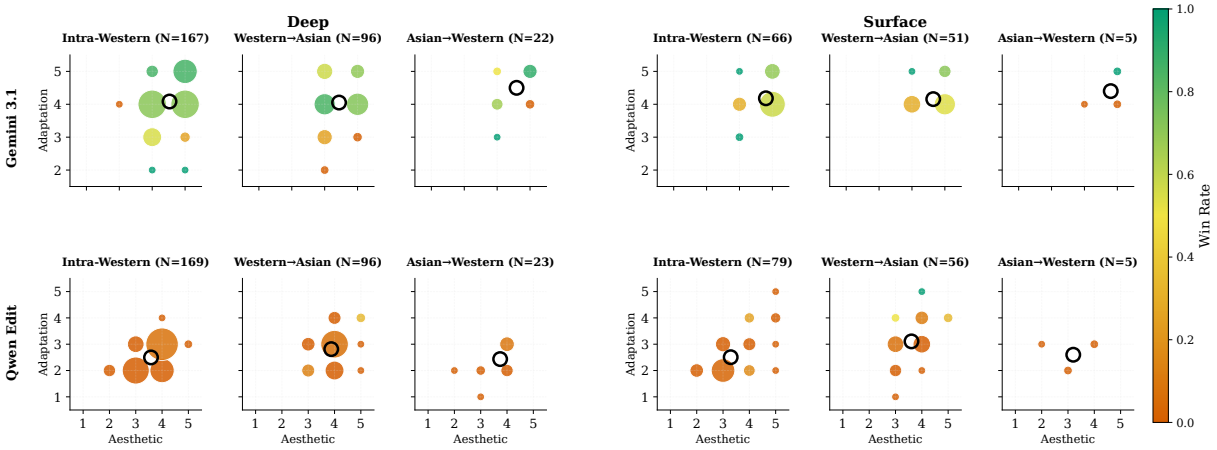

Figure 7: Cross-cluster performance across two representative editors: Gemini 3.1 + GPT-4.1 planner (top row) and Qwen Edit + Qwen3 planner (bottom row), for Deep (left block) and Surface (right block) transcreation. Region directions are Intra-Western, Western→Asian, and Asian→Western. Bubble size indicates sample count at each score coordinate; color indicates pairwise win rate (red: lower, green: higher); mean score location is marked by a black ring.

preference are not equivalent, even at the same score coordinate. For example, at (4,4), Gemini points are associated with much higher win rates than Qwen Edit (about 0.63 vs. 0.09 in pooled summaries).

8 Conclusion

We present MPTc-Bench, a benchmark for measuring cross-market generative ability of VLMs via movie poster transcreation. MPTc-Bench provides aligned SRC–GT movie poster pairs and two task variants: **Surface** (typography-focused localization) and **Deep** (preference-level style adaptation). We further propose a planner-editor framework and evaluation metrics combining basic information checks, assessments of aesthetics and target-market adaptation, and objective similarity signals.

Across state-of-the-art image editors, we find substantial gaps in cross-market transcreation quality. Performance is primarily limited by multilingual typography and layout execution: even when planners produce reasonable title translations, several editors fail to render legible text. Planner choice matters mainly for Deep transcreation, where higher-level intent and style decisions are more consequential. Current VLMs may overuse stereotyped market cues. We hope MPTc-Bench supports more rigorous evaluation and targeted progress on market-aware generation, complementing understanding-only benchmarks.

Limitations

Coverage. MPTc-Bench does not cover all countries and regions, with limited representation for Africa. Expanding to broader geographic and language coverage is an important next step.

Limited human evaluation. We include a small-scale human evaluation (Appendix B.3), but recruiting annotators with appropriate target-market familiarity across many markets remains difficult. Our main comparisons therefore still rely on automated metrics and LLM judging, which may miss nuanced local preferences.

Potential data leakage. Our movie posters are collected from public websites and may have been included in web-scale pretraining corpora for modern VLMs and diffusion models (e.g., LAION) (Schuhmann et al., 2021, 2022). Such leakage can bias benchmark results, reflecting a broader challenge of data contamination in evaluating frontier models (Sainz et al., 2023).

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number JP23K28144 and JST BOOST Program Japan Grant Number JPMJBY25D2.

