

# Can visual language models resolve textual ambiguity with visual cues? Let visual puns tell you!

Jiwan Chung Seungwon Lim Jaehyun Jeon Seungbeen Lee Youngjae Yu

Yonsei University

[jiwan.chung.research@gmail.com](mailto:jiwan.chung.research@gmail.com)

## Abstract

Humans possess *multimodal literacy*, allowing them to actively integrate information from various modalities to form reasoning. Faced with challenges like lexical ambiguity in text, we supplement this with other modalities, such as thumbnail images or textbook illustrations. Is it possible for machines to achieve a similar multimodal understanding capability?

In response, we present *Understanding Pun with Image Explanations* (🧠 UNPIE)<sup>1</sup>, a novel benchmark designed to assess the impact of multimodal inputs in resolving lexical ambiguities. Puns serve as the ideal subject for this evaluation due to their intrinsic ambiguity. Our dataset includes 1,000 puns, each accompanied by an image that explains both meanings. We pose three multimodal challenges with the annotations to assess different aspects of multimodal literacy; Pun Grounding, Disambiguation, and Reconstruction. The results<sup>2</sup> indicate that various Socratic Models and Visual-Language Models improve over the text-only models when given visual context, particularly as the complexity of the tasks increases.

## 1 Introduction

Humans can actively integrate information from multimodal sources without being explicitly told to. For example, a wink can reveal the insincerity behind a statement about dieting. Similarly, visual aids such as Venn diagrams help students understand abstract concepts such as set theory. This active understanding capacity is often denoted as *multimodal literacy* (Mills and Unsworth, 2017).

In contrast, current multimodal models lack this capacity for active understanding and typically operate under two assumptions: (1) all instructions require visual inputs, and (2) these inputs are relevant (Cui et al., 2023; Zhang et al., 2024). Such

<sup>1</sup>Data: [huggingface.co/datasets/jiwan-chung/VisualPun\\_UNPIE](https://huggingface.co/datasets/jiwan-chung/VisualPun_UNPIE)

<sup>2</sup>Code: [github.com/JiwanChung/VisualPun\\_UNPIE](https://github.com/JiwanChung/VisualPun_UNPIE)

Looks like a very serious  
leak under the sink

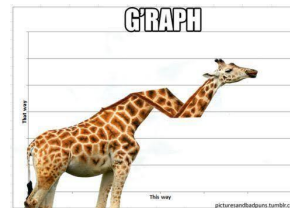


Figure 1: Puns naturally occur with images to enhance understanding (Zenner and Geeraerts, 2018), making them natural candidates for testing active multimodal understanding capacity of machines. Examples of puns accompanied by visual explanations from *r/puns* subreddit on Reddit.

limitations hinder their applicability in real-world scenarios, such as summarizing long blog posts, where irrelevant images must be excluded, and only contextually significant visuals should be used to enhance the understanding of disparate text segments.

An essential component of multimodal literacy is the ability to resolve multimodal ambiguities effectively, which refers to the capacity to disambiguate conflicting or unclear information in modality with information from another modality (Kottur et al., 2021; Guo et al., 2022). Owing to its explicit requirement of multimodal information gathering, disambiguation can serve as a controlled benchmark for evaluating multimodal literacy.

Puns stand as a unique challenge within ambiguity modeling. They are *intrinsically ambiguous* and understanding a pun requires grasping multiple interpretations of a single phrase or word simultaneously. Understanding puns can be difficult even for humans, often necessitating visual cues to clarify the intended interpretation, as demonstrated in Figure 1. Compared to verbose textual explanations, visual cues can deliver instant insight, preserving the humor and cleverness of the pun (Morreall, 1983). Therefore, puns provide an ideal testing ground for assessing models' capabilities in multi-



Figure 2: The UNPIE benchmark comprises three multimodal tasks: 1. Identifying the specific phrase in an English sentence that constitutes a pun, using the provided (a) *pun explanation* image; 2. Choosing the translation of the pun sentence that aligns more closely with the given (b) *pun disambiguator* image; and 3. Reconstructing the English pun sentence from its translated version, aided by the corresponding (a) *pun explanation* image.

modal interpretation.

In this work, we explore model capabilities in *resolving textual ambiguities through visual context*. To this end, we propose Understanding Pun with Image Explanations (🧠 UNPIE), a novel benchmark consisting of 1,000 text-based puns paired with illustrative images that highlight the incongruity within the puns. Additionally, our dataset approaches pun comprehension as a translation task with incomplete information. This method provides a tangible way to measure the often subjective skill of reconstructing puns. Each English pun is accompanied by translations in three different languages—German, French, and Korean—to capture the challenge of reconstructing the puns across diverse linguistic contexts.

We design three tests based on UNPIE to study how models can exploit visual context to aid pun understanding. Figure 2 summarizes the tasks comprising our benchmark. We first consider an English-only task *pun grounding* that challenges machines to identify the specific phrase in a sentence that forms a pun. Next, we formulate a multilingual challenge of *pun disambiguation* where models must choose the translation that best matches the image provided as a pun disambiguator. The final test, *pun reconstruction*, is a comprehensive task where models should recreate the original English pun sentence using a translated version with potentially no ambiguity. For both the pun grounding and reconstruction tasks, we additionally provide the *pun explanation* images as inputs to verify whether models can consider multimodal context when dealing with ambiguous text.

Our comprehensive experiments on UNPIE affirm the presence of multimodal literacy capacity

in two model types: monolithic Visual-Language Models and modular Socratic Models. Incorporating visual context consistently improved performance across our three pun comprehension tests. Notably, this improvement was more pronounced in more challenging tasks. Moreover, VLMs performed better than Socratic Models built on simple image captions. The result suggests that detailed visual understanding is necessary in our benchmark. Finally, fine-tuning with a standard multimodal machine translation dataset adversely affects performance in the pun reconstruction task. This degradation aligns with findings from prior studies (Futeral et al., 2023) stating that web-based multimodal translation datasets may not effectively capture visual dependencies.

Overall, our contributions are as follows:

1. 🧠 UNPIE, a novel benchmark for assessing the multimodal literacy capability of visual-language models. UNPIE is built on text with intrinsic ambiguity (puns), guaranteeing the benefit of visual context.
2. Three new tasks posed on the textual puns and the image annotations: pun grounding, disambiguation, and reconstruction.
3. Experimental results verifying multimodal literacy capability of both VLMs and Socratic Models concerning pun understanding.

## 2 Overview of 🧠 UNPIE Benchmark

UNPIE is a new multimodal multilingual benchmark. Its primary aim is to assess machines' capacity to actively integrate information from visual



Figure 3: Comparison of homographic (left) and heterographic (right) puns in UNPIE dataset along with the respective *disambiguator* visual annotations.


