Pun2Pun: Benchmarking LLMs on Textual-Visual Chinese-English Pun Translation via Pragmatics Model and Linguistic Reasoning

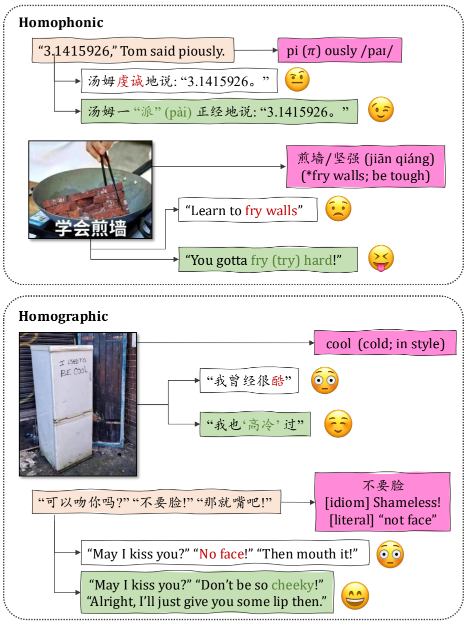
Yiran Rex Ma[1] Shan Huang[2] Yuting Xu[1] Ziyu Zhou[1] Yuanxi Wei[1*]
School of Humanities[1], School of Computer Science[2]
Beijing University of Posts and Telecommunications
{mayiran,weiyuanxi}@bupt.edu.cn

Abstract

Puns, as a unique form of linguistic creativity, present significant challenges in cross-lingual translation, particularly between linguistically distant languages like Chinese and English, where it's often considered a "mission impossible". We introduce Pun2Pun, a novel benchmark for quantitatively evaluating pun translation between Chinese and English while preserving both linguistic mechanisms and humorous effects. We propose the adaptation of Constant-Variable Optimization (CVO) Model for translation strategy and concomitant Overlap (Ovl) metric for translation quality assessment. Our approach provides a robust quantitative evaluation framework to assess models' complex linguistic and cultural reasoning capabilities in pun translation. Through extensive experiments on both textual and visual puns, we demonstrate that our translation strategy model significantly improves performance, particularly for better-performing models. Our findings reveal exciting potentials and current limitations of LLMs in preserving sophisticated humor across linguistic and cultural boundaries.[1]

1 Introduction

Puns, meaning plays on words exploiting dual meanings or similar sounds (Crystal, 2006; Abbott, 2002), represent unique manifestations of linguistic creativity. As shown in Figure 1, puns manifest as homophonic or homographic wordplay, whose translation has long been considered a "mission impossible" (Marina Ilari, 2021; Jakobson, 1959) between linguistically distant languages. This challenge stems from puns' reliance on language- and culture-specific features often absent in target languages (Delabastita, 2016; Cardford, 1975).



"3.1415926," Tom said piously.
pi (π) ously /paɪ/
汤姆虔诚地说: "3.1415926。"
汤姆一"派" (pài) 正经地说: "3.1415926。"

煎墙/坚强 (jiān qiáng)
(*fry walls; be tough)
"Learn to fry walls"
"You gotta fry (try) hard!"

cool (cold; in style)
"我曾经很酷"
"我也'高冷'过"
"可以吻你吗?" "不要脸!" "那就嘴吧!"
不要脸
[idiom] Shameless!
[literal] "not face"
"May I kiss you?" "No face!" "Then mouth it!"
"May I kiss you?" "Don't be so cheeky!"
"Alright, I'll just give you some lip then."

Figure 1: Categories of Puns in Textual and Visual Settings and Comparison of Literal Translation and Pun2Pun Translation.

Traditional approaches resort to suboptimal compromises (Delabastita, 2004), while computational methods, despite progress in detection (Yu et al., 2018; Arroubat, 2022) and generation (He et al., 2019), remain inadequate for translation (Dhanani et al., 2023). Current research mainly addresses closely related language pairs (Ermakova et al., 2022b, 2023b), leaving distant pairs like Chinese-English unexplored (Chen et al., 2023).

Recent advances in Large Language Models (LLMs) and Reasoning Language Models (RLMs) offer promise through sophisticated reasoning capabilities (Kojima et al., 2023; Wei et al., 2023; Besta et al., 2025). While LLMs show strong performance in computational humor (Hessel et al., 2023; Zhong et al., 2024), and existing benchmarks like MMLU (Hendrycks et al., 2020) and

[*] Corresponding author.

[1] Pun2Pun dataset, inference and evaluation scripts are available at https://github.com/rexera/Pun2Pun.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 331–354
July 28-29, 2025 ©2025 Association for Computational Linguistics

GSM8K (Cobbe et al., 2021) test general reasoning, language-specific reasoning remains untapped. Challenges persist in preserving wordplay effects (Weller and Seppi, 2020) and evaluation (Ermakova et al., 2023a).

We introduce Pun2Pun, a novel benchmark for cross-lingual pun translation between Chinese and English, with progressive sub-tasks from classification to translation. We propose the adaptation of Constant-Variable Optimization (CVO) Model (Zhao and An, 2020) for translation strategy and concomitant Overlap (Ovl) metric (Zhao, 2012) for evaluation. Through extensive experiments, we demonstrate improved translation quality while revealing current limitations in preserving humor across linguistic boundaries.

2 Related Work

2.1 Puns in Translation Studies

Puns set against general translation studies, Communicative Translation Theory (Newmark, 1988) prioritizes target-reader reception over literal fidelity, while Functional Equivalence (Nida and Taber, 1964) further underscores contextual reconfiguration to preserve rhetorical effects. As for puns' transferability, Delabastita (2004, 1993) established a foundational taxonomy of eight strategies, including PUN → PUN recreation, PUN → NON-PUN with dual meanings, PUN → RHETORICAL DEVICE, and PUN → ZERO with compensatory notes. Zhang (2000) advocate for pragmatic flexibility, proposing phonetic compensation in Chinese. Recent studies integrate cognitive-pragmatic models (Feng, 2019) to address the interplay of form, humor, and cultural semiotics in constrained contexts.

2.2 Computational Approaches to Puns

Early computational approaches evolved from rule-based systems (Mihalcea and Strapparava, 2005) to neural methods, with notable advances in detection (Arroubat, 2022), generation (Yu et al., 2018), and adversarial networks for controlled generation (Luo et al., 2019). For translation specifically, computer-assisted tools like PunCAT (Kolb and Miller, 2022) and CLEF JOKER workshop corpora (Ermakova et al., 2022a) advanced development, though primarily for closely related language pairs like English, Spanish, and French. Recent LLM-based approaches (Hessel et al., 2023; Zhong et al., 2024) show promise but face unique challenges in preserving wordplay effects (Weller and Seppi, 2020) and reliable evaluation (Albin and Paul, 2022).

2.3 Pun Translation and Complex Reasoning

Pun translation represents a complex reasoning chain: structural decomposition, cross-lingual feature mapping, and constrained creative generation. Recent RLMs (OpenAI, 2024a; DeepSeek-AI, 2025a; Qwen-Team, 2024b,a) leverage search heuristics (Monte Carlo Tree Search, beam search) and structured reasoning for such tasks. While existing benchmarks focus on general knowledge (MMLU (Hendrycks et al., 2020), IFEval (Zhou et al., 2023), GPQA (Rein et al., 2023)), mathematics (like MATH (Hendrycks et al., 2021)), and coding (SWE-Bench Verified (OpenAI, 2024b), LiveCodeBench (Jain et al., 2024)), language-specific complex reasoning remains underexplored.

3 Pun2Pun

3.1 Task Definition

Formulation Let $s = (w_1, w_2, ..., w_n)$ be a pun sentence with punning word w_{pun}. For homographic puns, define M_w as the meaning set of word w_i such that $M_i \rightarrow \{m_1, m_2, ..., m_n\}$, where $|M_i| \geq 2$. A homographic pun exploits dual meanings ($m_a, m_b \in M_{\text{pun}}$) through either polysemy (related meanings) or homonymy (unrelated meanings)[2]. For homophonic puns, let pronunciation ϕ_i correspond to word set $\Phi_i \rightarrow \{w_1, w_2, ..., w_n\}$, where $|\Phi_i| \geq 2$. A homophonic pun leverages phonetic identity/similarity ($w_a, w_b \in \Phi_{\text{pun}}$) to create wordplay. A pun can thus be defined as $P(p_1, p_2)$, where it's composed of two "elements" that shared homographic or homophonic relation. [3]

While we acknowledge that this binary classification may appear simplified compared to more granular linguistic taxonomies that distinguish polysemy, morphological play, cultural allusions, and other subtypes (Attardo, 2017), our approach is pragmatically motivated by the characteristics of available datasets and computational tractability. The source datasets we utilized (Liu, 2018; Chen

[2] We do not distinguish between polysemy and homonymy in this work due to their etymological obscurity.

[3] Note that (1) in practice puns can surely be both homophonic and homographic, while we approach them in isolation in this work; (2) this formulation is still *fuzzy* and subject to change due to complexity and richness of human language, of which we are always in awe.

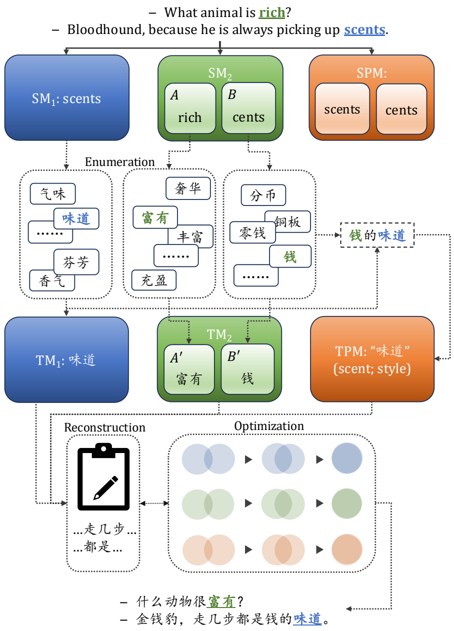


Figure 3: Constant-Variable Optimization (CVO) Model for Pun2Pun Translation. In CVO, Source Meanings (SM) are identified before *enumeration* for target meanings (TM), followed by target language *reconstruction* as well as Overlap *optimization* of three SM-TM pairs through TM word choice alterations, as indicated by three step-wise, overlapping circle pairs.

et al., 2024; Simpson et al., 2019) primarily employ this fundamental distinction, and our focus on cross-lingual translation between Chinese and English—languages with markedly different phonetic and semantic structures (detailed discussion in Section 4.3.3)—makes this binary framework particularly relevant for understanding mechanism transfer patterns.

Strategy Here, we introduce an adapted version of Constant-Variable Optimization (Zhao and An (2020), CVO, Figure 3) as the core approach for pun recreation in Pun2Pun. The CVO framework addresses the challenge of Pun2Pun translation by systematically decomposing the source pun into three essential components and then reconstructing them in the target language.

Decomposition Phase: A source pun is first analyzed into three source meaning (SM) constants: (1) SM_1 represents the core punning word w_{pun} with its dual elements p_1, p_2 that create the wordplay; (2) $\text{SM}_2 = (A, B)$ captures the contextual framework, where A serves as the semantic trigger that sets up the pun's potential and B is the support word that completes one interpretation; (3) $\text{SPM} = (p_1, p_2)$ represents the overall pragmatic effect—the humor mechanism that emerges from the interplay of dual meanings.

Translation Process: The translation achieves cross-lingual transfer by mapping these source components onto corresponding target meaning (TM) variables: TM_1, $\text{TM}_2 = (A', B')$, and $\text{TPM} = (p'_1, p'_2)$. This mapping follows a three-stage process: (1) *Enumeration*—identifying potential target language equivalents for each source component; (2) *Reconstruction*—combining target components to form a coherent pun while adapting to target-language constraints; (3) *Optimization*—refining word choices to maximize semantic and pragmatic overlap between source and target versions, measured by our Overlap metric (detailed in Section 3.3).

Sub-Tasks Building upon this, we designed a progression of tasks for both textual and visual puns, with input sentence s or caption-embedded image v, hereafter both as "puns"$\psi = P(p_1, p_2)$. **Classification** for tagging a pun as either homophonic or homographic: $t \leftarrow \pi(\psi)$; **Locating** the punning elements in the sentence: $w_{\text{pun}} \leftarrow \pi(\psi, t)$; **Decomposition** for extracting two elements of the pun and finish the mechanism: $p_1, p_2 \leftarrow \pi(\psi, t, w_{\text{pun}})$; for visual puns, **Appreci-**

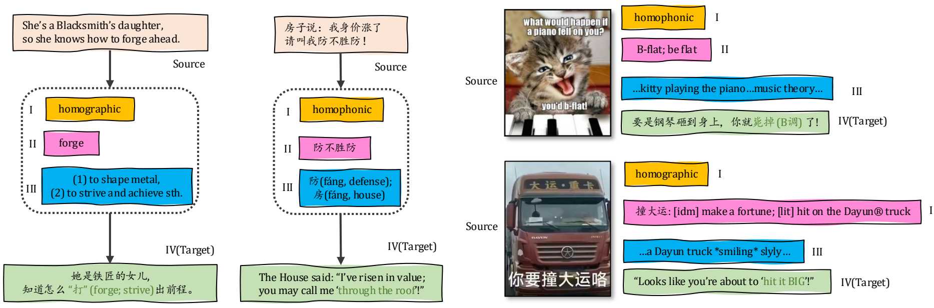

She's a Blacksmith's daughter,
so she knows how to forge ahead.
Source
I
homographic
II
forge
III
(1) to shape metal,
(2) to strive and achieve sth.
IV(Target)
她是铁匠的女儿，
知道怎么"打" (forge; strive) 出前程。
房子说：我身价涨了
请叫我防不胜防！
Source
I
homophonic
II
防不胜防
III
防(fáng, defense);
房(fáng, house)
IV(Target)
The House said: "I've risen in value;
you may call me 'through the roof'!"
Source
what would happen if
a piano fell on you?
you'd b-flat!
homophonic
I
B-flat; be flat
II
...kitty playing the piano ...music theory...
III
要是钢琴砸到身上，你就毙掉 (B调) 了！
IV(Target)
Source
大运·重卡
你要撞大运咯
homographic
I
撞大运: [idm] make a fortune; [lit] hit on the Dayun® truck
II
...a Dayun truck *smiling* slyly...
III
"Looks like you're about to 'hit it BIG'!"
IV(Target)

Figure 2: Progression of Four Sub-Tasks in Pun2Pun: *Classification*(I), *Locating/Decomposition*(II), *Decomposition/Appreciation*(III), and *Translation*(IV) for Textual/Visual Puns.

ation of the interplay of caption and image: $\alpha \leftarrow \pi(\psi, t, w_{\text{pun}}, p_1, p_2)$; finally, **Translation** for creating $\psi' = P(p_1', p_2')$ in target language such that both mechanism and pragmatic effect retain. Interchange from homophonic puns to homographic ones is allowed and vice versa.