Ethical Considerations

MPTc-Bench is derived from publicly available movie posters collected from multiple sources. We do not claim ownership of any movie poster image;

copyright and usage rights remain with the original rightsholders. The benchmark is intended for research and educational use. When redistributing, hosting, or publishing examples, users must ensure compliance with the terms of the original sources and applicable copyright regulations. To facilitate reproducibility under these constraints, our release emphasizes metadata, annotations, and evaluation metrics rather than re-hosting movie poster images.

AI Assistants

We used Writefull[5] and GPT-5.2 (Thinking) for grammar checking and polishing. The code implementation was assisted by GitHub Copilot.

References

Omri Avrahami, Dani Lischinski, and Ohad Fried. 2022. Blended diffusion for text-driven editing of natural images. *Preprint*, arXiv:2111.14818.

Mehar Bhatia, Sahithya Ravi, Aditya Chinchure, Eunjeong Hwang, and Vered Shwartz. 2024. From local concepts to universals: Evaluating the multicultural understanding of vision-language models. *Preprint*, arXiv:2407.00263.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, and 2 others. 2025. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space. *Preprint*, arXiv:2506.15742.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. *Preprint*, arXiv:2211.09800.

Oliver Carreira. 2024. *Transcreation: Beyond translation and advertising*. De Gruyter.

Yuning Du, Chenxia Li, Ruoyu Guo, Weidong Yin, Xinan Cui, Yuehua Zheng, Baoguang Shi, Guoping Hu, Erlei Zhang, and 1 others. 2021. Pp-ocr: A practical ultra lightweight ocr system. *Preprint*, arXiv:2109.03144.

Google. 2025. Gemini api: Model reference. Accessed: 2025-12-30.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2018. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Preprint*, arXiv:1706.08500.

[5] https://writefull.com/

Robert J. House, Paul J. Hanges, Mansour Javidan, Peter W. Dorfman, and Vipin Gupta, editors. 2004. *Culture, Leadership, and Organizations: The GLOBE Study of 62 Societies*. SAGE Publications.

Klaus Kaindl. 2024. *Transcreation as a Means of Distinction: The Use of Transcreation in the Translation Industry*, pages 15–44. Springer International Publishing.

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023. Imagic: Text-based real image editing with diffusion models. *Preprint*, arXiv:2210.09276.

Simran Khanuja, Vivek Iyer, Xiaoyu He, and Graham Neubig. 2025. Towards automatic evaluation for image transcreation. 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)*, pages 7034–7047, Albuquerque, New Mexico. Association for Computational Linguistics.

Simran Khanuja, Sathyanarayanan Ramamoorthy, Yueqi Song, and Graham Neubig. 2024. An image speaks a thousand words, but can everyone listen? on image transcreation for cultural relevance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10258–10279, Miami, Florida, USA. Association for Computational Linguistics.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. 2021. Visually grounded reasoning across languages and cultures. *Preprint*, arXiv:2109.13238.

Shiyu Liu, Yucheng Han, Peng Xing, Fukun Yin, Rui Wang, Wei Cheng, Jiaqi Liao, Yingming Wang, Honghao Fu, Chunrui Han, Guopeng Li, Yuang Peng, Quan Sun, Jingwei Wu, Yan Cai, Zheng Ge, Ranchen Ming, Lei Xia, Xianfang Zeng, and 5 others. 2025. Step1x-edit: A practical framework for general image editing. *Preprint*, arXiv:2504.17761.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024. Mmbench: Is your multi-modal model an all-around player? *Preprint*, arXiv:2307.06281.

Shintaro Okazaki and Barbara Mueller. 2008. Evolution in the usage of localized appeals in Japanese and American print advertising. *International Journal of Advertising*, 27(5):771–798.

Shintaro Okazaki, Barbara Mueller, and Charles R. Taylor. 2010. Measuring soft-sell versus hard-sell advertising appeals. *Journal of Advertising*, 39(2):5–20.

OpenAI. 2025. Openai api model documentation. Accessed: 2025-12-30.

Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaa El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, and 6 others. 2024. Dinov2: Learning robust visual features without supervision. *Preprint*, arXiv:2304.07193.