Dataset	Size	Ambiguous (%)	Gen
Multi30k	1000	2 %	✓
CoMMuTE	155	100 %	
 UNPIE	1000	100 %	✓

Table 1: Comparison of UNPIE against multimodal machine translation benchmarks. The statistics for Multi30k are from the *test-2017-flickr* subset. Gen denotes a generative benchmark.

sources to resolve ambiguity in text. Our dataset leverages puns that inherently contain such ambiguity to study the challenge of multimodal literacy in a natural environment.

UNPIE extends puns in two directions: visual context and multilingual translations. First, we collect images for each pun that 1. describes both meanings of the pun to explain it and 2. depicts only one meaning of the pun to disambiguate the pun (section 2.1). While one can naturally retrieve images for disambiguation from the web, images that illustrate the ambiguity of the pun in a single canvas are rare. Thus, we use an off-the-shelf text-to-image model (Betker et al., 2023) to generate such images. We then employ *human annotators* to filter the images so that they correctly explain the given pun. Secondly, we ask human annotators to translate the English pun sentences into multilingual targets (section 2.2). Importantly, the ambiguity should not carry on to the translation target.

## 2.1 Collecting Puns with Visual Context

**Base Text-Only Pun Data.** We build our multimodal multilingual benchmark on top of the text-only English pun dataset of SemEval 2017 Task 7 (Miller et al., 2017). The dataset bounds the pun understanding problem in two ways to rely less on external requirements: first, each sentence con-

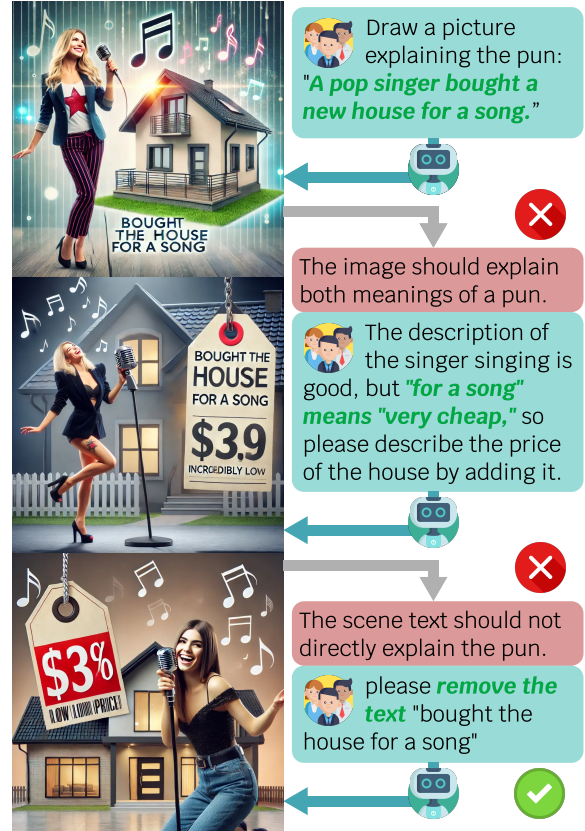


Figure 4: An example of our *pun explanation* image generation process. A human worker interacts with an off-the-shelf text-to-image model, iteratively guiding the model to produce an image that satisfies each specified criterion.

tains a maximum of one pun. Hence, a sentence’s lexical ambiguity is regulated, at least in terms of puns. Second, most pun has a lexical entry in WordNet 3.1 (81% of the whole data). This vocabulary limit keeps our pun generation problem from being dominated by many out-of-vocabulary words.

The data is divided into Homographic and Heterographic puns, depending on the surface form of the puns. As shown in Figure 3, homographic puns have identical spelling and pronunciation but different meanings, while heterographic puns differ in spelling and meanings. We inherit this categorization scheme and report our experiment results category-wise (Homographic and Heterographic). From the SemEval 2017 collection of 2,878 English pun sentences, we selected 500 homographic and 500 heterographic puns with concrete concepts that are more easily visualized through images.

**Generating Pun Explanation Images.** UNPIE is designed to assess a VLM’s capability to resolve lexical ambiguity with visual context. In terms of a pun, the context should depict both meanings

within the pun. Such images are hard to find among natural images due to their complex and sometimes ambivalent meanings. Further, such visual designs are typically proprietary, which contradicts our goal of an open-source dataset. Hence, we resort to creating new images that fit our requirements.

We recruited three NLP researchers to actively prompt the text-to-image generation model DALL-E 3 (Betker et al., 2023) to create images fitting our pun criteria while maintaining a natural appearance. The base text-only dataset provided the puns as data seeds (Miller et al., 2017). While we allow relative freedom in the choice of prompts, the workers reported that DALL-E 3 typically produced satisfactory images with straightforward instructions, as illustrated in Figure 4. Thanks to DALL-E 3’s multi-turn interface, the researchers could request further image revisions if the initial output was unsuitable. On average,  $\sim 24\%$  samples needed such multi-step modification. We obtained 1000 *pun explanation* images after this process.

**Retrieving Pun Disambiguator Images.** UNPIE offers an alternative visual context: per each pun, we attach two images that describe each meaning of the pun. These images disambiguate the pun and are intended to be used in the binary classification task of pun disambiguation explained in section 3.

As a *pun disambiguator* image is aligned to a single meaning, searching for the required image is easier compared to the *pun explanation* images that require encoding both meanings in the same image. Hence, we opt for image retrieval from the LAION 2B web image-text dataset (Schuhmann et al., 2022) rather than image generation. Using the CLIP (Radford et al., 2021)-based image search API (Beaumont, 2022), we retrieve ten images per the meaning of a pun. Then, we manually select the top one that best fits the description. We discard the whole sample when there is no suitable image. We considered two criteria when selecting the images: first, images that explicitly contain the meaning or the pun word itself as printed text are discouraged as such images reward OCR capability rather than general visual understanding. Second, images with watermarks are filtered out to avoid confusion.

## 2.2 Translating Puns to Multilingual Targets

Evaluating a machine’s ability to understand puns is a complex task. Without a rule-based algorithm to measure this capability, the assessment often relies on human judgment or other machines. However, relying on human evaluation can limit the scal-

Model	En→Fr Meaning		En→De Meaning	
	Freq↑	Freq↓	Freq↑	Freq↓
GPT4	68.5	75.9	71.2	73.4
+ Caption	73.4	77.8	74.6	76.6
		Accuracy (%)	Cohen Kappa ( $\kappa$ )	
GPT4Eval	78.1		0.39	

Table 2: Experiment on the effect of meaning frequency in puns. Top: division of pun reconstruction task results according to the commonality of meanings. Bottom: assessment of GPT-4-based meaning frequency ordering against an independent dataset with human-annotated meaning frequencies (Rice et al., 2019).

ability of the assessment process, while machine-based evaluation, such as using models like GPT-4 (OpenAI, 2023), may introduce undesirable biases (Liu et al., 2023c; Hada et al., 2023). To overcome these challenges, we suggest an alternative evaluation method via a downstream task in translation, intentionally aligning with previous research in the field of multimodal machine translation.