We assign four tasks each for textual (I. *Classification*, II. *Locating*, III. *Decomposition*, IV. *Translation*) and visual settings (I. *Classification*, II. *Decomposition*, III. *Appreciation*, IV. *Translation*), as shown in Figure 2.

3.2 Dataset Construction

Sources For textual data, we collected Chinese and English homophonic and homographic puns from multiple sources. Chinese puns were sourced from Liu (2018) and Chen et al. (2024), and English ones were from Simpson et al. (2019), with original statistics in Table 1. For visual data, since no relevant datasets exist, we manually curated a diverse collection of examples from both Chinese and English public social media sources, consisting of images paired with pun-based captions embedded in them.

Quality Assurance and Annotation We implemented a rigorous three-stage annotation process for textual puns, assisted by a `helper` model[4]:

- **Pun Verification:** `Helper` performed initial classification of puns as homophonic, homographic, or non-pun. With pre-labeled data in comparison, all outputs underwent thorough manual review when contradicting with predefined labels and leading to manual inspection of pun validity. Invalid and/or inappropriate examples were either modified to meet our criteria or removed.
- **Mechanism Verification:** `Helper` decomposed each pun's mechanism according to our formulation. We reviewed these outcomes, correcting any misanalysis and ensuring mechanism clarity. Examples lacking clear pun mechanisms after review were either strengthened or removed.
- **Finalization:** Three authors independently reviewed and curated each example following unified annotation guidelines for all Pun2Pun sub-tasks. Disagreements were resolved through team discussion, with challenging cases referred to external translation experts.

For visual puns, three authors manually collected, reviewed, and annotated the entire dataset based on unified standard and annotation guidelines. The final Pun2Pun dataset (statistics in Table 1) comprises 5.5k textual examples across English and Chinese, plus 1k caption-embedded images, all with high-quality, human-reviewed annotations for sub-tasks.

Category	Source	Modality	Phonic	Graphic
Chinese	Liu (2018)	Textual	947	528
	Chen et al. (2024)	Textual	524	528
English	Simpson et al. (2019)	Textual	1268	1610
Pun2Pun	Chinese	Textual	1154	1490
	English	Textual	1197	1661
	Chinese	Visual	426	74
	English	Visual	155	349

Table 1: Statistics of source datasets and our curated Pun2Pun textual dataset

3.3 Evaluation Methodology

Accuracy (Acc) Used for Task I to measure model performance in identifying homophonic and homographic puns.

Agent-Accuracy (AAcc) Applied to Task II and III. Uses a `judge` model[5] to verify consistency between model predictions and human annotations, scoring on a $[0, 10]$ scale.

Cosine Similarity (Cos) Measures semantic alignment in *translation* with an embedding model. Serves not as a determinant metrics but as a measurement for translation creativity.[6]

Hit Binary metric for *translation*, using a `judge` model for evaluating whether the translated sentence successfully contains a pun that is consistent with 1) our specified formulation; 2) target language mechanisms.

[4] `gpt-4o-mini` with vanilla settings and task-agnostic instructions, prompt is in Appendix A. During annotation, we have already found that `gpt-4o-mini` had its shortcomings such as mis-labeling and comprehension failures, particularly for Chinese data.

[5] `gpt-4o-mini`, the same for Hit and Ovl, prompts are in Appendix A. Note that the inherent inadequacy of LLM-as-a-Judge makes this evaluation consistent only within categories rather than comparable across all.

[6] We assume that LLMs would not generate irrelevant content. Since Cos represents superficial semantics, lower similarity with original pun represents better creativity for deviating from surface semantic concepts. In practice, we utilized `text-embedding-v3` from Qwen Team: `https://www.alibabacloud.com/help/en/model-studio/user-guide/embedding`.

Model	Strategy	English						Chinese					
		Hit↑		Ovl↑		Cos↓		Hit↑		Ovl↑		Cos↓	
gpt-4o	Vanilla	15.46	22.64	30.96	37.98	+8.82	+11.13	5.72	4.90	40.17	40.35	+4.76	+7.41
	1-Shot	23.39	26.63	32.76	38.29	+7.53	+10.72	10.92	7.58	35.66	43.56	+2.70	+6.12
	CVO	23.66	24.55	34.99	38.38	+7.53	+10.73	5.64	4.83	35.78	44.83	+4.45	+7.18
o1-mini	Vanilla	16.22	21.91	44.14	47.01	+9.57	+12.03	7.63	5.57	43.08	46.37	+5.28	+7.86
	1-Shot	15.64	22.70	41.91	46.08	+8.79	+11.58	7.26	6.51	46.68	46.05	+4.14	+7.07
	CVO	9.54	14.34	42.99	44.91	+9.45	+12.16	6.24	4.36	42.59	47.56	+4.74	+6.66
qwen-vl-max	Vanilla	3.84	5.96	39.86	44.23	+10.70	+13.18	2.17	2.55	35.35	41.58	+6.81	+8.79
	1-Shot	6.35	8.01	39.36	45.19	+9.83	+12.58	1.74	2.55	47.06	41.18	+5.91	+8.34
	CVO	3.93	7.16	42.69	43.94	+10.50	+13.15	1.81	1.80	36.94	50.10	+6.44	+8.86
qwq-32b-preview	Vanilla	9.52	14.58	41.82	46.79	+6.24	+9.15	4.95	3.63	42.68	49.58	+1.36	+2.99
	1-Shot	7.89	11.65	41.76	46.20	-0.65	+2.75	5.66	5.04	46.68	46.05	-0.11	+1.11
	CVO	14.67	21.86	38.98	46.56	+4.02	+6.90	5.82	4.99	39.36	47.56	+0.56	+2.11
deepseek-v3	Vanilla	10.94	15.41	**63.20**	47.55	+9.84	+12.11	3.56	3.49	39.36	49.44	+7.40	+9.02
	1-Shot	18.88	26.73	44.86	47.48	+9.45	+11.47	5.82	3.83	42.33	42.59	+4.45	+6.26
	CVO	**43.16**	**47.02**	59.43	**62.85**	**-0.93**	**-0.30**	4.26	3.56	40.59	38.80	+5.27	+6.94
deepseek-r1	Vanilla	40.13	24.82	62.30	59.83	+1.39	+4.32	23.89	22.21	**62.37**	**67.13**	+2.14	+4.03
	1-Shot	39.00	39.95	45.96	48.57	+6.12	+8.53	8.59	6.77	46.16	49.19	**-5.15**	**-3.32**
	CVO	34.84	41.47	50.34	49.15	+3.25	+9.30	**26.31**	**24.73**	60.76	65.58	-0.47	+1.26
claude-3.5-sonnet	Vanilla	30.91	33.84	46.60	52.82	+4.27	+7.34	14.73	13.16	47.79	52.82	+1.42	+3.87
	1-Shot	31.24	32.75	40.33	49.46	+5.17	+8.17	15.51	15.58	45.13	49.46	+1.12	+4.44
	CVO	30.16	31.07	44.66	48.58	+6.04	+9.14	16.12	11.42	43.74	48.58	+1.49	+5.10

Table 2: *Translation* Results on Pun2Pun Textual. All metrics are in homophonic(%) + homographic(%) order, with Cos being relative to 70.

Overlap (Ovl) This is concomitant with CVO model, as it is derived from *optimization* stage. For and only for those instances that hit, judge quantifies translation quality through weighted scoring: Ovl $= w_1\langle \mathrm{SM}_1, \mathrm{TM}_1\rangle + w_2\langle \mathrm{SM}_2, \mathrm{TM}_2\rangle + w_3\langle \mathrm{SPM}, \mathrm{TPM}\rangle$, where $w_1 = 0.25$, $w_2 = 0.25$, $w_3 = 0.50$ weight structure preservation, contextual reconstruction, and pragmatic retention respectively. Each component scored $[0, 100]$.

4 Experiments

4.1 Baselines

Models For textual puns, we evaluated various LLMs and RLMs in Pun2Pun, including gpt-4o, o1-mini(OpenAI, 2024a,c), deepseek-v3, deepseek-r1(DeepSeek-AI, 2025b,a), qwen-vl-max(Bai et al., 2023), qwq-32b-preview(Qwen-Team, 2024b), and claude-3.5-sonnet(Anthropic, 2024). As for visual puns, we evaluated gpt-4o, o3-mini(OpenAI, 2025), qwen-vl-max, qvq-72b-preview(Qwen-Team, 2024a), and claude-3.5-sonnet. All hyperparameters remained default.

Strategies 1) *Vanilla* followed a standard I/O with zero-shot Chain-of-Thought prompting ("Let's think step by step" , Wei et al. (2023)); 2) *1-Shot* offered one Pun2Pun translation CoT example in Figure 3; 3) *CVO* equipped models with a step-wise description of CVO translation model with the same example. Prompts for different settings are in Appendix A.

4.2 Results

Pun understanding generally constitutes no challenge. For textual puns, each model demonstrates varying capabilities in understanding puns (Task I-III) in both Chinese and English, with each excelling in different aspects. For visual puns, yet slightly underachieving in general than textual, similar pattern emerge. Interestingly, qwen model family have a strong tendency of identifying every pun as homophonic. Complete results and analysis are in Appendix B.

Pun2Pun translation is a complex challenge. Based on Table 2 and 3, we have the following discoveries:

1. **Hit and Ovl are generally unsatisfactory.** Even the best-performing models struggle with pun translation across languages, with hit rates rarely exceeding 40% for textual puns and 20% for visual puns, revealing significant room for improvement in preserving both linguistic mechanisms and pragmatic effects.

2. **Creativity is not bold enough.** Most models show positive cosine similarity values, indicating reluctance to deviate sufficiently from

Model	Strategy	Hit↑		Ovl↑		Cos↓	
gpt-4o	Vanilla	20.08	7.62	32.77	18.96	+1.26	-7.45
	1-Shot	**23.34**	11.02	32.14	22.78	+1.91	-7.43
	CVO	20.36	10.22	34.99	21.93	+1.86	-8.61
o3-mini	Vanilla	17.73	5.00	30.96	18.53	+3.09	-6.23
	1-Shot	17.69	5.80	31.35	19.62	+4.77	-6.21
	CVO	20.24	3.00	32.12	19.93	+4.14	-6.95
qwen	Vanilla	12.50	5.40	31.14	20.00	+4.11	-4.76
	1-Shot	10.91	5.00	30.69	20.71	+3.28	-5.56
	CVO	11.13	4.81	31.32	20.14	+3.58	-5.27
qvq	Vanilla	22.47	8.20	35.33	19.83	+2.28	-9.38
	1-Shot	17.20	7.41	33.46	22.44	+1.49	-6.92
	CVO	20.16	8.26	34.50	22.12	+1.14	-9.38
claude	Vanilla	14.29	6.20	33.06	20.04	+0.38	-11.59
	1-Shot	20.28	**17.80**	**40.21**	**23.66**	-0.56	-11.38
	CVO	19.48	13.20	35.64	21.96	**-2.11**	**-12.68**

Table 3: *Translation* Results on Pun2Pun Visual. All metrics are in English(%) + Chinese(%) order, with Cos being relative to 70. qwen, qvq, and claude stand for qwen-vl-max, qvq-72b-preview, and claude-3.5-sonnet respectively.

source semantics to craft effective target-language puns. The few instances of negative values (e.g., deepseek-v3/CVO for English and claude/CVO for visual puns) suggest that greater semantic divergence correlates with improved translation effectiveness.

3. **Homophonic puns are generally harder to translate well.** Across most models and languages, homographic puns consistently achieve higher hit rates and overlap scores than their homophonic ones. This disparity is particularly pronounced in English-to-Chinese translation.

4. **CoT deliberation and CVO strategy have nuanced impact.** While CVO shows notable improvements for certain models (notably deepseek-v3 for English and deepseek-r1 for Chinese), its effect varies significantly across model families. Both claude and qwq show mixed responses to structured reasoning approaches, suggesting that baseline reasoning capabilities and model architecture influence strategy effectiveness more than the strategy itself.