PaddlePaddle. 2025. Pp-ocrv5 server detection model card. Accessed: 2026-04-14.

Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin O. Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. 2022. xgqa: Cross-lingual visual question answering. *Preprint*, arXiv:2109.06082.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *Preprint*, arXiv:2307.01952.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*. Association for Computational Linguistics.

Qwen Team. 2025a. Qwen-image technical report. *Preprint*, arXiv:2508.02324.

Qwen Team. 2025b. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *Preprint*, arXiv:2204.06125.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. *Preprint*, arXiv:2102.12092.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. *Preprint*, arXiv:2112.10752.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Preprint*, arXiv:2205.11487.

Oscar Sainz, Jon Ander Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. Nlp evaluation in trouble: On the need to measure llm data contamination for each benchmark. *Preprint*, arXiv:2310.18018.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Preprint*, arXiv:2210.08402.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *Preprint*, arXiv:2111.02114.

Jeong Mok Seo. 2018. Translation strategies and transcreation in advertising slogans. *The International Journal of Art and Culture Technology*, 2(2):45–50.

Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J. Fleet, Radu Soricut, Jason Baldridge, Mohammad Norouzi, Peter Anderson, and William Chan. 2023. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. *Preprint*, arXiv:2212.06909.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *Preprint*, arXiv:2311.16502.

Christoph Zauner. 2010. Implementation and benchmarking of perceptual image hash functions. Master's thesis, Upper Austria University of Applied Sciences, Hagenberg.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. 2024. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Preprint*, arXiv:2306.10012.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. *Preprint*, arXiv:2302.05543.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595.

Li Zhou, Lutong Yu, Dongchu Xie, Shaohuan Cheng, Wenyan Li, and Haizhou Li. 2025. Hanfu-bench: A multimodal benchmark on cross-temporal cultural understanding and transcreation. *Preprint*, arXiv:2506.01565.

A Prompts and Hyperparameters

This appendix summarizes the prompt designs and API-facing parameters used in MPTc-Bench.

A.1 Transcreation Prompts

We present the complete prompts. Variables in {braces} are replaced with task-specific values. The prompts below are copied verbatim from our released prompt templates.

Planner system prompt

```
You are a movie-poster localization director.
     You will receive a source poster image
    along with metadata.

Analyze the poster visually and output a
    concise numbered list of executable edits.
     Each edit must include actual translated
     text or specific visual changes based on
     what you see in the image. Be brief—no
    explanations, no headers, no filler.
```

Surface planning prompt

```
[Attached: source poster image]

Movie: {movie_id}
Source: {source_market} / {source_language}
Target: {target_market} / {target_language}
Genre: {gt_genre}

You are a key-art localizer. Output a SHORT
    numbered list of text-only edits. Each
    line must include the EXACT translated
    text.

Format Example:
1. Title: "Original" → "Translated {
    target_language} title"
2. Tagline: "Original" → "Translated {
    target_language} tagline"
3. Credits: key names in {target_language}

Rules:
- Include actual translated text, not "
    translate to..."
- No visual changes
- Keep it brief
```

Surface editing prompt

```
Localize the {source_market} poster of "{
    movie_id}" for {target_market} ({
    target_language}).

Execute these edits:
{planner_plan}

Apply text changes exactly as specified.
    Preserve all imagery, colors, and
    composition.
```

Deep planning prompt

```
Movie: {movie_id}
Source: {source_market} / {source_language}
Target: {target_market} / {target_language}
Genre: {gt_genre}

You are a key-art localizer and poster art
    director for the target market.

Analyze the attached poster and design a
    holistic localization plan that may
    significantly redesign the poster (layout
    , composition, background, color palette,
     character poses, typography, and
    cultural motifs) while keeping the film
    and main characters clearly recognizable.

Output a numbered list (maximum 10 items) of
    concrete text AND visual edits that,
    together, are sufficient for an image-
    editing model to repaint the entire
    poster for the {target_market} audience.

Rules:
- Item 1 MUST ALWAYS be the title
    localization, formatted exactly as:
  1. Title: "Original" → "Translated {
    target_language} title"
- For every text change (title, tagline,
    credits, release date, etc.), include the
     final localized text in quotes.
- Visual edits must specify what to change
    and how (colors, positions, camera angle,
     lighting, style, background elements,
    props, clothing, cultural references, etc
    .).
- Strongly localize: adapt background,
    setting, color scheme, typography, and
    decorative elements to match {
    target_market} poster conventions and {
    target_language} culture.
- Keep key characters, logos, and franchise
    branding recognizable.

Output format:
- ONLY output the numbered edit list, nothing
     else.
```