**Translation with Machine Assistance.** We translate the original English pun sentence into three languages (German, French, and Korean). Note that we should ensure that the ambiguity in English does not carry over into the translated targets.

We here design a cooperative framework between machines and humans for pun translation. Per each language pair (*e.g.* En  $\rightarrow$  De), we recruit a bilingual worker whose native language is the target language (*e.g.* De). First, we use off-the-shelf translation models to generate three candidates. Then, the human workers select the best one and make further modifications to finalize the translation. This machine-assisted translation aligns with common practices in the industry (Federico et al., 2012). We chose machine-human cooperation for two reasons: firstly, we saw that our human translators find pun translation difficult. Machine suggestions can serve as starting points here. Secondly, this method expedited the annotation process and reduced costs.

**Addressing Lingering Ambiguity.** Certain cases arise where the ambiguity in the source language is retained in the translated text in literal translation. For example, consider the sentence: “A baseball player was a thief. He was always trying to steal.” The pun in this sentence relies on the dual meanings

of “steal”—“to take without permission” and “to steal a base in baseball.” The challenge in translation is twofold: Some languages contain equivalent idiomatic expressions (e.g., “stehlen” in German), which can result in similar ambiguities in the target text. To address this, translators were instructed to select alternative words that avoid unintended double meanings whenever possible. The pun’s humor is implied contextually within the first sentence, even if the pun word itself is not explicitly mentioned. For such instances, indirect translations were permitted, allowing human translators to render distinct interpretations of the pun without preserving its exact wording. To further refine the outputs, we applied text-based deduplication to eliminate closely matching translations. Refer to appendix B for more details.

### 2.3 Dataset Analysis

Our pipeline yields a dataset comprising 500 homographic and 500 heterographic pun sentences, each accompanied by one *pun explanation* image, two *pun disambiguator* images, and translations to three languages.

**How natural are the generated images?** Given the limited availability of real-world images accurately depicting puns, we opted to use AI-generated visuals. To gauge the difference between generated and authentic images, we conducted two human evaluation studies, comparing our generated images against natural image-pun pairs sourced from the web (<https://www.reddit.com/r/puns/>).

In the first study, human evaluators were asked to identify the correct text pun associated with each image from a set of potential matches. Results showed that natural images achieved an accuracy of 86%, while our generated images achieved a slightly higher accuracy of 92%. This test was conducted using a set of 50 randomly selected images. In the second study, we conducted an A/B comparison to assess the perceived naturalness of the images. To ensure consistency, natural images containing multiple panels, written text, or well-known characters were excluded from the evaluation. Across three independent evaluators, the naturalness test resulted in accuracy rates of 66%, 72%, and 74%, respectively, using another set of 50 random images. Overall, despite slight distributional differences between the generated and natural pun images, the disparity is considered acceptable. These findings indicate that evaluations performed within our benchmark can be reasonably

Metric	Translation	Homo	Hetero
Win Rate (%)	Plain	90.7	82.1
	Pun-aware	9.3	17.9
Score (Average)	Plain	94	93.9
	Pun-aware	88.8	87.8

Table 3: Statistical differences between unconditional translation and pun-aware translation, averaged across languages. Text similarity was evaluated using BERTScore (Zhang et al., 2019).

extrapolated to real-world settings.

**Common vs. uncommon meanings.** In UNPIE, each sample contains a pun phrase with two distinct meanings. This section explores how the popularity, or frequency, of each meaning influences downstream performance. To investigate, we rank the meanings of each word by their frequency using zero-shot GPT-4. To ensure the accuracy of GPT-4’s assessments, we cross-reference these with human-annotated frequency data from Rice et al. (Rice et al., 2019), which includes 890 homonyms with annotated frequencies. The lower section of Table 2 compares GPT-4’s frequency rankings with the human-annotated ground truth. Next, using GPT-4, we categorize our data into two groups based on more and less frequent meanings. This categorization is then analyzed through the *pun reconstruction* task outlined in section 3. As illustrated in the upper part of Table 2, the *pun reconstruction* task reveals that inputs with common meanings present more challenges than those with uncommon ones when using GPT-4. This suggests that texts with an uncommon meaning supplement the model’s inherent understanding of the more frequent meaning.

**How different are disambiguated translations from unconditional ones?** When disambiguation is enforced as a strict criterion, the resulting translations are expected to differ from straightforward, unconditional translations. To quantify the extent of this difference, we compare the unconditional translation  $\hat{y}_0$  against two baselines: (1) another unconditional translation produced by a different annotator ( $\hat{y}_1$ ), and (2) the disambiguated translation ( $y$ ). We measure text similarity scores for each pair:  $s_1 = sim(\hat{y}_0, \hat{y}_1)$  and  $s_2 = sim(\hat{y}_0, y)$ , and compute the win rate as the proportion of cases where  $s_2$  exceeds  $s_1$ . The results, summarized in Table 3, show that although disambiguation instructions lead to noticeable changes, the overall

	Model	Inputs	Homo	Hetero
LM	Vicuna	L	69.4	71.2
	Qwen-VL	L	43.8	57.8
	LLaVA	L	76.0	71.8
	GPT-4	L	95.4	92.0
SM	Vicuna	V + L	74.6 (↑ 5.2)	76.6 (↑ 5.4)
	GPT-4	V + L	96.0 (↑ 0.6)	92.4 (↑ 0.4)
VLM	Qwen-VL	V + L	63.6 (↑ <b>19.8</b> )	70.8 (↑ <b>13.0</b> )
	LLaVA	V + L	81.8 (↑ 5.8)	73.0 (↑ 1.2)
	GPT-4	V + L	97.6 (↑ 2.2)	94.0 (↑ 2.0)

Table 4: Results on the pun grounding task. We report the exact match accuracy of the generated pun phrase. ↑ denotes the performance gain from visual context.

difference remains relatively small. Further details can be found in appendix B.

### 3 Task Overview

We pose three multimodal pun understanding tasks on the collected annotations to test models’ capability to use visual context in addressing lexical ambiguity, as illustrated in Figure 2. Each task evaluates different aspects: the easier *Pun Grounding* task can be solved without image input. It is aimed at determining if less advanced models, which might not fully resolve such challenges, can enhance their performance with added visual information. The second task of *pun disambiguation* is designed to necessitate the usage of visual context. Finally, the *pun reconstruction* task replicates a practical multimodal literacy scenario. This task necessitates that models not only use the given translation but also infer or extract the underlying pun meaning that the translation does not explicitly convey, potentially drawing on visual inputs to do so.