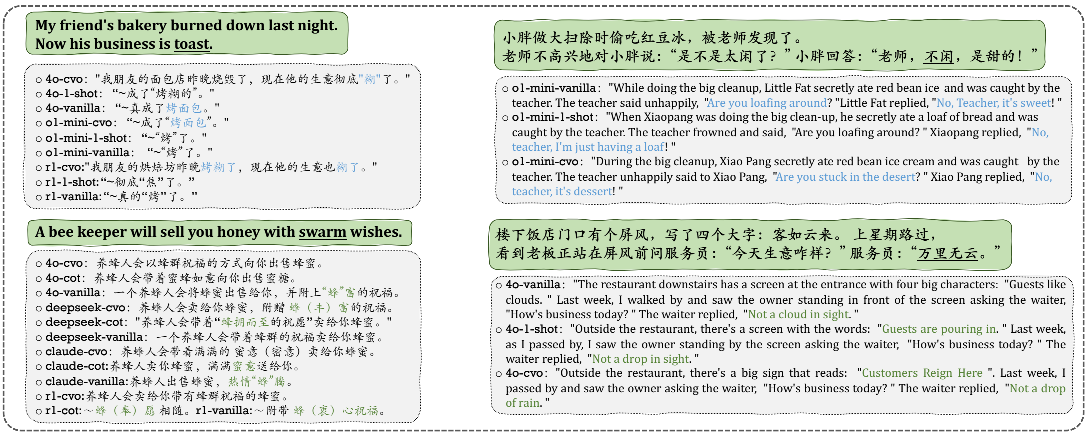
4.3 Discussion

4.3.1 Close Reading

Figure 4 and 6 showcase *Translation* performance on textual and visual puns. Meticulous close reading reveals:

Vanilla < 1-Shot < CVO We observe a general progression of translation quality, as denoted in blue. For instance, gpt-4o/cvo produces " 生意彻底糊了" instead of " 成了烤面包", successfully preserving the wordplay through the character 糊 (burnt/ruined). Similarly, in " 不闲，是甜的！ " (playing on " 闲/咸" meaning "free time/salty"), o1-mini/cvo translates it to "Are you stuck in the desert?" "No, teacher, it's dessert!" - cleverly adapting the wordplay to English phonetics while maintaining the pragmatic effect. Similar pattern exists in visual puns like "down key, looking down on others" (claude/cvo), " 猫教徒 “(qvq/cvo), and " 猫奴" (gpt-4o/cvo).

Vanilla ≥ 1-Shot ≥ CVO In other cases, however, deliberation or CVO procedure might impede translation by over-abstracting or neglecting source text nuances, as denoted in green. As for "swarm wishes", gpt-4o/vanilla produces “ ‘蜂’ 富的祝福”, cleverly playing on " 蜂/丰" (bee/abundant). Meanwhile, CVO translations often produce overly literal renderings like " 以蜂群



My friend's bakery burned down last night.
Now his business is toast.

○ 4o-cvo："我朋友的面包店昨晚烧毁了，现在他的生意彻底"糊"了。"
○ 4o-1-shot：“~成了“烤糊的”。"
○ 4o-vanilla：“~真成了烤面包。"
○ o1-mini-cvo：“~成了“烤面包”。"
○ o1-mini-1-shot：“~“烤”了。"
○ o1-mini-vanilla：“~“烤”了。"
○ r1-cvo:"我朋友的烘焙坊昨晚烤糊了，现在他的生意也糊了。"
○ r1-1-shot:“~彻底“焦”了。”
○ r1-vanilla:“~真的“烤”了。”

A bee keeper will sell you honey with swarm wishes.

○ 4o-cvo：养蜂人会以蜂群祝福的方式向你出售蜂蜜。
○ 4o-cot：养蜂人会带着蜜蜂如意向你出售蜜糖。
○ 4o-vanilla：一个养蜂人会将蜂蜜出售给你，并附上“蜂”富的祝福。
○ deepseek-cvo：养蜂人会卖给你蜂蜜，附赠蜂（丰）富的祝福。
○ deepseek-cot："养蜂人会带着“蜂拥而至的祝愿”卖给你蜂蜜。"
○ deepseek-vanilla：一个养蜂人会带着蜂群的祝福卖给你蜂蜜。
○ claude-cvo：养蜂人会带着满满的蜜意（密意）卖给你蜂蜜。
○ claude-cot:养蜂人卖你蜂蜜，满满蜜意送给你。
○ claude-vanilla:养蜂人出售蜂蜜，热情“蜂”腾。
○ r1-cvo:养蜂人会卖给你带有蜂群祝福的蜂蜜。
○ r1-cot:~蜂（奉）愿相随。r1-vanilla:~附带蜂（衷）心祝福。

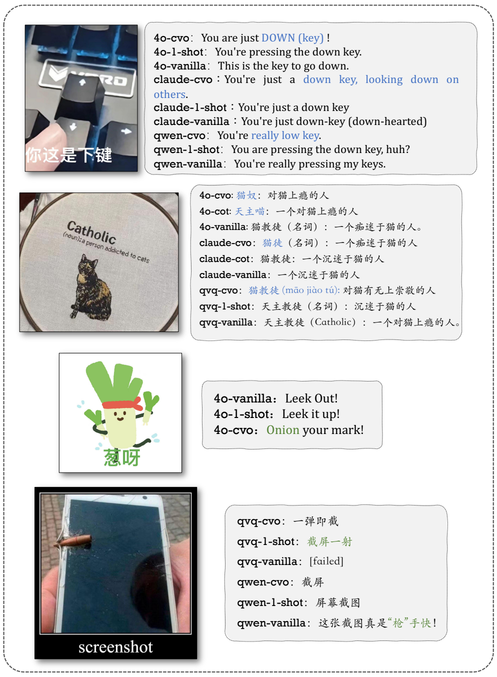
小胖做大扫除时偷吃红豆冰，被老师发现了。
老师不高兴地对小胖说：“是不是太闲了？”小胖回答：“老师，不闲，是甜的！”

○ o1-mini-vanilla："While doing the big cleanup, Little Fat secretly ate red bean ice and was caught by the teacher. The teacher said unhappily, "Are you loafing around? "Little Fat replied, "No, Teacher, it's sweet! "
○ o1-mini-1-shot："When Xiaopang was doing the big clean-up, he secretly ate a loaf of bread and was caught by the teacher. The teacher frowned and said, "Are you loafing around? " Xiaopang replied, "No, teacher, I'm just having a loaf! "
○ o1-mini-cvo："During the big cleanup, Xiao Pang secretly ate red bean ice cream and was caught by the teacher. The teacher unhappily said to Xiao Pang, "Are you stuck in the desert? " Xiao Pang replied, "No, teacher, it's dessert! "

楼下饭店门口有个屏风，写了四个大字：客如云来。上星期路过，
看到老板正站在屏风前问服务员：“今天生意咋样？”服务员：“万里无云。”

○ 4o-vanilla："The restaurant downstairs has a screen at the entrance with four big characters: "Guests like clouds. " Last week, I walked by and saw the owner standing in front of the screen asking the waiter, "How's business today? " The waiter replied, "Not a cloud in sight. "
○ 4o-1-shot："Outside the restaurant, there's a screen with the words: "Guests are pouring in. " Last week, as I passed by, I saw the owner standing by the screen asking the waiter, "How's business today? " The waiter replied, "Not a drop in sight. "
○ 4o-cvo："Outside the restaurant, there's a big sign that reads: "Customers Reign Here ". Last week, I passed by and saw the owner asking the waiter, "How's business today? " The waiter replied, "Not a drop of rain. "

Figure 4: Close Reading on Textual *Translation* Performance



4o-cvo: You are just DOWN (key) !
4o-1-shot: You're pressing the down key.
4o-vanilla: This is the key to go down.
claude-cvo：You're just a down key, looking down on others.
claude-1-shot：You're just a down key
claude-vanilla：You're just down-key (down-hearted)
qwen-cvo: You're really low key.
qwen-1-shot: You are pressing the down key, huh?
qwen-vanilla: You're really pressing my keys.

4o-cvo: 猫奴：对猫上瘾的人
4o-cot: 天主喵：一个对猫上瘾的人
4o-vanilla: 猫教徒（名词）：一个痴迷于猫的人。
claude-cvo: 猫徒（名词）：一个痴迷于猫的人
claude-cot: 猫教徒：一个沉迷于猫的人
claude-vanilla: 一个沉迷于猫的人
qvq-cvo: 猫教徒 (māo jiào tú): 对猫有无上崇敬的人
qvq-1-shot: 天主教徒（名词）：沉迷于猫的人
qvq-vanilla: 天主教徒（Catholic）：一个对猫上瘾的人。

4o-vanilla: Leek Out!
4o-1-shot: Leek it up!
4o-cvo: Onion your mark!

qvq-cvo: 一弹即截
qvq-1-shot: 截屏一射
qvq-vanilla: [failed]
qwen-cvo: 截屏
qwen-1-shot: 屏幕截图
qwen-vanilla: 这张截图真是"枪"手快!

Figure 6: Close Reading on Visual *Translation* Performance

祝福的方式". Similarly, for " 客如云来" (guests arrive like clouds) and " 万里无云" (clear sky), Vanilla's "Guests like clouds" and "Not a cloud in sight" preserves the original wordplay more faithfully than CVO's "Customers Reign Here" and "Not a drop of rain," which inappropriately shifts the conceptual framework. The same holds true for visual puns, as shown in "Onion your mark" (`gpt-4o/cvo`), " 截屏" (`qwen/cvo`), and " 一弹即截" (`qvq/cvo`).

Interesting Findings a) CVO shows potential in transferring surface concepts and improving adaptability in certain cases (a case process is offered in Appendix C); b) model performance varies significantly; c) conceptual overlap between languages facilitates translation—puns involving concepts with cross-cultural equivalents (like "web/网" or "grilled/烤") translate more effectively, while language-specific concepts (like Chinese " 碰酒杯" or English "shakes pear") resist translation; d) visual puns generally prove more challenging than textual ones due to their multimodal nature and cultural embeddedness; e) strategic interchange between pun mechanisms emerges as a potentially effective technique when direct mechanism preservation is impossible, which is further discussed in Section 4.3.3. A detailed analysis of those with cases is in Appendix C.

4.3.2 Optimization Study

Since CVO's essence lies in iterative optimization, we conducted an mechanical iteration study to examine whether simple, repeated refinement could enhance translation quality. We randomly selected 20 textual examples from Pun2Pun dataset and implemented a naive optimization pipeline with `deepseek-r1`, subjecting each translation to five consecutive refinement iterations. Two authors independently evaluated the results using a 5-point scale across three dimensions: innovativeness, content retention, and target language fluency (detailed rubrics are in Appendix C). Results in Figure 5 proved disappointing—while it occasionally showed marked improvement, the overall pattern revealed minimal systematic gains across iterations. This indicates that effective pun translation optimization requires more sophisticated approaches than simple iteration, potentially including reward designs, multi-agent systems, or structured reasoning frameworks that can more intelligently navigate the complex semantic space between languages.

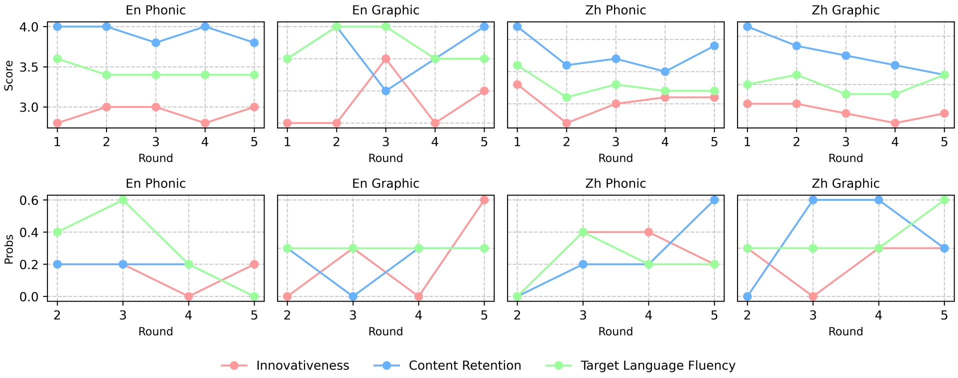


Figure 5: Optimization Study with Naive `deepseek-r1` Iterative Pipeline

4.3.3 Interchange Study

From linguistic intuitions, Chinese and English exhibit fundamentally different characteristics that shape their pun mechanisms. Chinese, with its abundance of homophones (different characters sharing identical pronunciations), naturally favors homophonic puns. By contrast, English, with its rich polysemy but fewer homophones, tends toward homographic wordplay. This linguistic divergence creates an intriguing translation challenge: could models effectively translate puns by switching mechanisms when necessary?

To investigate this phenomenon, we designed an experiment analyzing mechanism interchange patterns using our best-performing models—`deepseek-r1/CVO` for Chinese and `deepseek-v3/CVO` for English. We tracked how pun types transformed during translation, examining whether homophonic puns remained homophonic or converted to homographic, and vice versa. Figure 7 presents our findings as a Sankey diagram. When translating English homophonic puns to Chinese, models frequently convert them to homographic puns. Similarly, Chinese homographic puns often transform into English homophonic puns. Interestingly, Chinese homophonic puns and English homographic puns predominantly retain their mechanism when translated, presumably showing a trajectory dependency. The observed interchange patterns confirm that successful cross-lingual pun translation often requires pragmatic mechanism adaptation rather than rigid structural preservation.