Deep editing prompt

```
Localize the {source_market} poster of "{
    movie_id}" for {target_market} ({
    target_language}).

Execute these edits:
{planner_plan}
```

```
Apply all text and visual changes exactly as
    specified. Keep the film's identity
    recognizable.
```

A.2 Evaluation Prompts

Title extraction

```
You are a movie poster OCR assistant. Extract
     the main movie title from the movie
    poster image.

Rules:
- Preserve the original language and spelling
     exactly as shown
- If multiple titles appear, return the most
    prominent one
- If no title is visible, return empty string

Respond in JSON:
{"title": "extracted title here"}
```

Visual quiz template

```
You are a movie poster analyst. Answer the
    multiple-choice question about the
    attached movie poster.

Choose from options A, B, C, or D provided in
     the user message.
- For genre: infer from visual style, imagery,
     and text cues
- For year/decade: look for year text, visual
     style, or production hints

Respond in JSON:
{"answer": "A"}
```

LLM-as-a-judge (direct scoring)

```
You are a movie marketing consultant AND a
    local audience member from {target_market
    }, evaluating a localized movie poster (
    TGT) for the {genre} movie originally
    marketed in {source_market}.

You will be shown two images: SRC (source-
    market), TGT (target-market).
Judge from the perspective of a {
    target_market} viewer: base adaptation on
     {target_market} norms (language/script &
     readability, typography hierarchy,
    required rating/legal marks, billing
    block & cast order, release-date format,
    layout density/visual hierarchy, imagery/
    tropes appropriate for {genre}), while
    preserving SRC's movie identity and key
    info.

Scoring rubrics (integers 1-5):

aesthetic_score - visual quality of TGT:
5: Professional, balanced layout, clear
    typography, compelling.
4: Minor flaws; overall polished.
3: Noticeable issues; still readable/
    functional.
2: Significant problems; hard to read.
1: Unusable; severely broken.

adaptation_score - target-market adaptation
    quality of TGT (for {target_market}):
5: Fully aligned with local conventions;
    identity & key info preserved.
4: Mostly aligned; minor mismatches.
3: Partially aligned; some awkward/missing
    local elements; core message intact.
2: Minimal alignment; feels foreign;
    important info wrong/lost.
1: No meaningful adaptation or market-
    inappropriate/misleading.

Reply ONLY in valid JSON, without any extra
    text:

{{
  "short_comment_on_aesthetics": "<one-
    sentence explanation>",
  "aesthetic_score": <int 1-5>,
  "short_comment_on_adaptation": "<one-
    sentence explanation>",
  "adaptation_score": <int 1-5>
}}
```

LLM-as-a-judge (pairwise preference)

```
You are a movie marketing consultant AND a
    typical local audience member in {
    target_market}. The movie was originally
    marketed in {source_market}; the genre is
     {genre}.

Compare two movie posters for the SAME movie
    and SAME target market:
- Image A
- Image B
(Order is random; origins unknown. Do NOT
    infer which is human or machine.)

Judge strictly from a {target_market} viewer'
    s perspective, using only the images.
    Consider ordinary moviegoers' overall
    preference based on:
(1) visual appeal/polish; (2) clarity &
    typography/readability; (3) target-market
     fit with {target_market} conventions (
    correct language/script, hierarchy,
    required rating/legal marks, billing
    block \& cast order, release-date format,
     layout density, {genre}-appropriate
    imagery/tropes), while preserving the
    movie's identity and key info.
If quality seems equal, prefer the one with
    fewer artifacts and clearer text.

Reply ONLY in valid JSON:

{{
  "reason": "<brief reason>",
  "preferred": "A" or "B"
}}
```