**Pun Grounding.** The first step in understanding a pun is to identify it. Our initial task examines whether visual context aids models in identifying pun phrases within sentences. Given the whole English sentence  $x^i = [x_0^i, \dots, x_t^i]$  containing a pun phrase  $s^i = [x_k^i, \dots, x_l^i]$  and its corresponding *pun explanation* image  $v_e^i$ , the model returns a pun phrase candidate  $\bar{s}^i$ . Note that while the actual target phrase  $s^i$  is part of the full sentence  $x^i$ , the model’s output  $\bar{s}^i$  is not bound by this constraint. We purposefully formulate this task as a sequence-to-sequence problem to facilitate zero-shot evaluation across various baselines. The model’s output is then assessed for exact text match with the actual pun phrase to determine accuracy.

**Pun Disambiguation.** Once models pinpoint a

pun’s location, they must then interpret its semantics. Understanding a pun hinges on recognizing the different meanings of the pun phrase, as its humor lies in this ambiguity. In this task, we assess the models’ proficiency in correlating each meaning of the pun with its associated visual context. Given the English sentence  $x^i$  and the *pun disambiguator* image  $v_d^i$  aligned with one of the meanings constructing the pun, the model should produce a translation of the sentence into a target language (e.g. German  $\bar{y}_{De}^i$ ). Notably, the translated text should be free of any ambiguity stemming from the pun, closely aligning with the meaning depicted in the provided image. We compare the model-generated translation  $\bar{y}_{De}^{i,j}$  with two translation targets  $y_{De}^{i,0}, y_{De}^{i,1}$ , each corresponding to a different meaning of the pun. The model’s output is considered correct if it more closely resembles the ground-truth translation  $y_{De}^{i,j}$  that corresponds to the meaning depicted in the image  $v_d^{i,j}$ ,  $j \in 0, 1$ . Refer to section 4.3 for the implementation details.

**Pun Reconstruction.** The final task is to reconstruct the complete pun sentence. To make the problem deterministic, we provide two types of inputs to the model: a non-English language translation of the original pun sentence that has been clarified of any ambiguities (e.g. German  $y_{De}^{i,j}$ ) and the related *pun explanation* image  $v_e^i$ . The model then generates an output  $\bar{x}^i$ , which we compare with the original English pun sentence  $x^i$  to determine if both English sentences encapsulate the same pun. It is a complex task to determine whether two sentences contain the same pun, and we resort to machine-based evaluation with GPT-4 to obtain the binary decision. We verify GPTEval’s validity here using human evaluation in appendix D.

## 4 Experiments on UNPIE benchmark

### 4.1 Models

**LM.** To measure the effectiveness of multimodal modeling, we establish baselines using unimodal text-only language models. We incorporate an open-source model (Vicuna-13B (Chiang et al., 2023)) and the advanced proprietary language model (GPT-4 (OpenAI, 2023)). Furthermore, we appropriate a visual-language model, LLaVA, for a text-only scenario by inputting only text prompts without the images. This approach assesses the concept of *multimodal alignment tax* (Chen et al., 2023) in the context of pun interpretation, implying that fine-tuning a model on visual data might

Model	Inputs	En → De			En → Fr			En → Ko			
		Homo	Hetero	All	Homo	Hetero	All	Homo	Hetero	All	
Random		50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	
SM	Vicuna	V + L	59.4	64.4	61.9	61.4	<u>72.2</u>	66.8	55.4	55.2	55.3
	GPT-4	V + L	<b>68.2</b>	<b>74.6</b>	<b>71.4</b>	<b>69.0</b>	<b>76.8</b>	<b>72.9</b>	<b>65.4</b>	<b>66.2</b>	<b>65.8</b>
VLM	Qwen-VL	V + L	60.7	64.4	62.6	61.7	71.4	66.5	55.4	<u>57.2</u>	56.3
	LLaVA	V + L	<u>65.1</u>	<u>70.8</u>	<u>68.0</u>	61.1	70.6	65.8	<u>58.1</u>	56.9	<u>57.5</u>
	LLaVA-MMT	V + L	63.5	68.0	65.7	<u>64.1</u>	70.0	<u>67.0</u>	56.6	56.1	56.4

Table 5: Experimental results on the pun disambiguation task. All scores are reported in terms of binary classification accuracy. The best scores are **bolded** and the second-best ones are underlined.

impair its original linguistic capabilities. We do not test LM baselines against *pun disambiguation* as the task necessitates visual context.

**SM (Socratic Models).** SM (Zeng et al., 2022), also called pipelining (Bitton-Guetta et al., 2023), is a two-staged framework extending text-only LMs to multimodal tasks by first encoding the multimodal context to textual descriptions. To implement SMs, we employ the same language models as previously mentioned and use BLIP-2 OPT 2.7B (Li et al., 2023) as the visual description generator to encode the images into textual captions.

**VLM.** Monolithic visual-language models directly take the raw images and user queries as inputs to produce textual responses. We employ two popular and high-performing VLMs for this purpose: LLaVA 1.5 13B (Liu et al., 2023a) and Qwen-VL-Chat 7B (Bai et al., 2023). (We refer to Qwen-VL-Chat as Qwen-VL in result tables due to space constraints.) For the tasks of *pun disambiguation* and *pun reconstruction*, we also introduce a machine translation baseline. We thus fine-tune LLaVA with the Multi30k multimodal machine translation dataset (Elliott et al., 2016), yielding the LLaVA-MMT variant. We choose LoRA (Hu et al., 2021) over full fine-tuning for efficient implementation.

## 4.2 Do Images Help Pun Grounding?

**Metrics.** We report accuracy based on the equality of the model-estimated pun phrase and the ground-truth pun phrase. To check the equality, we use the exact match of the surface text form and report the accuracy of the outputs.

**Results.** As anticipated, the incorporation of visual context led to a consistent improvement in pun grounding performance across all models, including Socratic Models and Visual-Language Models (refer to Table 4). Also, GPT-4, a stronger model,

could solve the task even without visual context, verifying our original intention of proposing this task to test the helpfulness of visuals where the task is straightforward but the models are less capable. For evaluation fairness, we employed a standard prompt template across all models (details in appendix E). Note that while careful prompt engineering can further improve the scores, our findings focus on understanding the role of visual context in realistic scenarios rather than extracting the maximum potential from each model.

## 4.3 Can VLMs Disambiguate with Images?

**Metrics.** We conduct a generative evaluation for the pun disambiguation test. The task for the machines is to translate a given pun sentence into a target language, using the accompanying image as a guide to disambiguate the meaning of the pun phrase. In this generative test, the model generates a sequence of text, which is then evaluated against two potential translation targets. The model’s output is considered accurate if it aligns more closely with the translation that corresponds to the context of the provided image. We use BERTScore (Zhang et al., 2019) to measure the text similarity following the human evaluation results in appendix C.