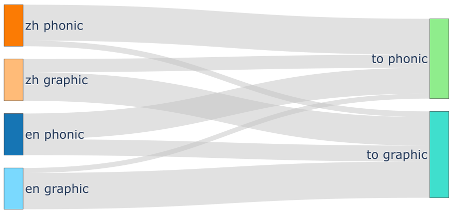


Figure 7: Phonic-Graphic Interchange Study

5 Conclusion

In this work, we introduced Pun2Pun, a novel benchmark for evaluating cross-lingual pun translation between Chinese and English. We established a comprehensive evaluation framework with Constant-Variable Optimization (CVO) Model for translation strategy and concomitant Overlap (Ovl) metric for quality assessment. Through extensive experiments on both textual and visual puns, we observed that our CVO translation strategy shows improvements for certain model families, though overall performance remains modest with hit rates rarely exceeding 40% for textual puns and 20% for visual puns. Our analysis reveals interesting patterns such as mechanism interchange between homophonic and homographic puns as a potential adaptation technique, though this approach requires further investigation to establish its broader effectiveness.

Our findings highlight the substantial challenges that current LLMs face in preserving sophisticated humor across linguistic boundaries, particularly in handling culturally embedded visual puns and maintaining pragmatic effects. While our benchmark provides a foundation for systematic evaluation of cross-lingual pun translation, the modest performance levels achieved suggest that this remains a challenging task requiring continued research effort. These insights contribute to our understanding of the limitations and potential directions for improvement in cross-lingual creative text generation.

Limitations

Data Construction and Subjectivity The inherent subjectivity of humor and pun appreciation introduces challenges in objective data curation. While we employed a three-stage annotation process with multiple author review and external expert consultation for challenging cases, we did not systematically quantify the consistency of annotations across annotators or measure agreement rates. This absence of inter-rater reliability metrics makes it difficult to assess the stability and replicability of our annotation framework.

CVO Implementation While we introduce the CVO framework conceptually, our implementation represents only a rudimentary approximation of its theoretical potential. Future work could develop more sophisticated implementations that better leverage the theoretical underpinnings of this approach, potentially through delicate reward designs, multi-agent systems, or more structured reasoning frameworks. Our current approach does not fully capitalize on the optimization aspects of the CVO model, as evidenced by our optimization study results.

Model Selection Constraints Our evaluation focuses primarily on large-scale and proprietary models, which limits insights into the performance characteristics of smaller, open-source models.

Prompting Strategy Limitations Our investigation of few-shot learning approaches was particularly superficial, without systematic exploration of exemplar variance or impact. Moreover, our prompting strategies also lacked exploration of more sophisticated techniques such as multi-step reasoning frameworks or structured decomposition.

Evaluation Methodology Our heavy reliance on LLM-as-a-judge methods introduces potential biases and consistency issues. The use of `gpt-4o-mini` as our primary judge model creates a systematic dependency that could propagate model-specific biases. While we found these metrics provide useful comparative signals within our experimental framework, they should be interpreted with caution regarding absolute performance levels. The absence of human judgment undermines the reliability and validity of our quantitative results, rendering under-justified whether our automated judgments align with human perceptions of pun quality and humor effectiveness.

The lack of gold-standard reference translations further compounds the issue, though creating high-quality human references for pun translation is exceptionally challenging and resource-intensive given the creative and subjective nature of humor.

Our automated metrics are most reliable for comparing relative performance across models and strategies rather than providing definitive assessments of translation quality, and future work should prioritize establishing human evaluation benchmarks to validate automated approaches.

An intriguing direction for future investigation involves examining how traditional machine translation metrics such as BLEU(Papineni et al., 2001) or COMET(Rei et al., 2020) would evaluate pun translations. Since these metrics typically favor literal semantic alignment, they might systematically penalize the creative deviations and semantic divergence that our analysis shows are often necessary for effective pun translation. Comparing literal machine translations with our more creative pun translations using these conventional metrics could provide valuable insights into the tension between translation fidelity and creative adaptation in humor translation.

Contextual Isolation Our benchmark isolates puns from their broader contextual environments, whereas in natural settings, puns typically serve specific communicative functions within larger discourse contexts. This decontextualization, while methodologically necessary, limits ecological validity and may not reflect the challenges of translating puns within natural conversational or literary contexts.

Limited Language and Cultural Scope Our benchmark focuses exclusively on Chinese-English pun translation, which limits the generalizability of our findings. Our results may or may not extend to other language pairs with different typological relationships. Expanding to other Asian languages, European language pairs, or languages with different writing systems would strengthen the validity of our conclusions and provide broader insights into cross-lingual pun translation mechanisms.

Ethics and Broader Impact Statement

We employed meticulous filtering procedures to minimize biased content during data construction and evaluation. However, given the inherent ambiguity and subjectivity of puns, particularly ones that rely on cultural or symbolic interpretations, we cannot guarantee complete neutrality. We acknowledge that some data samples may contain ethically sensitive, offensive, or culturally aggressive content. We do not endorse such language or implication that may appear in the dataset. Our aim is to improve model performance in challenging linguistic and cultural contexts, not to reinforce or propagate harmful stereotypes or inappropriate humor. We encourage future researchers to continue improving model alignment, cultural sensitivity, and content safety in similar multilingual multimodal settings.

References

Barbara Abbott. 2002. Puns and the structure of language. *Journal of Literary Semantics*, 31(3):233–251.

Digue Albin and Campen Paul. 2022. Automatic Translation of Wordplay.

Anthropic. 2024. Introducing Claude 3.5 Sonnet. `https://www.anthropic.com/news/claude-3-5-sonnet`. Accessed: 2025-4-6.

Hakima Arroubat. 2022. Wordplay location and interpretation with deep learning methods. Proceedings of the Working Notes of CLEF 2022: Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, 2022.

Salvatore Attardo, editor. 2017. *The Routledge Handbook of Language and Humor*, 1 edition. Routledge, New York, NY : Routledge, [2017] | Series: Routledge handbooks in linguistics.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond.

Maciej Besta, Julia Barth, Eric Schreiber, Ales Kubicek, Afonso Catarino, Robert Gerstenberger, Piotr Nyczyk, Patrick Iff, Yueling Li, Sam Houliston, Tomasz Sternal, Marcin Copik, Grzegorz Kwaniewski, Jürgen Müller, ukasz Flis, Hannes Eberhard, Hubert Niewiadomski, and Torsten Hoefler. 2025. Reasoning Language Models: A Blueprint. ArXiv:2501.11223 [cs].

A. Cardford. 1975. *Translation and Untranslatability*. Oxford University Press, Oxford.

Yang Chen, Chong Yang, Tu Hu, Xinhao Chen, Man Lan, Li Cai, Xinlin Zhuang, Xuan Lin, Xin Lu, and Aimin Zhou. 2024. Are U a joke master? pun generation via multi-stage curriculum learning towards a humor LLM. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 878–890, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Yuyan Chen, Zhixu Li, Jiaqing Liang, Yanghua Xiao, Bang Liu, and Yunwen Chen. 2023. Can Pre-trained Language Models Understand Chinese Humor? In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pages 465–480, Singapore Singapore. ACM.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168.

David Crystal. 2006. *The Cambridge Encyclopedia of the English Language*, 2nd edition. Cambridge University Press, Cambridge.

DeepSeek-AI. 2025a. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. ArXiv:2501.12948 [cs].

DeepSeek-AI. 2025b. Deepseek-v3 technical report.

D. Delabastita. 2016. *Traductio: Essays on punning and translation*. Taylor Francis.

Dirk Delabastita. 1993. There's a double tongue: An investigation into the translation of puns. *Target*, 5(2):221–242.

Dirk Delabastita. 2004. Wordplay as a translation problem: A linguistic perspective. In Harald Kittel, Armin Paul Frank, Norbert Greiner, Theo Hermans, Werner Koller, José Lambert, and Fritz Paul, editors, *Übersetzung*, pages 600–606. Walter de Gruyter.

Farhan Dhanani, Muhammad Ra, and Muhammad Atif Tahir. 2023. Humour Translation with Transformers.

Liana Ermakova, Anne-Gwenn Bosser, Adam Jatowt, and Tristan Miller. 2023a. The JOKER Corpus: English-French Parallel Data for Multilingual Wordplay Recognition. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2796–2806, Taipei Taiwan, China. ACM.

Liana Ermakova, Tristan Miller, Anne-Gwenn Bosser, Victor Manuel Palma Preciado, Grigori Sidorov, and Adam Jatowt. 2023b. Overview of JOKER –CLEF-2023 Track on Automatic Wordplay Analysis. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, volume 14163, pages 397–415, Cham. Springer Nature Switzerland. Series Title: Lecture Notes in Computer Science.

Liana Ermakova, Tristan Miller, Orlane Puchalski, Fabio Regattin, Élise Mathurin, Sílvia Araújo, Anne-Gwenn Bosser, Claudine Borg, Monika Bokiniec, Gaelle Le Corre, Benoît Jeanjean, Radia Hannachi, or Mallia, Gordan Matas, and Mohamed Saki. 2022a. CLEF Workshop JOKER: Automatic Wordplay and Humour Translation.

Liana Ermakova, Fabio Regattin, Tristan Miller, Anne-Gwenn Bosser, Claudine Borg, Benoît Jeanjean, Elise Mathurin, Gaelle Le Corre, Radia Hannachi, Sílvia Araújo, Julien Boccou, Albin Digue, and Aurianne Damoy. 2022b. Overview of the CLEF 2022 JOKER Task 3: Pun Translation from English into French.

Quangong Feng. 2019. Cognitive-pragmatic approaches to pun translation. *Foreign Language Research*, 36(3):45–52.

He He, Nanyun Peng, and Percy Liang. 2019. Pun Generation with Surprise. ArXiv:1904.06828 [cs].

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the MATH dataset. In *NeurIPS Datasets and Benchmarks*.

Jack Hessel, Ana Marasovic, Jena D. Hwang, Lillian Lee, Jeff Da, Rowan Zellers, Robert Mankoff, and Yejin Choi. 2023. Do Androids Laugh at Electric Sheep? Humor "Understanding" Benchmarks from The New Yorker Caption Contest. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 688–714, Toronto, Canada. Association for Computational Linguistics.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2024. Livecodebench: Holistic and contamination free evaluation of large language models for code. *CoRR*, abs/2403.07974.

Roman Jakobson. 1959. On linguistic aspects of translation. *Topics in the Theory of Signs and Communication*, 1:114–130.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large Language Models are Zero-Shot Reasoners. ArXiv:2205.11916 [cs].

Waltraud Kolb and Tristan Miller. 2022. Human–computer interaction in pun translation. In *Using Technologies for Creative-Text Translation*, 1 edition, pages 66–88. Routledge, New York.

Huanyong Liu. 2018. Chinesehumorsentiment: Chinese humor sentiment mining including corpus build and nlp methods. `https://github.com/liuhuanyong/ChineseHumorSentiment`. GitHub repository.

Fuli Luo, Shunyao Li, Pengcheng Yang, Lei li, Baobao Chang, Zhifang Sui, and Xu Sun. 2019. Pun-GAN: Generative Adversarial Network for Pun Generation. ArXiv:1910.10950 [cs].

CT Marina Ilari. 2021. Translating humor is a serious business. Accessed: 2025-01-30; By ATA Chronicle of American Translators Association.

Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: investigations in automatic humor recognition. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT '05*, pages 531–538, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Peter Newmark. 1988. *A Textbook of Translation*. Prentice Hall.

Eugene A. Nida and Charles R. Taber. 1964. *The Theory and Practice of Translation*. Brill.

OpenAI. 2024a. Introducing openai o1. `https://openai.com/o1/`. Accessed: 2025-01-30.

OpenAI. 2024b. Introducing SWE-bench verified we' re releasing a human-validated subset of swe-bench that more.

OpenAI. 2024c. Openai o1-mini system card. Accessed: 2025-04-04.

OpenAI. 2025. Openai o3-mini system card. Accessed: 2025-04-04.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, page 311, Philadelphia, Pennsylvania. Association for Computational Linguistics.

Qwen-Team. 2024a. Qvq: To see the world with wisdom. `https://qwenlm.github.io/blog/qvq-72b-preview/`. Accessed: 2025-01-30.

Qwen-Team. 2024b. Qwq: Reflect deeply on the boundaries of the unknown. `https://qwenlm.github.io/blog/qwq-32b-preview/`. Accessed: 2025-01-30.

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.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2023. GPQA: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.

Edwin Simpson, Erik-Lân Do Dinh, Tristan Miller, and Iryna Gurevych. 2019. Predicting humorousness and metaphor novelty with Gaussian process preference learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 5716–5728.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. ArXiv:2201.11903 [cs].

Orion Weller and Kevin Seppi. 2020. The rJokes Dataset: a Large Scale Humor Collection.

Zhiwei Yu, Jiwei Tan, and Xiaojun Wan. 2018. A Neural Approach to Pun Generation. pages 1650–1660, Melbourne, Australia. Association for Computational Linguistics.

Nanfeng Zhang. 2000. On the untranslatability and retranslatability of puns. *Chinese Translators Journal*, 21(4):32–37.

Huijun Zhao. 2012. A quantitative model for pragmatic translation of puns. *Foreign Languages Research*, (5):72–76. 13 citations(CNKI)[6-1-2024].

Huijun Zhao and Yan An. 2020. The meaning optimization of variables in translation of puns. *Foreign Language Research*, (6):92–98.

Shanshan Zhong, Zhongzhan Huang, Shanghua Gao, Wushao Wen, Liang Lin, Marinka Zitnik, and Pan Zhou. 2024. Let's Think Outside the Box: Exploring Leap-of-Thought in Large Language Models with Creative Humor Generation. ArXiv:2312.02439 [cs].