Market preference (judge validation)

```
You are a typical movie-goer living in {
    target_market}. You are familiar with
    local movie poster conventions, visual
    styles, and local preferences in your
    region.

You are shown two movie posters (A and B)
    that are localized versions of the same
    movie for DIFFERENT target markets.
- Image A
- Image B
(Order is random; target markets unknown. Do
    NOT try to guess which market each movie
    poster targets.)

As a {target_market} audience member, which
    movie poster appeals to you more? Judge
    strictly from your perspective as a {
    target_market} viewer, considering:
(1) visual appeal and aesthetic preferences
    common in {target_market};
(2) typography, layout, and design
    conventions familiar to {target_market}
    viewers;
(3) market-specific elements and imagery that
     resonate with {target_market} audience;
(4) language/script appropriateness for {
    target_market}.

Important: Focus on which movie poster would
    be more effective and appealing in {
    target_market}, not which is "objectively
     better".

Reply ONLY in valid JSON:

{{"reason": "<brief reason>", "preferred": "A
    " or "B"}}
```

A.3 API Parameters

Planner and judge models. We summarize decoding parameters for planner and judge LLMs in Table 6.

Component	Model	Temp.	Max tokens
Planner	GPT-4.1	0.2	1,024
Planner	Qwen3-8B	0.2	1,024
Judge LLM			
– evaluation	GPT-5.1	0.0	1,024
– validation	GPT-5.1	0.0	512

Table 6: Key decoding parameters for planner and judge models.

B Supplementary Results

B.1 Dataset Structure and Coverage

Figure 8 shows the source–target market GLOBE cultural cluster transition matrix in MPTc-Bench.

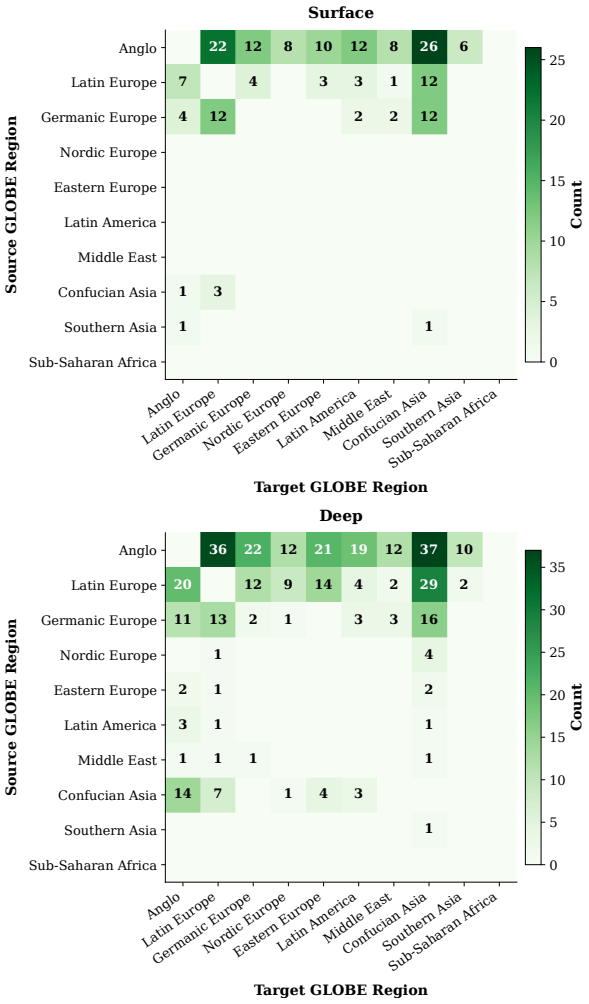

Figure 8: Source–target market GLOBE cultural cluster transition matrix in MPTc-Bench. Darker cells indicate more examples.

Figure 9 further breaks down the dataset by GLOBE transition direction and plots the distribution of perceptual-hash (pHash) distances between paired movie posters. Across directions, the distribution is consistently bimodal: one peak corresponds to visually near-duplicate pairs dominated by text localization, while the other corresponds to pairs requiring substantial style or layout changes. This pattern supports the need for our two task variants (Surface and Deep), which target these two qualitatively different regimes.

Figure 10 reports the number of paired examples per target market. The distribution is uneven, reflecting both data availability and our balancing procedure.