**Results.** All the considered baselines have demonstrated their ability to disambiguate translation outputs based on visual context, as illustrated in Table 5. Both strengthening the language model (Vicuna vs. GPT-4) and improving visual context processing (Vicuna with image captions from BLIP-2 vs. LLaVA) led to more accurate disambiguation. Still, comprehending puns in the textual form was a more decisive factor for pun disambiguation than a stronger visual understanding, as GPT-4 with image captions outperforms all other models. Interestingly, fine-tuning with the Multi30k multi-

De→En											
Model	Inputs	Correct (%)	Homo Bleu-4	METEOR	Correct (%)	Hetero Bleu-4	METEOR	Correct (%)	All Bleu-4	METEOR	
LM	Vicuna	L	27.9	28.8	56.6	16.0	29.1	65.1	22.0	29.0	60.9
	GPT-4	L	43.1	30.1	66.1	45.2	30.7	70.9	44.2	30.4	68.5
	Qwen-VL	L	30.3	29.4	58.8	20.3	30.0	66.7	25.3	29.7	62.8
	LLaVA	L	31.7	27.7	57.9	19.0	29.9	65.6	25.4	28.8	61.8
SM	Vicuna	V + L	35.0(↑7.1)	25.6	51.7	19.1(↑3.1)	26.3	57.5	27.1(↑5.1)	26.0	54.6
	GPT-4	V + L	62.9(↑19.8)	298	655	45.9(↑0.7)	307	685	54.4(↑10.2)	30.3	67.0
VLM	Qwen-VL	V + L	34.3(↑4.0)	28.5	54.2	19.9(↓0.4)	29.7	58.2	27.1(↑1.8)	29.1	56.2
	LLaVA	V + L	33.2(↑1.5)	28.7	55.1	20.1(↑1.1)	29.2	61.2	26.7(↑1.3)	26.0	58.2
	GPT-4	V + L	<b>65.2(↑22.1)</b>	29.9	63.8	50.6(↑5.4)	29.3	65.3	57.9(↑13.7)	29.6	64.6
	LLaVA-MMT	V + L	27.0	12.3	38.1	31.5	25.6	45.7	29.3	18.5	41.9
Fr→En											
LM	Vicuna	L	28.7	28.0	57.6	19.2	29.5	66.6	24.0	28.8	62.1
	GPT-4	L	60.0	30.0	66.2	44.5	30.1	70.3	52.3	30.1	68.3
	Qwen-VL	L	31.5	29.2	59.2	19.9	30.4	67.7	25.7	29.8	63.5
	LLaVA	L	32.6	27.9	58.4	21.0	29.3	67.7	26.8	28.6	63.1
SM	Vicuna	V + L	38.4(↑9.7)	24.0	50.5	18.1(↓1.1)	25.3	55.9	28.3(↑4.3)	24.7	53.2
	GPT-4	V + L	63.6(↑3.6)	29.2	65.1	45.2(↑0.7)	30.7	68.1	54.4(↑2.1)	30.0	66.6
VLM	Qwen-VL	V + L	37.0(↑5.5)	28.1	55.7	22.4(↑2.5)	29.6	61.7	29.7(↑4.0)	28.9	58.7
	LLaVA	V + L	34.3(↑1.7)	28.5	55.3	23.7(↑2.7)	29.6	63.3	29(↑2.2)	29.1	59.3
	GPT-4	V + L	65.6(↑5.6)	29.8	63.0	46.1(↑1.6)	29.3	65.6	55.9(↑3.6)	29.6	64.3
	LLaVA-MMT	V + L	33.3	12.9	39.3	27.0	24.3	43.2	30.2	17.8	41.3
Ko→En											
LM	Vicuna	L	26.3	25.4	48.3	11.1	25.8	48.6	18.7	25.6	48.5
	GPT-4	L	62.7	30.9	69.5	41.8	29.5	65.5	52.3	30.2	67.5
	Qwen-VL	L	26.6	28.8	51.0	12.5	28.1	51.7	19.6	28.5	51.4
	LLaVA	L	27.9	25.4	55.0	11.9	25.5	50.6	19.9	25.5	52.8
SM	Vicuna	V + L	31.9(↑5.6)	20.3	38.7	16.6(↑5.5)	20.3	35.4	24.3(↑5.6)	20.3	37.1
	GPT-4	V + L	68.1(↑5.4)	30.7	69.9	46.4(↑4.6)	29.3	64.4	57.3(↑5.0)	30.0	67.2
VLM	Qwen-VL	V + L	35.5(↑8.9)	26.8	46.2	18.3(↑5.8)	26.7	45.9	26.9(↑7.3)	26.8	46.1
	LLaVA	V + L	30.2(↑2.3)	23.4	41.3	16.4(↑5.0)	23.1	41.0	23.3(↑3.4)	23.3	41.2
	GPT-4	V + L	70.2(↑7.5)	30.1	65.7	52.3(↑10.5)	29.5	61.3	61.3(↑9.0)	29.8	63.5
	LLaVA-MMT	V + L	28.0	6.3	38.5	18.3	15.0	46.7	23.3	10.7	42.6

Table 6: Outcomes for the pun reconstruction task, where ↑ and ↓ signify the performance change attributed to the inclusion of visual context. The model with the largest performance increase is marked **bold** in each language.

modal machine translation dataset (Elliott et al., 2016) harmed the accuracy of visual alignment. The fine-tuned model (LLaVA-MMT) underperforms the zero-shot LLaVA in nearly all aspects, except in the English-to-French translation of heterographic puns. This finding echoes previous research (Futeral et al., 2023), which suggests that multimodal machine translation datasets cannot properly evaluate multimodal literacy capability.

#### 4.4 Do Images Help Pun Reconstruction?

**Metrics.** The pun reconstruction task involves machines using both the human-translated text and the image context to recreate the original pun sentence. Then, the reconstructed pun is compared with the original sentence for consistency in puns. Still, determining whether two sentences share the same pun is a complex task. To tackle this, we use a machine-based evaluation method with GPT-4 (OpenAI, 2023) to determine if the puns in both

sentences are equivalent. To ensure the validity of this approach, known as GPTEval, we further compare it with human annotations in appendix D. Additionally, we report on common text evaluation metrics, such as Bleu-4 (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005)—metrics widely used in the machine translation domain.

**Results.** The results in Table 6 affirm that visual context significantly enhances machines’ ability to reconstruct puns and manage their inherent ambiguity. For all tested models, the inclusion of images consistently improved the accuracy of pun reconstruction. The only exception was the weakest model in both language processing and visual comprehension (SM based on Vicuna). Notably, unlike the main metric of correctness, the automatic text evaluation scores (Bleu-4 and METEOR) did not reflect a clear trend. Through manual inspection of the generated outputs, we saw that such scores were more aligned with changes in the surface form



of the text, which did not necessarily correlate with the accurate identification of puns. This resonates with previous reports stating that such text scores are not fully effective outside of their original domain of machine translation (Liu et al., 2016).