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

A Prompt

A.1 Helper and Judge Model

Prompts of Helper and Judge in all phases are offered in Figure 8, 9, and 10. Note that 1) pun definition in Helper is reused; 2) one Pun2Pun process and outcome example is included in *Theoretical Framework* section of Judge prompt and reused for *CVO* strategy (cvotheory in prompt).

A.2 Task Prompt

```
# Classification
{pun_definition}
Please determine if this
   sentence contains a
   homophonic pun or a
   homographic pun. Output
   'phonic' for homophonic puns
   and 'graphic' for homographic
   puns.

# Locating
{pun_definition}
Please identify where the pun is
   in this sentence.

# Decomposition
{pun_definition}
Please explain the mechanism of
   this pun. For homophonic
   puns, explain how the
   pronunciation is similar or
   identical. For homographic
   puns, explain how multiple
   meanings are formed from a
   single word.

# Appreciation
{pun_definition}
Please explain the image-text
   relationship, cultural
   background, and usage
   scenarios of this pun.

# Translation
{pun_definition}
Your task:
If the original text is in
   Chinese, translate this pun
   into English while preserving
   the original pun effect or
   creating a new pun in the
   target language. Vice versa.
```

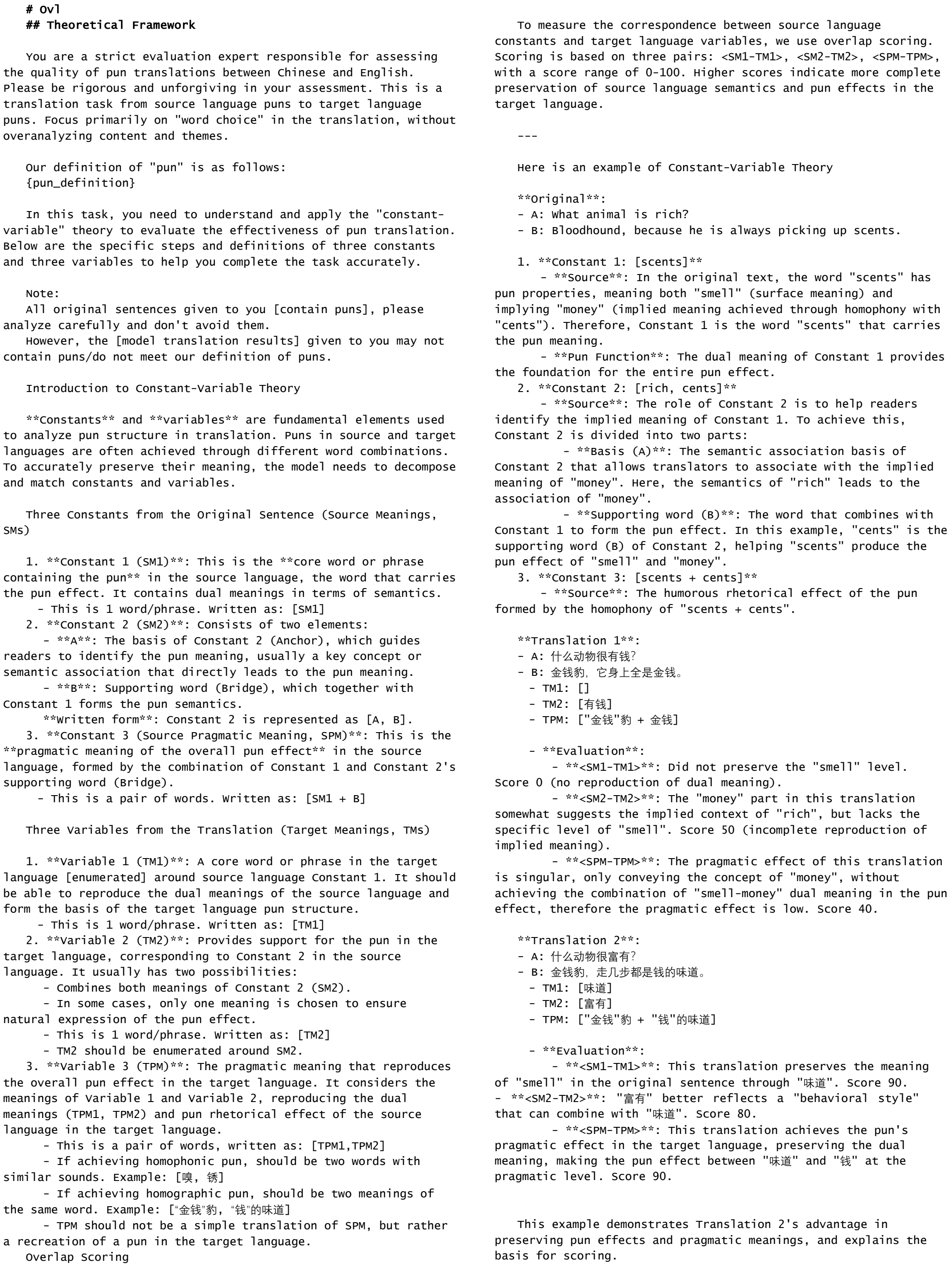
A.3 Strategy Prompt

```
# Vanilla
Let's think step by step like
   this:

Analysis:
...
Final Answer:
...

# 1-Shot
Here is a Pun2Pun Translation
   example:

 Original:
- What animal is rich?
- Bloodhound, because he is
   always picking up scents.

Translation:
- 什么动物很富有？
- 金钱豹，走几步都是钱的味道。

Let's think step by step like
   this:

Analysis:
...
Final Answer:
...

# CVO
The following Constant-Variable
   Optimization Theory can help
   you finish the task.
```

```
Helper Model:

# Pun Definitions:
- Homophonic pun: A pun where two words have the same or similar
pronunciation but different meanings, creating wordplay.
- Homographic pun: A pun where a single word can be understood in
two different ways, or where two words have the same or similar
form but different meanings, creating wordplay.

# Pun Classification

You are a linguistic expert specializing in pun analysis. I will
provide you with a text that may contain a pun, and I need you to
classify it.

Determine whether the text contains a pun, and if so, classify it
as either:
- Homophonic: relying on words that sound the same or similar but
have different meanings
- Homographic: relying on words with the same form that have
multiple meanings (polysemy or homonymy)
- Not a pun: if you believe the text doesn't contain wordplay

OUTPUT FORMAT:
Classification: [Homophonic/Homographic/Not a pun]

Analyze thoroughly before providing your answer. If the text is in
Chinese, pay special attention to potential homophones based on
tone and pronunciation similarities.

# Mechanism Identification

You are a linguistic expert specializing in pun analysis. I will
provide you with a text that contains a pun, and I need you to
identify its mechanism.

Please:
1. Locate the specific punning word or phrase
2. Explain the dual meanings being exploited:
   - For homophonic puns: identify the words that sound similar
and their respective meanings
   - For homographic puns: identify the multiple meanings of the
same word/phrase

OUTPUT FORMAT:
Punning element: [word or phrase]
Meaning 1: [first meaning]
Meaning 2: [second meaning]

Analyze thoroughly before providing your answer. If the text is in
Chinese, pay special attention to potential homophones based on
tone and pronunciation similarities.

# Pun Explanation

You are a linguistic expert specializing in pun analysis. I will
provide you with a text that contains a pun, and I need you to
explain how it works.

Briefly explain how the pun works in 1-2 sentences, highlighting:
- The linguistic mechanism (homophonic or homographic)
- The contextual trigger that activates the dual meanings
- How the ambiguity creates humor

OUTPUT FORMAT:
Mechanism: [brief explanation]

Analyze thoroughly before providing your answer. If the text is in
Chinese, pay special attention to the cultural context that might
affect interpretation.
```

```
Judge Model:

# Locating, Decomposition, Appreciation

You are a helpful assistant that determines if the model
prediction covers the annotation.
Score the model's prediction on a scale of 0-10.
Focus only on content and semantics, ignore the style. Minor
differences or extended explanations are acceptable if it does hit
the annotation.

# Hit
You are a translation expert and native English speaker,
responsible for determining whether the model output contains
valid puns and evaluating their appropriateness and fluency in
English. Please be strict and ensure accurate judgment.

   The model's task is to translate Chinese puns into English
puns (vice versa). Your task is to determine if the given
translation is valid.

   The definition of puns is as follows:
   {pun_definition}

   For homophonic puns, the translation must contain words with
the same or similar pronunciation but different meanings.
   For homographic puns, the translation must contain words with
the same or similar form but different meanings.

   You will be given the original sentence and its translation.
You need to judge according to the following steps:

   1. **Check Translation Fluency**:
      - Determine if the translation follows English grammar
structure and flows naturally. If the translation is unnatural or
doesn't conform to English language conventions, immediately
answer "No" and briefly explain the issues.

   2. **Determine if a Pun Exists**:
      - For homophonic puns, are there words with same/similar
pronunciation but different meanings? If the pronunciation
difference is too large, answer "No" directly.
      - For homographic puns, are there words with same/similar
form but different meanings?

   3. **Analyze Pun Appropriateness**:
      - If a pun exists in the translation, analyze whether it's
appropriate and can be naturally understood in English.
      - For homophonic puns, explain the words with similar/same
pronunciation and their different meanings.
      - For homographic puns, explain the words with similar form
and their different meanings.

   4. **Cultural and Contextual Considerations**:
      - Ensure your judgment considers native English speakers'
comprehension and acceptance. If the pun is unnatural or fails to
create effective humor or double meaning in English, answer "No".
      - We allow translating a source language homophonic pun into
a homographic pun, or a source language homographic pun into a
homophonic pun.
      - We do not allow using parenthetical annotations to convey
the original pun's meaning, nor directly translating both meanings
from the source language.

   Final Answer: Yes/No

   # Ovl (to be continued)
```

Figure 8: Helper and Judge Prompt (Partial)



```
   # Ovl
   ## Theoretical Framework

   You are a strict evaluation expert responsible for assessing
the quality of pun translations between Chinese and English.
Please be rigorous and unforgiving in your assessment. This is a
translation task from source language puns to target language
puns. Focus primarily on "word choice" in the translation, without
overanalyzing content and themes.

   Our definition of "pun" is as follows:
   {pun_definition}

   In this task, you need to understand and apply the "constant-
variable" theory to evaluate the effectiveness of pun translation.
Below are the specific steps and definitions of three constants
and three variables to help you complete the task accurately.

   Note:
   All original sentences given to you [contain puns], please
analyze carefully and don't avoid them.
   However, the [model translation results] given to you may not
contain puns/do not meet our definition of puns.

   Introduction to Constant-Variable Theory

   **Constants** and **variables** are fundamental elements used
to analyze pun structure in translation. Puns in source and target
languages are often achieved through different word combinations.
To accurately preserve their meaning, the model needs to decompose
and match constants and variables.

   Three Constants from the Original Sentence (Source Meanings,
SMs)

   1. **Constant 1 (SM1)**: This is the **core word or phrase
containing the pun** in the source language, the word that carries
the pun effect. It contains dual meanings in terms of semantics.
     - This is 1 word/phrase. Written as: [SM1]
   2. **Constant 2 (SM2)**: Consists of two elements:
      - **A**: The basis of Constant 2 (Anchor), which guides
readers to identify the pun meaning, usually a key concept or
semantic association that directly leads to the pun meaning.
      - **B**: Supporting word (Bridge), which together with
Constant 1 forms the pun semantics.
      **Written form**: Constant 2 is represented as [A, B].
   3. **Constant 3 (Source Pragmatic Meaning, SPM)**: This is the
**pragmatic meaning of the overall pun effect** in the source
language, formed by the combination of Constant 1 and Constant 2's
supporting word (Bridge).
     - This is a pair of words. Written as: [SM1 + B]

   Three Variables from the Translation (Target Meanings, TMs)

   1. **Variable 1 (TM1)**: A core word or phrase in the target
language [enumerated] around source language Constant 1. It should
be able to reproduce the dual meanings of the source language and
form the basis of the target language pun structure.
     - This is 1 word/phrase. Written as: [TM1]
   2. **Variable 2 (TM2)**: Provides support for the pun in the
target language, corresponding to Constant 2 in the source
language. It usually has two possibilities:
      - Combines both meanings of Constant 2 (SM2).
      - In some cases, only one meaning is chosen to ensure
natural expression of the pun effect.
      - This is 1 word/phrase. Written as: [TM2]
      - TM2 should be enumerated around SM2.
   3. **Variable 3 (TPM)**: The pragmatic meaning that reproduces
the overall pun effect in the target language. It considers the
meanings of Variable 1 and Variable 2, reproducing the dual
meanings (TPM1, TPM2) and pun rhetorical effect of the source
language in the target language.
      - This is a pair of words, written as: [TPM1,TPM2]
      - If achieving homophonic pun, should be two words with
similar sounds. Example: [嗅, 锈]
      - If achieving homographic pun, should be two meanings of
the same word. Example: [“金钱”豹, “钱”的味道]
      - TPM should not be a simple translation of SPM, but rather
a recreation of a pun in the target language.
   Overlap Scoring

   To measure the correspondence between source language
constants and target language variables, we use overlap scoring.
Scoring is based on three pairs: <SM1-TM1>, <SM2-TM2>, <SPM-TPM>,
with a score range of 0-100. Higher scores indicate more complete
preservation of source language semantics and pun effects in the
target language.

   ---

   Here is an example of Constant-Variable Theory

   **Original**:
   - A: What animal is rich?
   - B: Bloodhound, because he is always picking up scents.

   1. **Constant 1: [scents]**
      - **Source**: In the original text, the word "scents" has
pun properties, meaning both "smell" (surface meaning) and
implying "money" (implied meaning achieved through homophony with
"cents"). Therefore, Constant 1 is the word "scents" that carries
the pun meaning.
      - **Pun Function**: The dual meaning of Constant 1 provides
the foundation for the entire pun effect.
   2. **Constant 2: [rich, cents]**
      - **Source**: The role of Constant 2 is to help readers
identify the implied meaning of Constant 1. To achieve this,
Constant 2 is divided into two parts:
         - **Basis (A)**: The semantic association basis of
Constant 2 that allows translators to associate with the implied
meaning of "money". Here, the semantics of "rich" leads to the
association of "money".
         - **Supporting word (B)**: The word that combines with
Constant 1 to form the pun effect. In this example, "cents" is the
supporting word (B) of Constant 2, helping "scents" produce the
pun effect of "smell" and "money".
   3. **Constant 3: [scents + cents]**
      - **Source**: The humorous rhetorical effect of the pun
formed by the homophony of "scents + cents".

   **Translation 1**:
   - A: 什么动物很有钱?
   - B: 金钱豹，它身上全是金钱。
     - TM1: []
     - TM2: [有钱]
     - TPM: ["金钱"豹 + 金钱]

     - **Evaluation**:
        - **<SM1-TM1>**: Did not preserve the "smell" level.
Score 0 (no reproduction of dual meaning).
        - **<SM2-TM2>**: The "money" part in this translation
somewhat suggests the implied context of "rich", but lacks the
specific level of "smell". Score 50 (incomplete reproduction of
implied meaning).
        - **<SPM-TPM>**: The pragmatic effect of this translation
is singular, only conveying the concept of "money", without
achieving the combination of "smell-money" dual meaning in the pun
effect, therefore the pragmatic effect is low. Score 40.

   **Translation 2**:
   - A: 什么动物很富有?
   - B: 金钱豹，走几步都是钱的味道。
     - TM1: [味道]
     - TM2: [富有]
     - TPM: ["金钱"豹 + "钱"的味道]

     - **Evaluation**:
        - **<SM1-TM1>**: This translation preserves the meaning
of "smell" in the original sentence through "味道". Score 90.
- **<SM2-TM2>**: "富有" better reflects a "behavioral style"
that can combine with "味道". Score 80.
        - **<SPM-TPM>**: This translation achieves the pun's
pragmatic effect in the target language, preserving the dual
meaning, making the pun effect between "味道" and "钱" at the
pragmatic level. Score 90.


   This example demonstrates Translation 2's advantage in
preserving pun effects and pragmatic meanings, and explains the
basis for scoring.
```