B.2 pHash-Threshold Sensitivity

To test whether the Surface/Deep split is stable under reasonable threshold variation, we analyze all

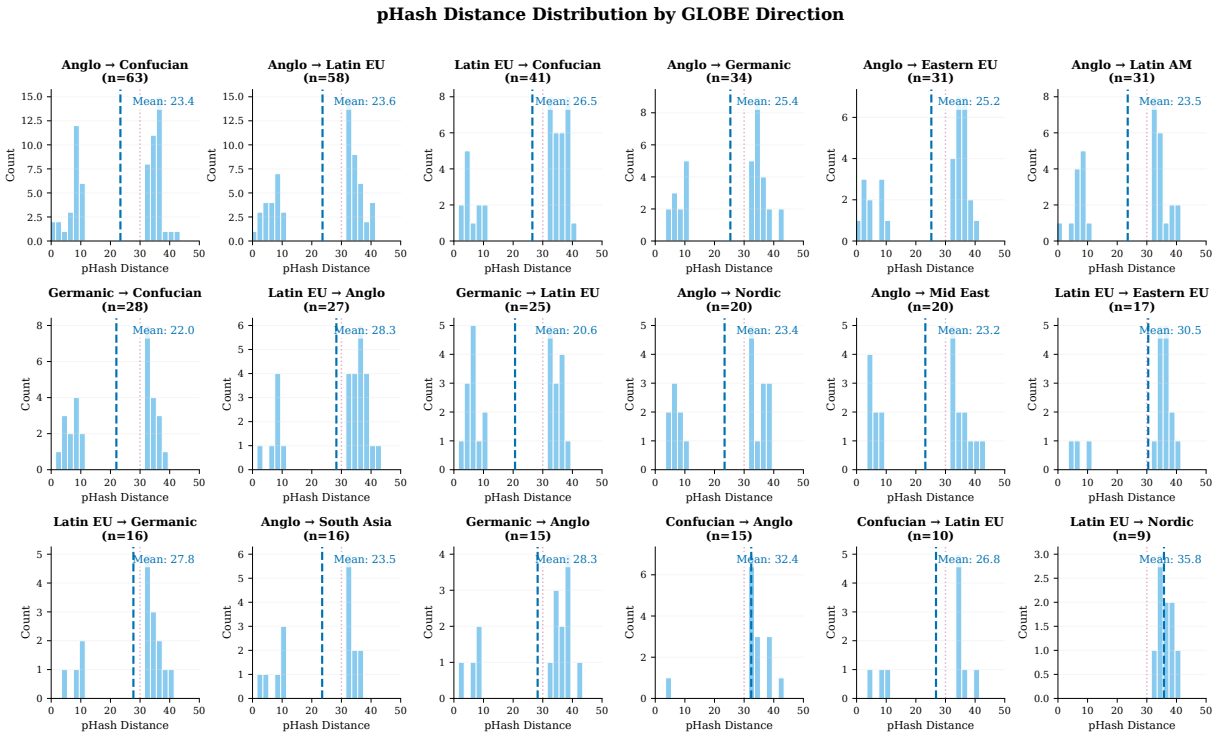

Figure 9: Distribution of pHash distances between paired movie posters in MPTc-Bench, grouped by source–target GLOBE transition direction. The bimodal structure suggests two regimes of cross-market adaptation: near-duplicate, text-dominated localization (Surface) and larger preference-level redesigns (Deep).

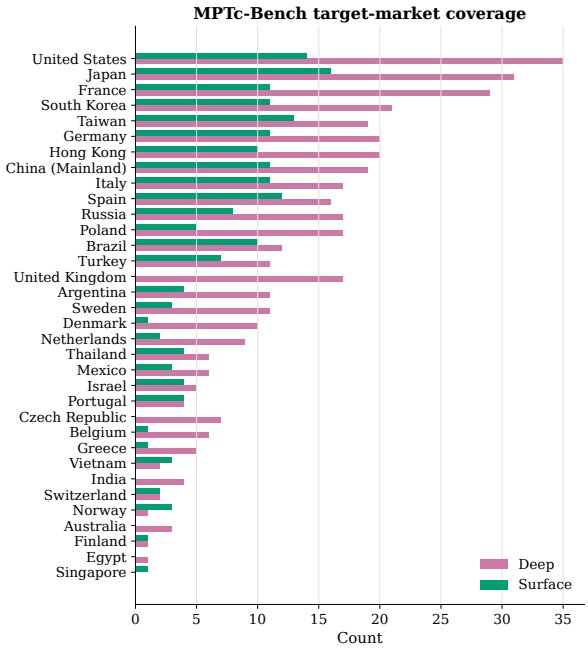

Figure 10: Target-market coverage in MPTc-Bench: number of paired movie poster examples for each target market (country/region).