Due to the differences in their forms, models found it more challenging to reconstruct heterographic puns than homographic ones. Notably, incorporating visual context in these more complex scenarios led to significant improvements. Furthermore, the benefit of visual context became even more evident when dealing with Korean inputs; a language typically considered more divergent from English than either German or French. This reinforces the idea that machines depend more on visual cues when tackling complex linguistic tasks. Finally, as in the pun disambiguation task, the fine-tuned LLaVA-MMT suffered from a decline in performance compared to the zero-shot LLaVA. This further supports the notion that visual understanding is necessary to handle UNPIE.

## 5 Related Work

**Multimodal Machine Translation.** By integrating backtranslation as a downstream task, UNPIE contributes to the literature on Multimodal Machine Translation (MMT), a widely studied area that extends neural machine translation with additional visual contexts (Specia et al., 2016; Elliott et al., 2017; Barrault et al., 2018). Previous research argues that visual information can help resolve ambiguities in the source text (Li et al., 2022; Hatami et al., 2022). However, the primary dataset for MMT, Multi30K (Elliott et al., 2016), has limited examples of such ambiguities, leading to questions about the use of MMT for assessing multimodal literacy capacity (Elliott, 2018; Wu et al., 2021; Futeral et al., 2023). Another benchmark counteracts this phenomenon with manual annotation (Futeral et al., 2023; Bawden et al., 2018). Nevertheless, this dataset is relatively small (155 samples) due to the difficulty in pinpointing ambiguities within sentences. Additionally, the benchmark is limited to classification models.

**Computational Pun Understanding.** After early research (Ritchie, 2005) pointed out ambiguity as a key in pun generation, numerous studies have investigated automatic pun generation regarding heterographic puns, which slackens the surface form identity requirement for each meaning of the pun (He et al., 2019; Yu et al., 2020; Mittal et al.,

2022). Other research explored homographic pun generation which is based on multiple meanings of a polysemous word (Yu et al., 2018; Luo et al., 2019; Tian et al., 2022). Recently, Sun et al. (Sun et al., 2022) extended the pun generation problem to consider contextual cues. We extend this line of research with multimodal understanding.

**Visual-Language Models.** The field has seen rapid growth since Flamingo (Alayrac et al., 2022) illustrated the advantages of applying large language models to the visual domain. BLIP-2 (Li et al., 2023), utilizing the OPT language model (Zhang et al., 2022), made significant strides in image captioning. The introduction of a stronger language model (Touvron et al., 2023) further enabled prompt-based control of the models. MiniGPT-4 (Zhu et al., 2023) and LLaVA (Liu et al., 2023b) pioneered the field of visual instruction tuning. InstructBLIP (Dai et al., 2023), an extension of BLIP-2, improved its capability to follow instructions more accurately. Further developments in this domain include other models such as LLAMA-Adapter (Zhang et al., 2023) and Qwen-VL (Bai et al., 2023). Our research puts visual language models (VLMs) to the test regarding their multimodal literacy capabilities.

## 6 Conclusion

We introduced 🍌 UNPIE, a new benchmark for the multimodal literacy capability. Based on UNPIE, we craft three tests to measure how machines can utilize visual context to resolve inherent ambiguity in puns. Our findings indicate that machines can indeed leverage visual information to enhance their understanding of text, as shown by their improved performance across all tasks.

However, achieving human proficiency in multimodal literacy is still a challenge. While our results are encouraging, there remains a considerable gap in machine capability to fully grasp and interpret the intricate relationship between text and visuals, particularly in more complex tasks like *pun reconstruction*. Therefore, we envision UNPIE as not only a platform for testing but also as a starting point for the development of future multimodal models to actively navigate and integrate information from multiple modalities.

## Limitations and Ethical Considerations

UNPIE, while being a multilingual dataset, is built on the English-only pun corpus (Miller et al., 2017).

As such, it primarily models lexical ambiguities unique to English, stemming from polysemies or similar surface forms of the language. To enhance its linguistic diversity and applicability, expanding the dataset to include ambiguities inherent in other languages would be beneficial. Such expansion would not only diversify the linguistic challenges in the dataset but also offer deeper insights into how lexical ambiguities manifest differently across various languages and cultures.

Although UNPIE’s size is much larger than that of the previous multimodal literacy dataset that features explicit ambiguities (Futeral et al., 2023), its total size is insufficient for creating a training split suitable for fine-tuning. This limitation stems from the scarcity of puns, which are inherently challenging for humans to create as well and are not readily available in large quantities online. We thus plan to expand the dataset for multilingual puns in the future.

**Ethical Considerations.** UNPIE, constructed using existing English puns, may inadvertently perpetuate cultural biases and stereotypes present within the humor. Although human annotators were instructed to eliminate any puns expressing explicit hatred, subtle biases can still be perpetuated through seemingly innocuous humor.

To address ethical concerns in the data curation process, we confirmed that all human annotators either volunteered willingly or were compensated fairly for their contributions. We defer the details to appendix B.

## Acknowledgment

This work was partly supported by an IITP grant funded by the Korean Government (MSIT) (No. RS-2020-II201361, Artificial Intelligence Graduate School Program (Yonsei University) and RS-2024-00353131) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2024-00354218).

## References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou,

and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

- Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.
- Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *THIRD CONFERENCE ON MACHINE TRANSLATION (WMT18)*, volume 2, pages 308–327.
- Rachel Bawden, Rico Sennrich, Alexandra Birch, and Barry Haddow. 2018. Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1304–1313.
- Romain Beaumont. 2022. Clip retrieval: Easily compute clip embeddings and build a clip retrieval system with them. <https://github.com/rom1504/clip-retrieval>.
- James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. 2023. [Improving image generation with better captions](#).
- Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. 2023. Breaking common sense: Whoops! a vision-and-language benchmark of synthetic and compositional images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2616–2627.
- Delong Chen, Jianfeng Liu, Wenliang Dai, and Baoyuan Wang. 2023. Visual instruction tuning with polite flamingo. *arXiv preprint arXiv:2307.01003*.
- Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. [Vicuna: An open-source chatbot impressing gpt-4 with 90%\\* chatgpt quality](#).
- Chenheng Cui, Yiyang Zhou, Xinyu Yang, Shirley Wu, Linjun Zhang, James Zou, and Huaxiu Yao. 2023. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. *arXiv preprint arXiv:2311.03287*.
- Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tjong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. [InstructBLIP: Towards](#)