Figure 9: (Continued) Judge Prompt

````
# Ovl
## Step 1: Extract 3 Pairs

   Please first read the following theory:
   ---------------------------------
   {ovl_theory}
   ---------------------------------

   Your task:

   Please analyze the original text and translation of the
following pun, identifying all constants and variables. Output
only a JSON object containing the following fields:

      "SM1": str,
      "SM2": str,
      "SPM": str,
      "TM1": str,
      "TM2": str,
      "TPM": str
   ---------------------------------
   Here are two examples:

   Original:
   - A: What animal is rich?
   - B: Bloodhound, because he is always picking up scents.

   Translation:
   - A: 什么动物很富有？
   - B: 金钱豹，走几步都是钱的味道。

   "SM1": "scents", "SM2": "rich, cents", "SPM": "scents +
cents", "TM1": "气味", "TM2": "金钱", "TPM": "嗅，锈"

   Original:
   ''3.14159265,'' Tom said piously.

   Translation:
   ''3.14159265,'' 汤姆虔诚地说，仿佛在念老天"\pi"的经。

   "SM1": "piously", "SM2": "3.14159265, pi", "SPM": "piously +
pi", "TM1": "虔诚地", "TM2": "π经", "TPM": "\pi，派"
   ---------------------------------

   Finally output one line of jsonl, without ```json``` wrapping.
   Note: We allow type conversion between homophonic puns and
homographic puns during translation. Please identify if there is
type conversion in the translated sentence, do not misjudge it as
having no pun. Please output all the above fields without
omission.
   Please Analyze step by step, output format as follows: (Please
use English prompts "Analysis" and "Extraction", do not wrap
prompts with **, extraction results do not need ```jsonl```
wrapping)

   Preliminaries:

   This is a [homophonic/homographic] pun, playing on the
[homophonic/homographic] relationship between [SPM1] and [SPM2].

   Now, for three source meanings:

   Analysis:
   1. SM1: ...
   2. SM2: ...
   3. SPM: ...
   ...

   Now, for three target meanings:

   Analysis:
   1. TM1: ...(how it came into being through enumeration)
   2. TM2: ...
   3. TPM: ...(how the two parts constitute
homophonic/homographic pun)
   ...

   Extraction:
````

````
   # Ovl
   ## Step 2: Score Overlap

   Please first read the following theory:
   ---------------------------------
   {ovl_theory}
   ---------------------------------

   Your task:

   Based on the extracted pairs, evaluate the overlap between
<SM1-TM1>, <SM2-TM2>, and <SPM-TPM>. The scoring criteria are as
follows:

   1. <SM1-TM1> Scoring Criteria (0-100):
      - 90-100: Completely preserves the dual meanings of the
original pun word, with natural expression
      - 70-89: Basically preserves dual meanings, but expression
is slightly awkward
      - 40-69: Only partially preserves meanings
      - 0-39: Completely loses the dual meanings of the pun word

   2. <SM2-TM2> Scoring Criteria (0-100):
      - 90-100: Completely preserves the contextual support and
semantic association of the original
      - 70-89: Basically preserves contextual support, but
association is weaker
      - 40-69: Contextual support is incomplete
      - 0-39: Completely loses contextual support function

   3. <SPM-TPM> Scoring Criteria (0-100):
      - 90-100: Perfectly recreates pun effect and conforms to
target language expression habits
      - 70-89: Successfully constructs pun but slightly awkward
      - 40-69: Pun effect is weak or expression is unnatural
      - 0-39: Fails to construct pun effect

   Reminder: Please score strictly and keep overall scores low.

   Please analyze step by step, output format as follows: (Please
use English prompts "Analysis" and "Scores", do not wrap prompts
with **, final scores do not need ```jsonl``` wrapping)

   Analysis for SM1-TM1:
   1. ...
   2. ...
   ...
   ovl1: ...

   Analysis for SM2-TM2:
   1. ...
   2. ...
   ...
   ovl2: ...

   Analysis for SPM-TPM:
   1. ...
   2. ...
   ...
   ovl3: ...

   Scores:
   '{"ovl1": float, "ovl2": float, "ovl3": float}'
````

Figure 10: (Continued) Judge Prompt

```
{cvotheory}
Let's think step by step like
    this:

Analysis:
...
Final Answer:
...
```

B Results on Pun Understanding

The results for pun understanding tasks (Tasks I-III, as in Table 4 and 5) demonstrate strong performance across models, though with notable variations in specific capabilities and task types.

Classification Performance For textual puns, most models achieve high accuracy in classification (Task I), with several exceeding 90% accuracy. `claude-3.5-sonnet` shows particularly strong performance on English homophonic puns and Chinese homographic puns. `deepseek-r1` maintains consistent high performance across both languages, achieving over 90% accuracy in most settings.

Interestingly, the `qwen` model family shows a strong bias toward classifying puns as homophonic, particularly evident in their performance disparity between homophonic and homographic classifications. For instance, `qwen-vl-max` achieves high accuracy on English homophonic puns but significantly lower performance on homographic ones with vanilla strategy.

Locating and Decomposition In Tasks II and III (locating and decomposition), models generally maintain strong performance, though with more variation than in classification. `deepseek-v3` and `deepseek-r1` consistently achieve high AAcc scores across both tasks and languages. The CVO strategy often helps improve performance on these tasks, particularly evident in `gpt-4o`'s results where AAcc scores increase by several percentage points with CVO implementation.

Visual Pun Understanding For visual puns, while performance is generally lower than textual puns, models still demonstrate reasonable understanding capabilities. `o3-mini` achieves notably high classification accuracy for both English and Chinese with vanilla strategy, though its performance drops in subsequent tasks. `qvq-72b-preview` shows more balanced performance across all three tasks, maintaining consistent accuracy in classification and reasonable AAcc scores in locating and decomposition.

Strategy Impact The impact of different strategies (vanilla vs. 1-shot vs. CVO) varies across models and tasks. While CVO generally improves performance for more complex tasks (II and III), its benefit is not universal. Some models, particularly those in the `qwen` family, show more consistent performance with simpler strategies.

Overall, these results suggest that current LLMs have strong capabilities in understanding and analyzing puns, though with room for improvement in visual pun processing and more complex decomposition tasks. The varying impact of different strategies across models indicates that pun understanding capabilities may be more inherent to model architecture and training than dependent on prompt engineering.

C Close Reading

Here, we elaborate on five interesting findings stated in Section 4.3.1, offering detailed analysis with concrete examples from model performance on Pun2Pun Textual (together with analysis) and Visual (images are altogether in Figure 12).

C.1 Findings and Analysis

C.1.1 Enhanced Creativity and Adaptability through CVO

When implemented effectively, CVO encourages models to transcend literal translations, instead prioritizing functional equivalence that preserves humor while adapting to target language constraints. An exemplary case in Pun2Pun Textual is offered in Figure 11.

Consider the English homophonic pun: "Life is a puzzle; look here for the missing peace." Using CVO, `gpt-4o/cvo` translated this as " 人生如谜，有时鞋和不了谐" (Life is a puzzle, sometimes shoes can't harmonize). This translation cleverly replaces the "peace/piece" homophonic wordplay with " 鞋/谐" (shoes/harmony) in Chinese, demonstrating cultural adaptation rather than literal translation.

Similarly, for the Chinese homographic pun: " 家有小屁孩，特皮，特烦。我忍无可忍，大喝一声：’滚！’’我又不是轮胎，你叫我怎么滚呀。’" (Having a naughty child at home who is particularly mischievous and annoying. I couldn't tolerate it anymore and shouted: "Roll away!" "I'm not

Okay, I now need to work on this pun translation task provided by the user. First of all, I need to carefully understand the user's requirements and the theoretical framework provided. The user wants to translate a Chinese pun sentence into English while preserving or recreating the pun effect as much as possible, analyzed using the “constant-variable” theory.

First, I need to analyze the original sentence, “**一个饰品公司的董事长，姓古。他的自我介绍是：‘大家都叫我古董，但我的事业很新潮。**’” The key pun here is the word “antique.” In Chinese, “古董” refers to both the chairman of the board, whose last name is Gu (because “董” also means director in Chinese), and “古董” as an ancient object, which is the same as “新潮” in the second half of the sentence. In the second half of the sentence, it is contrasted with “新潮”, which means “new wave”.

Next, according to CVO theory, I need to identify three constants: SM1 is “古董”, which consists of the surname “古” and the position “董” in “董事长”; SPM is “古董” versus “新潮”. The English equivalent of “古董” is “antique”, and the surname needs to be punned with the position. Perhaps a similar sounding surname, such as “Good”, could be considered to mimic “古” and then combine it with the pun on “antique”.

The next step is the enumeration of variables. TM1 may need a word that can be used as a surname and has a related meaning. For example, “Olden” as a surname is close to “old”, and “antique” also means old. For positions, “Chairman” could be retained, but puns would need to be found to combine with “Olden”, e.g. “Chairman Olden” sounds like “antique”. sounds like “antique”, but may need to be restructured.

TM2 then needs to support puns, possibly translating “董事长” as “chairman” while finding words associated with “olden”, for example, “olden” itself means old, echoing “antique”. “The TPM needs to present the contrast between “antique” and “trendy”, possibly with the word “olden”. TPM needs to present a contrast between “antique” and “trendy”, possibly using a combination of “antique” and “trendy”.

Finally, adjust the wording to ensure that the pun is natural. For example, “Chairman Olden” sounds like “antique”, but a more direct connection may be needed. Perhaps keep the last name as “Chairman Gu” and then look for words in English that sound like “Gu” and have a pun potential, such as “Goo”, but that may not be ideal. Or “Mr. Gu”. Or “Mr. Antique” is straightforward, but may not be natural.

All things considered, there may be a need to find a better balance between the surname and the position to ensure that the pun is evident in English. For example, using “Mr. Age” as the last name and “Chairman” as the job title makes “Chairman Age”, which then puns on The word “age” means age and is related to “antique” and the word for career is “trendy”. This might be clearer, but “Age” is not very common as a surname. Alternatively, “Elder” could be used as a surname, “Chairman Elder”, but again this may not be natural.

Ultimately, it may be necessary to compromise by choosing a word that is closer to the pronunciation and has some pun. For example, “Gu” is transliterated as “Goo”, but this may remind the reader of glue, which is not appropriate. Perhaps it would be better to just keep “Gu” as a family name and then explain the pun, but the translation task requires that the pun be recreated, so the English equivalent must be found. Perhaps a better choice would be “Mr. Classic”, because “classic” has the meaning of classic and old, and it might be more natural as a surname, and then the cause would be “cutting-edge”, which is a contrast. edge” for contrast. This may not be enough of a pun, but it makes more sense.

On second thought, the key to the original sentence is the combination of the surname “古” and the post “董” to form “古董”, while the cause is “新潮”. and career is “new wave”. Perhaps in English, the combination of “Mr. Old” as the surname and “Chairman” as the position would be “Chairman Old”, but it doesn't sound like “antique”. “antique”, which may need to be adjusted. Or use “Antique” directly as a surname, “Mr. Antique”, but it may not sound natural. Or look for harmonizing words, such as “Oldham” as a surname, which is close to “old”, but may not be direct enough.