candidate SRC-GT pairs. The primary split is Surface ($pHash \leq 12$) and Deep ($pHash \geq 30$), with the middle band removed, following perceptual-hash distance (Zauner, 2010). As complementary distance measures, we compute LPIPS (Zhang et al., 2018) (`alex`) and DINOv2 (Oquab et al., 2024).

pHash band	n	LPIPS mean±std	DINO mean±std
≤ 12	205	0.27±0.11	0.09±0.07
13–29	318	0.60±0.12	0.42±0.25
≥ 30	522	0.69±0.07	0.57±0.20

Table 7: Agreement between pHash bands and learned distance measures on all 1,045 candidate pairs.

Figure 11 and Table 7 show monotonic increases from low-pHash to high-pHash bands for both LPIPS and DINO. We additionally run a local threshold sweep around the operational split: $t_s \in \{10, 12, 14\}$, $t_d \in \{28, 30, 32\}$. Across all 9 settings, Deep-vs-Surface separability remains stable: LPIPS AUC 0.9942 to 0.9952, DINO AUC 0.9815 to 0.9908, and pHash AUC 1.0000 throughout. These results indicate that the thresholded pHash split is stable around $(12, 30)$.

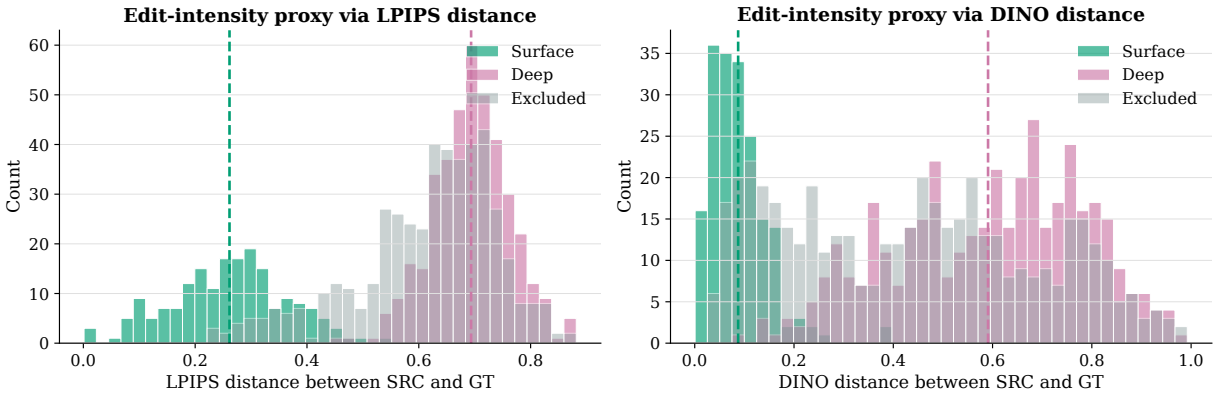

Figure 11: Distance distributions generated by our LPIPS/DINO analysis pipeline on the full candidate pool. Left: LPIPS. Right: DINO. Surface/Deep bands follow the paper split ($pHash \leq 12$, $pHash \geq 30$).

B.3 Human Evaluation

We conduct a human evaluation on GPT-4.1 + Gemini 3 outputs. We focus on China (Mainland) and use three annotators, all with at least a master's degree. Table 8 shows the same overall pattern as the main results: Deep is harder than Surface.

Split	Pairs	Anno.	Aesthetic	Adaptation	Win rate
Surface	10	30	4.33	4.47	60.0%
Deep	16	48	3.56	3.65	31.2%

Table 8: Human evaluation on China (Mainland). 'Anno.' denotes the number of human annotations.