- general-purpose vision-language models with instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Desmond Elliott. 2018. Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2974–2978.
- Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. *arXiv preprint arXiv:1710.07177*.
- Desmond Elliott, Stella Frank, Khalil Sima’an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. *arXiv preprint arXiv:1605.00459*.
- Marcello Federico, Alessandro Cattelan, and Marco Trombetti. 2012. Measuring user productivity in machine translation enhanced computer assisted translation. In *Proceedings of the 10th Conference of the Association for Machine Translation in the Americas: Research Papers*. Association for Machine Translation in the Americas.
- Matthieu Futral, Cordelia Schmid, Ivan Laptev, Benoît Sagot, and Rachel Bawden. 2023. Tackling ambiguity with images: Improved multimodal machine translation and contrastive evaluation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5394–5413.
- Danfeng Guo, Arpit Gupta, Sanchit Agarwal, Jiun-Yu Kao, Shuyang Gao, Arijit Biswas, Chien-Wei Lin, Tagyoung Chung, and Mohit Bansal. 2022. Gravlbert: Graphical visual-linguistic representations for multimodal coreference resolution. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 285–297.
- Rishav Hada, Varun Gumma, Adrian de Wynter, Harshita Diddee, Mohamed Ahmed, Monojit Choudhury, Kalika Bali, and Sunayana Sitaram. 2023. Are large language model-based evaluators the solution to scaling up multilingual evaluation? *arXiv preprint arXiv:2309.07462*.
- Ali Hatami, Paul Buitelaar, and Mihael Arcan. 2022. Analysing the correlation between lexical ambiguity and translation quality in a multimodal setting using wordnet. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 89–95.
- He He, Nanyun Peng, and Percy Liang. 2019. Pun generation with surprise. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1734–1744.
- Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.
- Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. Simmc 2.0: A task-oriented dialog dataset for immersive multimodal conversations. *arXiv preprint arXiv:2104.08667*.
- Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.
- Yi Li, Rameswar Panda, Yoon Kim, Chun-Fu Richard Chen, Rogerio S Feris, David Cox, and Nuno Vasconcelos. 2022. Valhalla: Visual hallucination for machine translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5216–5226.
- Chia-Wei Liu, Ryan Lowe, Iulian Vlad Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.
- Yang Liu, Dan Iter, Yichong Xu, Shuhang Wang, Ruo Chen Xu, and Chenguang Zhu. 2023c. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.
- Fuli Luo, Shun Yao Li, Pengcheng Yang, Lei Li, Baobao Chang, Zhifang Sui, and Xu Sun. 2019. Pun-gan: Generative adversarial network for pun generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3388–3393.
- Tristan Miller, Christian F Hempelmann, and Iryna Gurevych. 2017. Semeval-2017 task 7: Detection and interpretation of english puns. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 58–68.
- Kathy A Mills and Len Unsworth. 2017. Multimodal literacy. Oxford University Press.
- Anirudh Mittal, Yufei Tian, and Nanyun Peng. 2022. Ambipun: Generating humorous puns with ambiguous context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1053–1062.

- John Morreall. 1983. *Taking laughter seriously*. Suny Press.
- OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.
- Caitlin A Rice, Barend Beekhuizen, Vladimir Dubrovsky, Suzanne Stevenson, and Blair C Armstrong. 2019. A comparison of homonym meaning frequency estimates derived from movie and television subtitles, free association, and explicit ratings. *Behavior research methods*, 51:1399–1425.
- Graeme Ritchie. 2005. Computational mechanisms for pun generation. In *Proceedings of the Tenth European Workshop on Natural Language Generation (ENLG-05)*.
- Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.
- Lucia Specia, Stella Frank, Khalil Sima’an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553.
- Jiao Sun, Anjali Narayan-Chen, Shereen Oraby, Shuyang Gao, Tagyoung Chung, Jing Huang, Yang Liu, and Nanyun Peng. 2022. Context-situated pun generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4635–4648.
- Yufei Tian, Divyanshu Sheth, and Nanyun Peng. 2022. A unified framework for pun generation with humor principles. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3253–3261.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Zhiyong Wu, Lingpeng Kong, Wei Bi, Xiang Li, and Ben Kao. 2021. Good for misconceived reasons: An empirical revisiting on the need for visual context in multimodal machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6153–6166.
- Zhiwei Yu, Jiwei Tan, and Xiaojun Wan. 2018. A neural approach to pun generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1650–1660.
- Zhiwei Yu, Hongyu Zang, and Xiaojun Wan. 2020. Homophonic pun generation with lexically constrained rewriting. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2870–2876.
- Andy Zeng, Maria Attarian, Krzysztof Marcin Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael S Ryoo, Vikas Sindhwani, Johnny Lee, et al. 2022. Socratic models: Composing zero-shot multimodal reasoning with language. In *The Eleventh International Conference on Learning Representations*.
- Eline Zenner and Dirk Geeraerts. 2018. One does not simply process memes: Image macros as multimodal constructions. *Cultures and traditions of wordplay and wordplay research*, pages 167–194.
- Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. 2023. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.
- Yichi Zhang, Yao Huang, Yitong Sun, Chang Liu, Zhe Zhao, Zhengwei Fang, Yifan Wang, Huanran Chen, Xiao Yang, Xingxing Wei, et al. 2024. Benchmarking trustworthiness of multimodal large language models: A comprehensive study. *arXiv preprint arXiv:2406.07057*.
- Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

## A Hyperparameters & Setup

**Models.** In all GPT-4 (OpenAI, 2023) usage, we use the `gpt-4-0613` endpoint. When conducting experiments with open-source models, we leverage the official implementation codes in conjunction with publicly available weights from the Huggingface Hub (<https://huggingface.co/models>). For our work with the Vicuna (Chiang et al., 2023) language model, we employed the `lmsys/vicuna-7b-v1.5` endpoint. Additionally, for the LLaVA 1.5 13B (Liu et al., 2023a) and Qwen-VL-Chat (Bai et al., 2023) visual-language models, we used the following model parameters: `liuhaotian/llava-v1.5-13b` and `Qwen/Qwen-VL-Chat`, respectively.

**Text Generation.** We use deterministic greedy sampling for all experiments, introducing no randomness or external hyperparameter in the text generation process. Except for GPT-4, all models were allowed to generate up to 200 tokens with the freedom of an early stopping.

**Computational Resources & Fine-tuning.** We used the OpenAI API for inferring GPT-4 (OpenAI, 2023) outputs. For open-source models such as Vicuna (Chiang et al., 2023), LLaVA (Liu et al., 2023a), and Qwen-VL-Chat (Bai et al., 2023), we use a single NVIDIA A100 40GB GPU for inference. While the exact inference speed varies depending on the length of prompts and responses, a query takes about  $\sim 0.8$  seconds to terminate when utilizing batch processing. Fine-tuning LLaVA was also possible in a single A100 40GB GPU thanks to the efficient LoRA-based implementation (Hu et al., 2021). We trained each translation model for ten epochs in the training split of the Multi30k dataset (Elliott et al., 2016) with early stopping, which took  $\sim 20$  hours on average.

## B Data Collection Details

**Generating Pun Explanation Images.** We rely on the DALL-E 3 (Betker et al., 2023) image-to-text model to generate images that explain both meanings of the pun at the same time. DALL-E 3 can be accessed either from the GPT-4 (<https://chat.openai.com/>) or Bing Image Generator (<https://www.bing.com/create>) web interface. Our annotators employed both interfaces.