Or use “Mr. Age” as a surname, the post “Chairman Age”, and then explain it as “antique” vs. “trendy”. Or consider other terms, such as “Elder” as a surname and “Chairman Elder”, but again this may not be straightforward.

To summarize, it may be necessary to accept a certain degree of paraphrasing to find a combination of surname and position that creates a pun in English, while maintaining the contrast with the original meaning. Ultimately, the choice of “Chairman Vint” or a similar structure ensures that the pun is obvious and natural at the same time.

Analysis:

SM1: 古董 (homophonic pivot: 古[姓] + 董[职位])

SM2: [董事长(position anchor), 董(bridge term)]

SPM: [古董(antique) + 新潮(trendy)] → semantic contrast

TM1: Olden (surname + "old" semantic tie)

TM2: [Chairman (position) + Olden (bridge)] → "Chairman Olden" ≈ "antique"

TPM: [Olden(name)/olden(archaic), trendy] → preserved contrast

Final Answer:

The chairman of a jewelry company is surnamed **Olden**. He introduces himself by saying, "They all call me **Chairman Olden**, yet my business stays cutting-edge."

Figure 11: A Case of deepseek-r1/CVO Textual *Translation*

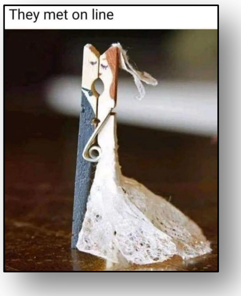
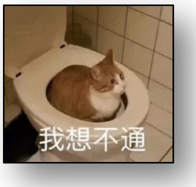
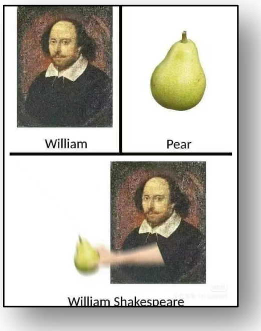
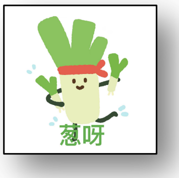
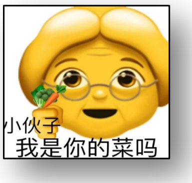
	4o-vanilla: 他们在线上相遇。 4o-cvo: 他们在“线上”相遇。 4o-1-shot: 他们在线上相识。 o3-mini-vanilla：他们在线上相遇了。 o3-mini-cvo：他们在线上相遇。 o3-mini-1-shot：他们在线上相遇。 qvq-cvo：他们在线上相遇 qvq-1-shot：他们在线上相遇了。 qvq-vanilla：他们在线上相遇。
	4o-cvo：I'm stuck. 4o-vanilla：I'm confused. o3-mini-1-shot：I can't pass through" or "I'm stuck o3-mini-cvo：I can't pass through. claude-1-shot：This is so blocked-ing frustrating! claude-cvo：This is beyond my comprehension qwen-1-shot：I can't think through it. qvq-cvo：I'm blocked.
	4o-cvo: 威廉·李斯沃兹 4o-1-shot: 威廉扔梨 4o-vanilla: 威廉 挥梨（威廉·莎士比亚） claude-cvo：威廉摇梨 claude-1-shot：威廉莎士比亚 claude-vanilla：威廉梨思比尔 o3-mini-cvo：威廉·梨斯比尔 o3-mini-1-shot：威廉·莎士比亚”变成“威廉·梨摇” qvq-cvo：当‘威廉’遇上‘梨’，就成了‘威廉·莎士比亚’! qwen-cvo：威廉梨 qwen-1-shot：威廉·梨士比（William Pear-shakespeare） qwen-vanilla：威廉·莎士比亚
	4o-vanilla：Leek Out! 4o-1-shot：Leek it up! 4o-cvo：Onion your mark! claude-vanilla：Lettuce say hi! claude-1-shot：Spring on! claude-cvo：Leeks like I'm in a hurry! o3-mini-vanilla：Leek, huh? o3-mini-1-shot：Leeks, huh? o3-mini-cvo：Leek, huh?
	4o-vanilla：Am I your type? 4o-1-shot：Am I your type? Or am I on your plate? 4o-cvo：Hey cutie, am I your type? o3-mini-1-shot：Am I your dish? claude-1-shot：Hey handsome, am I your cup of tea? qvq-cvo：Am I your type, love? qwen-vanilla：Young man, am I your type of dish? qwen-1-shot：Hey young man, am I your dish? qwen-cvo：Hey young man, am I your type?

Figure 12: Cases of Visual *Translation*

Model	Strategy	English						Chinese					
		I		II		III		I		II		III	
		Acc		AAcc		AAcc		Acc		AAcc		AAcc	
gpt-4o	Vanilla	82.29	69.11	70.76	67.73	78.36	65.08	92.46	54.73	81.98	79.46	86.83	55.57
	1-Shot	85.88	94.94	74.35	75.66	76.52	77.48	91.51	70.49	79.64	73.89	81.37	53.83
	CVO	82.04	96.33	77.44	80.61	72.43	79.83	78.86	71.97	84.84	83.02	83.10	70.27
o1-mini	Vanilla	81.70	91.63	77.76	82.24	71.68	86.39	89.08	58.72	83.71	77.92	80.07	54.63
	1-Shot	81.45	85.73	75.25	84.83	71.82	84.53	90.64	51.07	83.71	80.47	82.06	50.74
	CVO	81.87	86.57	80.60	88.80	50.67	62.49	88.65	52.28	84.66	83.69	69.76	48.72
qwen-vl-max	Vanilla	92.40	23.42	70.43	71.22	71.18	70.98	97.66	6.38	68.72	69.17	85.22	72.95
	1-Shot	79.45	59.96	69.26	64.84	59.31	65.08	93.59	30.20	68.09	67.00	76.98	55.62
	CVO	93.07	16.38	65.83	51.78	44.44	83.74	87.18	20.13	70.07	68.25	75.73	75.48
qwq-32b-preview	Vanilla	76.36	2.11	55.81	64.24	52.13	55.57	87.69	14.10	86.87	83.33	81.26	45.80
	1-Shot	72.35	15.77	49.37	58.04	49.46	64.78	80.16	19.73	80.24	78.26	76.09	48.55
	CVO	75.44	22.70	72.43	78.92	53.30	72.94	89.60	13.02	88.70	86.19	78.46	59.25
deepseek-v3	Vanilla	75.69	52.98	74.32	71.46	72.49	89.89	78.34	47.62	78.80	72.06	92.18	77.59
	1-Shot	75.69	54.12	74.23	74.31	72.01	92.41	91.25	70.82	77.93	73.86	89.94	85.08
	CVO	73.35	92.17	74.90	76.04	61.40	92.19	87.95	72.01	74.11	75.00	82.52	83.49
deepseek-r1	Vanilla	90.90	90.01	74.69	72.37	76.78	83.07	94.97	78.14	77.93	74.61	91.57	54.91
	1-Shot	76.73	93.74	75.86	71.70	70.84	87.78	90.62	78.92	78.71	74.06	85.23	71.37
	CVO	72.18	91.15	73.60	75.38	58.90	88.20	89.29	73.19	75.24	75.13	80.09	69.51
claude-3.5-sonnet	Vanilla	94.65	25.29	70.84	74.65	85.13	90.07	90.99	66.55	69.63	65.77	89.51	76.17
	1-Shot	84.62	97.59	74.60	72.85	79.45	87.24	90.49	70.44	71.23	66.85	84.75	67.48
	CVO	87.05	96.38	74.02	76.94	80.95	85.55	89.05	75.57	76.78	73.22	87.89	71.33

Table 4: Pun2Pun Textual Results on Task I-III. All metrics are in homophonic(%) + homographic(%) order.

Model	Strategy	I		II		III	
		Acc		AAcc		AAcc	
gpt4o	Vanilla	70.44	65.40	79.37	65.20	69.84	52.20
	1-Shot CoT	77.38	41.80	64.88	55.20	34.92	41.00
	CVO CoT	58.73	68.80	65.87	47.00	23.02	27.80
o3-mini	Vanilla	98.21	96.00	65.48	28.60	54.37	19.40
	1-Shot CoT	62.70	71.00	47.82	28.20	24.40	17.40
	CVO CoT	52.38	71.80	48.41	26.60	22.42	16.20
qwen	Vanilla	37.70	83.60	63.40	55.40	57.40	48.10
	1-Shot CoT	31.55	83.00	50.40	45.20	18.80	20.40
	CVO CoT	32.54	83.00	53.60	43.17	11.40	15.83
qvq	Vanilla	92.03	80.20	77.91	58.52	55.82	41.80
	1-Shot CoT	91.47	94.20	80.80	51.62	48.20	28.51
	CVO CoT	94.05	84.00	80.76	60.20	43.69	28.60
claude	Vanilla	62.70	68.00	73.02	49.20	65.08	43.00
	1-Shot CoT	52.18	62.20	62.50	28.40	30.56	28.80
	CVO CoT	74.60	77.60	60.91	41.00	32.14	30.00

Table 5: Pun2Pun visual results on Task I-III. All metrics are in English(%) + Chinese(%) order. qwen, qvq, and claude stand for qwen-vl-max, qvq-72b-preview, and claude-3.5-sonnet respectively.

a tire, how am I supposed to roll?"), qwen/1-shot rendered it as: "Having a little brat at home, so naughty, so annoying. I couldn't take it anymore and shouted, 'Get lost!' 'But I'm not a map, how am I supposed to get lost?'" This translation innovatively maps the Chinese conceptual framework of " 滚" (roll) and " 轮胎" (tire) to the English "get lost" and "map" - maintaining the pun structure while adapting to cultural context, though with some reduction in situational plausibility.

For Chinese homophonic puns, the CVO approach similarly demonstrates creative adaptation. In example: " 女友跟我说，晚上给我妈买箱水。我接完电话马上搬了箱冰露送过去了。刚才女友打电话过来一阵暴怒：啊，让你买香水你买一箱矿泉水！ " (My girlfriend told me to buy a box of water for my mom in the evening. After hanging up, I immediately delivered a box of Binglu [bottled water]. Just now, my girlfriend called, furious: "I asked you to buy perfume, not a box of mineral water!"), o1-mini/cvo translated it as: "My girlfriend told me to buy a bottle of 'perfume' for my mom tonight. After hanging up, I quickly grabbed a bottle of 'sent' and delivered it. Just now, my girlfriend called me furiously: 'I asked you to buy perfume, not a bottle of 'sent'!'" This translation attempts to preserve the phonetic confusion between " 香水" (perfume) and " 箱水" (box of water) by using "perfume" and "sent" (approximating "scent"), though this adaptation somewhat detaches from the original context by omitting the specific reference to mineral water.

These findings confirm our quantitative findings where CVO-enabled translations generally showed lower cosine similarity scores, indicating greater willingness to diverge semantically from source text when necessary to preserve humor. However, as demonstrated particularly in the last example, this creative liberty sometimes results in translations that, while innovative, may sacrifice some contextual coherence or cultural specificity of the original text.

C.1.2 Performance Variation Across Models

We revealed significant performance variations across different models for pun translation. Overall, gpt-4o and o1-mini demonstrated superior creative capabilities, followed by deepseek-r1, claude, and deepseek-v3, while qwen-vl-max and qwq models showed more limited effectiveness.

This pattern becomes evident when examining specific examples. For instance, in translating the Chinese homographic pun: "小明在做眼保健操时对同桌说：我们的新老师好像对我们很客气，以后日子好过了。新老师正好走过来，小明闭着眼，没看到，老师听到后说：别听他的，他闭着眼，说瞎话呢。" (While doing eye exercises, Xiaoming said to his deskmate: "Our new teacher seems very polite to us; life will be better from now on." The teacher happened to walk by, and Xiaoming, with his eyes closed, didn't see her. The teacher heard and said: "Don't listen to him; he's talking nonsense with his eyes closed.")

The pun hinges on "瞎话" which can mean both "talking with eyes closed" and "talking nonsense." Using CVO, gpt-4o, o1-mini, claude, and deepseek-r1 all employed "blindly" in their translations, successfully preserving the dual meaning while maintaining natural English expression. For example, gpt-4o's translation ended with "Don't listen to him; he's speaking blindly with his eyes closed," and claude rendered it as "Don't listen to him - he's talking blind nonsense!" These translations effectively captured both the literal situation and the wordplay.

In contrast, qwen and deepseek-v3 produced more literal translations that missed the punning opportunity, using phrases like "talking nonsense with his eyes closed" that failed to convey the clever dual meaning of the original Chinese expression. The qwq model's attempt ("he's blind right now, telling blind lies") appeared stylistically awkward with repetitive wording and questionable logical coherence.