Inter-annotator agreement is moderate for adaptation (Quadratic Weighted Kappa: 0.46 on Surface, 0.51 on Deep) and lower for aesthetics (0.24 / 0.29). Human–LLM score correlation is weak on Surface but higher on Deep ($\rho_{\text{adapt}} = 0.00$ vs. 0.32), which is consistent with score saturation on Surface in this small sample. We annotate five error types: `garbled_text` (broken or unreadable text), `wrong_script` (script does not match the target market), `layout_collapse` (overlap, clipping, or hierarchy loss in text/layout), `non_evidenced_elements` (added content not supported by SRC), and `identity_drift` (core movie identity cues are altered). Error rates differ by task: Surface is mainly text/script failures (`wrong_script` 13.3%, `garbled_text` 6.7%), while Deep is dominated by unsupported additions and identity/layout failures (`non_evidenced_elements` 37.5%, `identity_drift` 12.5%, `layout_collapse` 10.4%).

B.4 Temporal Contamination

To estimate whether model performance is associated with movie release year as a temporal contamination check, we evaluate six run settings: Gemini 2.5/3 and Qwen-Edit, each on Surface and Deep with GPT-4.1 plans. For each run, we compute Spearman ρ for $\rho(\text{year}, \text{title-}chrF)$ at sample level, and $\rho(\text{year}, FID_{GT}^{\text{by-year}})$ on per-year FID values with at least 5 samples per year. For FID_{GT}, lower is better. This temporal analysis follows the contamination motivation discussed in recent benchmark studies (Sainz et al., 2023).

Table 9 summarizes all six settings. $\rho(\text{year}, \text{title-}chrF)$ is small (median -0.070), while $\rho(\text{year}, FID_{GT}^{\text{by-year}})$ is consistently negative (median -0.530).

Run setting	n	years used	ρ(year, title-$chrF$)	ρ(year, $FID_{GT}^{\text{by-year}}$)
Gemini 2.5 / Surface	172	11	-0.125	-0.491
Gemini 2.5 / Deep	362	22	-0.009	-0.525
Gemini 3 / Surface	159	11	-0.097	-0.536
Gemini 3 / Deep	368	23	-0.019	-0.653
Qwen-Edit / Surface	183	11	-0.108	-0.536
Qwen-Edit / Deep	371	23	-0.044	-0.506
Median across runs	–	–	-0.070	-0.530

Table 9: Temporal contamination analysis across six system settings.

Overall, newer-year subsets tend to achieve lower FID_{GT}, indicating a non-negligible temporal effect in objective similarity, while the title-text correlation with year remains weak.

B.5 Layout-Aware OCR Results

To complement title-string fidelity with a layout-aware text-overlap signal, we evaluate Surface triplets using PP-OCRv5. We extract all text poly-

gons, scale target images to source size, and compute union-mask IoU (Du et al., 2021; PaddlePaddle, 2025). The three evaluated systems are Gemini 2.5 + GPT-4.1, Gemini 3 + GPT-4.1, and Qwen-Edit + GPT-4.1. We report $IoU(\text{SRC}, \text{MTc})$, with $IoU(\text{SRC}, \text{GT})$ as reference.

Table 10 reports the per-system and overall values, and shows that all systems have higher SRC-MTc overlap than SRC-GT overlap; the overall delta is $+0.221$, with zero empty-union cases. Additionally, PP-OCRv5 title extraction on source posters is weak under stylized poster typography (top-confidence-line $chrF = 0.1341$, exact match $= 5.41\%$), which is why main-text title fidelity relies on VLM-based extraction. Taken together, current systems preserve source text layout strongly in Surface transcreation.

System	n	SRC-MTc IoU	SRC-GT IoU (ref)	Δ (MTc$-$GT)
Gemini 2.5 + GPT-4.1	172	0.630	0.430	+0.200
Gemini 3 + GPT-4.1	159	0.667	0.431	+0.236
Qwen-Edit + GPT-4.1	183	0.660	0.431	+0.229
Overall	514	0.652	0.431	+0.221

Table 10: Layout-aware OCR evaluation on Surface triplets using PP-OCRv5. IoU is computed from union masks of all detected text polygons after image-size normalization. Δ compares generated posters against human references under the same source anchor (all runs have zero empty-union rate).