> 4o-cvo: Xiaoming told his deskmate, "Our new teacher seems to be very polite to us; things will get better from now on." The new teacher happened to walk by, and Xiaoming, with his eyes closed, didn't see him. The teacher heard and said, "Don't listen to him; he's speaking blindly with his eyes closed."

o1-mini-cvo: Xiaoming was doing eye exercises and said to his deskmate, "Our new teacher seems to be very kind to us; life is going to be easier now." Just then, the new teacher walked by. Xiaoming closed his eyes and didn't see him. Hearing this, the teacher said, "Don't listen to him, he's closing his eyes and blindly lying."

deepseek-r1-cvo: Xiao Ming was doing eye exercises and whispered to his deskmate, "Our new teacher seems so polite. Life's gonna be smooth from now on." Just then, the new teacher walked past. With his eyes shut, Xiao Ming didn't notice her. Hearing this, the teacher quipped: "Don't listen to him. He's talking blind!"

claude-cvo: During eye exercises, Ming said to his deskmate: "Our new teacher seems very nice, life will be easier." The teacher happened to pass by, and seeing Ming with his eyes closed, said: "Don't listen to him - he's talking blind nonsense!"

deepseek-v3-cvo: Xiao Ming was doing eye exercises and said to his deskmate, "Our new teacher seems to be very polite to us; our days will be easier from now on." The new teacher happened to walk by, and Xiao Ming, with his eyes closed, didn't see her. Hearing this, the teacher said, "Don't listen to him; he's talking nonsense with his eyes closed."

qwen-cvo: Xiao Ming was doing eye exercises when he said to his deskmate, "Our new teacher seems very polite to us, so life will be better from now on." The new teacher happened to walk over at that moment. Xiao Ming, with his eyes closed, didn't see the teacher. The teacher heard him and said, "Don't listen to him; he's talking nonsense with his eyes closed."

qwq-cvo: Xiaoming, during eye exercises, tells his desk mate: "Our new teacher seems very friendly towards us; things will be easier from now on." The

> new teacher happens to walk by, and Xiaoming, with his eyes closed, doesn't see her. The teacher hears this and says: "Don't believe him; he's blind right now, telling blind lies."

The performance gap becomes even more pronounced with homophonic puns. A Chinese pun plays on " 胡" (a surname) and " 幸福" (happiness): " 某姓胡夫妇去旅游，安检人员对他说：你姓胡？然后他回答：不，我一点也不幸福。安检人员：我问你是不是姓胡？他：我不幸福啊！ " Here, `o1-mini` excelled by transforming the original completely: "Mr. Payne and his wife went traveling. At security, the officer asked him, 'Is your last name Payne?' He replied, 'I'm not in pain!'" This creative adaptation replaced the original Chinese wordplay with an English equivalent that preserves both the humor mechanism and situational context.

> `4o-cvo`: One day, a couple with the last name "Hu" were going through airport security. The officer asked the husband, "Is your last name Hu?" He replied, "No, I'm not happy at all." The officer repeated, "I asked if your last name is Hu." He insisted, "But I'm not happy!"

> `o1-mini-cvo`: Mr. Payne and his wife went traveling. At security, the officer asked him, "Is your last name Payne?" He replied, "I'm not in pain!"

> `deepseek-r1-cvo`: Security officer: "Are you Mr. Woo?" Husband: "No, I'm not who!" Security officer: "I mean, is your surname Woo?" Husband: "I'm not who, I told you!"

> `claude-cvo`: Security: Are you Hugh? - Passenger: No, I'm not happy. - Security: I asked if your name is Hugh! - Passenger: And I said I'm not happy!

> `deepseek-v3-cvo`: A couple with the surname Hu went on a trip. At the security checkpoint, the officer asked, "Are you Hu?" He replied, "No, I'm not happy at all." The officer clarified, "I asked if you are Hu." He responded, "I'm not happy!"

> `qwen-cvo`: Security officer: Are you Mr. Hoo? Mr. Hoo: No, I'm not happy at all. Security officer: I asked if you were Mr. Hoo. Mr. Hoo: I'm not happy!

> `qwq-cvo`: 安检人员说："Do you happen to be 姓 Hu?" 回答者说："No, happiness doesn't happen to me." 安检人员说："I asked if you are 姓 Hu。" 回答者说："I'm not happy, you see."

Similarly, `deepseek-r1` demonstrated creative adaptation by using the surname "Woo" to create phonetic wordplay with "who," maintaining the core humor mechanism. However, `gpt-4o`, `claude`, and `deepseek-v3` struggled with this example, often producing translations that failed to create effective wordplay, merely preserving the literal "Hu/happy" connection which doesn't work as well in English. The `qwen` and `qwq` models produced even less effective translations, with `qwq` notably failing to maintain proper English formatting by mixing Chinese and English inappropriately.

C.1.3 Superior Translation of Shared Concepts

Puns based on concepts shared between Chinese and English cultures tend to translate more effectively than those relying on language-specific features. When the underlying mechanism or cultural reference of a pun has equivalents in both languages, models can more successfully preserve both humor and meaning.

This pattern was particularly evident with homographic puns that rely on polysemy (multiple meanings of words). For example, the English pun "Before he was hired as a short order cook they grilled him" plays on "grilled" having both cooking and interrogation meanings. Most models successfully translated this by employing the Chinese character " 烤" (to roast/grill) in combination with examination-related terms like " 考验" (test) or " 烤问" (a clever blend of "roast" and "question"). The success rate was remarkably high, with 18 out of 21 model-strategy combinations producing effective translations. Models like `gpt-4o`, `o1-mini`, `claude`, `deepseek-v3`, and `deepseek-r1` all maintained the dual meanings consistently across different strategies.

Similarly, translations thrived when conceptual frameworks aligned across cultures. A Chinese pun about a spider and butterfly where the spider is rejected because it "hangs around the web all day" was effectively rendered in English by both `gpt-4o` and `deepseek-v3`. The wordplay on "

网" (web/internet) worked equally well in English with "web/web-surfing," requiring minimal adaptation since the dual meaning exists in both languages.

Another successful example involved a Chinese family joke where everyone likes different animals, but "dad loves the ’ 狐狸精’ next door." The term " 狐狸精" (fox spirit/seductress) was aptly translated as "vixen" by qwen and "foxy lady" by deepseek-v3, both preserving the dual meaning of an actual fox and an attractive, potentially troublesome woman. These translations succeeded because the fox-as-seductress metaphor exists in both Chinese and English cultural frameworks.

In visual puns, we observed similar patterns. The English visual pun with "on line" was successfully translated to Chinese by most models as " 线上" or " 在线上", which preserves both the literal meaning (physically on a line) and the figurative one (online/on the internet).

For a Chinese visual pun showing a toilet with the caption " 我想不通" (literally "I can’t think it through" but visually depicting "I can’t pass through"), models across all three strategies frequently produced apt translations like "I’m stuck," "I can’t pass through," or "I’m blocked." These translations effectively convey both the physical blockage shown in the image and the mental state of confusion or frustration, maintaining the dual meaning present in the original.

These examples demonstrate that when puns rely on semantic or conceptual overlap that exists in both languages rather than language-specific features like phonetics or orthography, models can translate them with relatively high fidelity.

C.1.4 Translation Challenges for Language-Specific Concepts

Certain puns based on language-specific features or cultural idioms presented significant translation challenges for all models, regardless of strategy. These "untranslatable" puns often relied on features unique to the source language with no equivalent mechanism in the target language.

A clear example of this challenge appeared in a Chinese homographic pun where a character wears gloves while drinking because " 我的私人医生已不允许我的手再碰酒杯了" (My personal doctor doesn’t allow my hands to touch wine glasses anymore). The humor hinges on " 碰酒杯," which in Chinese can mean both physically touching glasses and the idiomatic sense of drinking alcohol. When gpt-4o/vanilla translated this using vanilla strategy as "My personal doctor doesn’t allow my hands to even touch a glass anymore," the wordplay was lost because English lacks a similar dual meaning for "touch glasses."

Chinese homophonic puns proved especially resistant to effective translation. For instance, a pun about a child in a spider costume saying " 我是蜘蛛" (I am a spider), which when spoken quickly sounds like " 是只猪" (is a pig), prompted the father to joke, " 猪怎么有八只脚啊?" (Since when does a pig have eight legs?). gpt-4o attempted to preserve this with "I’m a spider (’spy-der’)!" and "Since when does a pig (’spy-d’) have eight legs?" But this invented pronunciation connection fails to create an authentic English pun, as the phonetic similarity that works in Chinese has no natural English equivalent.

Similarly, a Chinese pun playing on " 肉眼" (naked eye) and " 右眼" (right eye) proved untranslatable. A dialogue where a sister warns about bacteria invisible to the "naked eye" (肉眼) and the brother responds he’ll use his "left eye" instead created humor through the similar pronunciation of " 肉" (meat/naked) and " 右" (right). deepseek-r1/cvo translated this as "bacteria are invisible to the naked eye!" with the response "Then I’ll use my *left* eye," which preserves the literal meaning but loses the phonetic wordplay that made the original funny.

Visual puns with culturally specific references faced similar obstacles. A Chinese visual pun featuring the phrase " 有两把刷子" (literally "having two brushes") failed in translation because its idiomatic meaning of "having skill/ability" has no English equivalent. Models like gpt-4o could only produce literal translations ("Two brushes? Tooth brushes!" or "There really are two brushes"), missing the idiomatic dimension entirely.

Conversely, English puns based on specific phonetic patterns also challenged models when translated to Chinese. A visual pun showing "William Shakespeare" represented by "William" with a pear (playing on "shake a pear" sounding like "Shakespeare") proved impossible to render effectively in Chinese. While models like qvq successfully explained the mechanism (" 当’ 威廉’ 遇上’ 梨’, 就成了’ 威廉·莎士比亚’!"), none could create an authentic Chinese pun that preserved both the phonetic play and the visual element. Various attempts resulted in awkward constructions like " 威廉 · 李斯沃兹," " 威廉扔梨," or " 威廉摇梨"

that explained rather than recreated the wordplay.

These examples highlight a fundamental limitation in cross-linguistic pun translation: when the humorous effect depends on linguistic features unique to the source language (specific phonetic patterns, cultural idioms, or language-specific polysemy), even the most sophisticated models struggle to find functional equivalents. In such cases, models typically resort to either literal translation (losing the wordplay) or explanatory notes (losing the spontaneous humor), demonstrating that some aspects of linguistic humor remain resistant to direct cross-cultural translation.

C.1.5 Visual Puns Present Greater Translation Challenges than Textual Puns

Our analysis reveals that visual pun translation consistently underperforms compared to textual pun translation across all models and strategies. This performance gap stems from the inherent complexity of visual puns, which require simultaneous processing of both visual and linguistic elements. Visual puns operate through the interplay between caption text and image content, creating a multimodal semantic space that demands cultural adaptation on multiple levels. When translating visual puns, models must not only negotiate linguistic differences between source and target languages but also reconfigure visual references that may have entirely different cultural interpretations or associations. The image itself remains unchanged during translation, creating a fixed constraint that limits the translator's freedom compared to purely textual contexts. Additionally, visual puns often rely on culturally-specific visual metaphors, symbols, or references that may not exist in the target culture, further complicating the translation process. This multimodal complexity explains why even the most sophisticated models struggle to maintain both humor and coherence when translating visual puns across linguistic and cultural boundaries.

C.1.6 Interchange as an Effective Cross-Linguistic Translation Strategy

Transforming homophonic puns to homographic ones or vice versa—emerges as a particularly effective strategy for cross-linguistic pun translation. This approach accommodates the inherent structural differences between Chinese and English. For instance, when translating the English homophonic pun "A busy barber is quite harried," `gpt-4o/vanilla` transformed it into a Chinese homographic pun: " 忙碌的理发师真是’ 发’ 愁," leveraging the dual meanings of " 发" (hair/to become). Similarly, "The young pine sapling was admonished by his father. Apparently he'd been knotty" was effectively rendered as " 小松树苗被他的父亲责备了，显然他有点儿’ 节外生枝’ 了," converting sound-based wordplay into meaning-based wordplay on literal and figurative interpretations.

The reverse transformation proved equally valuable. When translating the English homographic pun "The prospector didn't think his career would pan out," successful models created a Chinese homophonic pun: " 这位勘探者没想到他的事业最终会小有’ 金’ 喜," where " 金" (gold) creates sound play with its homophone in " 惊喜" (pleasant surprise). Similarly, "A fisherman who was also a pianist was an expert with scales" became " 一个既是渔夫又是钢琴家的人，在’ 调’（钓）上堪称高手," (`deepseek-v3/cvo`) transforming meaning-based wordplay to sound-based play on " 调" (tune/tone) and " 钓" (fishing).

This strategic interchange acknowledges the distinct linguistic features of each language—Chinese with its abundance of homophones and characters with multiple meanings, and English with its rich polysemy but more limited homophony. Models implementing this approach successfully bridge the seeming untranslatability of language-specific humor by reconfiguring not just the lexical components but the fundamental mechanism of the wordplay itself. This finding suggests that the most effective pun translations prioritize functional equivalence of humorous effect over strict preservation of the original wordplay mechanism, allowing greater creative latitude to achieve cross-cultural resonance.

C.2 Rubrics for Optimization Study

Innovativeness (0-5 scale)

- 0: No attempt at creative adaptation; direct word-for-word translation only
- 1: Minimal creativity; slight modification but no effective wordplay
- 2: Basic attempt at wordplay that doesn't fully capture the humor mechanism
- 3: Moderate creativity with functional wordplay that partially preserves humor

- 4: High creativity with effective adaptation of the pun to target language
- 5: Exceptional creativity; creates equivalent or enhanced humor effect with culturally resonant wordplay

Content Retention (0-5 scale)

- 0: Complete content loss; translation bears no relation to original meaning
- 1: Severe content loss; only minimal preservation of original context
- 2: Significant content distortion; core situation partially preserved
- 3: Moderate content preservation; main scenario retained with some alterations
- 4: Strong content preservation; most context elements successfully transferred
- 5: Complete content retention; all key elements of original context preserved

Target Language Fluency (0-5 scale)

- 0: Incomprehensible in target language; broken syntax and nonsensical phrasing
- 1: Poor fluency; awkward phrasing with significant grammatical errors
- 2: Below average fluency; understandable but with unnatural expressions
- 3: Average fluency; generally natural phrasing with minor awkwardness
- 4: Good fluency; natural phrasing that sounds authentic to native speakers
- 5: Excellent fluency; indistinguishable from content written by native speakers