In contrast to previous studies that employed designers for image generation tasks involving machine-human collaboration (Bitton-Guetta et al., 2023), we chose to recruit three NLP researchers to

generate images with DALL-E 3 and curate the outputs. This decision was driven by the necessity for an accurate representation of the puns’ meanings rather than the artistic quality of the images, considering the specific requirements of our research task. The NLP researchers participated voluntarily in this annotation task.

**Translating Puns.** To streamline the translation process while ensuring clarity in meaning, we adopted a machine-assisted translation approach. This method simplifies the task for human annotators, who are required to ensure that the translations clearly reflect the intended meaning of the text. Initially, we provide three machine-generated translation options for the annotators to select and refine. These options are created using GPT-4, which we prompt in two distinct ways, and DeepL (<https://www.deepl.com/translator>), a proprietary translation service.

Human annotators then use these machine-provided translations as a base to craft the final version of the translated pun sentence, as depicted in the user interface shown in Figure 5. For each of the three target languages (German, French, and Korean), we engaged a native speaker who is also proficient in English, ensuring both linguistic accuracy and fidelity to the original pun’s meaning. The extent for how much the translations written by human annotators were drifted from unconditional translation is reported in Table 7. We pay each annotator 12–15\$ per hour. We have received approval from the university department and conducted data collection.

## C Testing the Validity of BERTScore

In the Pun Disambiguation task, we require the models to generate the disambiguated text. Thus, we need an automatic algorithm to decide whether the generated output aligns with the intended pun meaning. We formulate this as a text-only problem of matching the output with the ground-truth disambiguated translation result.

We consider various options here: three translation metrics (Bleu, METEOR, and Rouge-L) and two model-based metrics (BERTScore (Zhang et al., 2019) and GPT-based Evaluation). Note that GPTEval here does not receive images as inputs following the other options. Given the generation output of a VLM (LLaVA) or a Socratic Model (GPT4), we ask human annotators for a ternary classification: Match, No Match, and Invalid. A

Metrics	Translation	De		Fr		Ko	
		Homo	Hetero	Homo	Hetero	Homo	Hetero
Win Rate (%)	Plain	93.2	80.0	92.4	81.0	86.4	85.4
	Pun-aware	6.8	20.0	7.6	19.0	13.6	14.6
Score (Average)	Plain	95.2	95.0	95.5	95.4	91.3	91.4
	Pun-aware	84.5	88.0	96.6	89.6	85.0	85.9

Table 7: Statistical differences between unconditional translation and pun-aware translation. Text similarity was evaluated using BERTScore (Zhang et al., 2019).

Method	LLaVA		GPT4	
	Acc (%)	$\phi$	Acc (%)	$\phi$
BLEU	84	57	84	59
METEOR	82	52	80	49
Rouge-L	84	41	82	54
BERTScore	<b>93</b>	<b>81</b>	<b>89</b>	<b>72</b>
GPTEval	76	41	76	43

Table 8: We compare each metric with human judgments on a set of 100 samples.  $\phi$  denotes the Phi correlation coefficient.

sample is classified as matched if it is better aligned to the intended translation target than the other one and as not matched vice versa. We also let the annotators mark invalid outputs in which the text does not align with either target. Finally, we compare the human decisions with automatic algorithms on 100 valid samples. We filtered out 32 invalid outputs from LLaVA and none from GPT4.

The results show that BERTScore greatly improves over the traditional translation metric baselines. As the disambiguation of a pun sentence typically lies in the correct translation of salient phrases, conventional metrics without semantic understanding are not sufficient for the task. On the other hand, BERTScore shows an acceptable correlation with human annotations. We thus employ it as the metric for the Pun Disambiguation task. Perhaps surprisingly, the strong LLM backbone of GPTEval yields the worst outcome. Upon qualitative examination, we saw that GPT tends to favor certain targets regardless of the input. We leave alleviating this bias of GPTEval to future research.

## D Testing the Validity of GPTEval

To ensure the reliability of our machine-based evaluation method, we conducted a comparative analysis with human assessments. This involved man-

ually annotating 100 sample outputs from a multi-modal approach (GPT-4 with image captions) and a text-only approach (GPT-4). The findings in Table 9 indicate that GPT-4’s evaluations can be considered dependable for assessing the performance of machine-generated outputs in the pun reconstruction task.

## E Prompt Templates

### E.1 Pun Grounding

- LM

```

###Variables
PUN_SENTENCE

###PROMPT
[sentence]: {PUN_SENTENCE}
This is a pun sentence. Identify the
↪ specific word or phrase
that creates the pun. Respond with
↪ only the word or phrase that
makes it a pun, without any
↪ explanation.
[answer]:

```

- SM or VLM

```

###Variables
PUN_SENTENCE, IMAGE

###PROMPT
[sentence]: {PUN_SENTENCE}
This is a pun sentence. Identify the
↪ specific word or phrase
that creates the pun, given the image
↪ as context "{IMAGE}".
Respond with only the word or phrase
↪ that makes it a pun,
without any explanation.
[answer]:

```

### E.2 Pun Disambiguation

- SM or VLM

	Multimodal					Text-Only				
	Precision	Recall	F1	Acc (%)	$\rho$	Precision	Recall	F1	Acc (%)	$\rho$
De → En	0.91	0.95	0.93	89	0.65	0.91	0.92	0.92	87	0.62
Fr → En	0.92	0.92	0.92	88	0.62	0.91	0.90	0.91	86	0.62
Ko → En	0.84	0.86	0.85	80	0.54	0.88	0.88	0.88	85	0.66

Table 9: Test results of machine-based evaluation scheme using GPT-4 in the pun reconstruction task. We compare GPT-4’s decision with human judgments on a set of 100 samples.  $\rho$  denotes the Pearson correlation coefficient.

```

###Variables
PUN_SENTENCE, IMAGE, LANGUAGE

###PROMPT
[English]: {PUN_SENTENCE}
Translate the given English sentence
↳ into {LANGUAGE}, given the
image as context "{IMAGE}". Please
↳ respond using the format
below:
[LANGUAGE]:

```

```

###Variables
PUN_SENTENCE_IN_LANGUAGE, LANGUAGE,
↳ IMAGE

###PROMPT
Please translate the following
↳ sentence from {LANGUAGE} into
English, ensuring that the
↳ translation contains a pun. I
↳ will
provide you with a sentence in {
↳ LANGUAGE} using the format
[LANGUAGE]: "{
↳ PUN_SENTENCE_IN_LANGUAGE}".

[Image Description]: {IMAGE}
This image description is about two
↳ meanings of the word that
you are expected to create. Use this
↳ information to craft your
pun-inclusive English translation.
↳ Please respond using the
format below:
[English]:

```

### E.3 Pun Reconstruction

- LM

```

###Variables
PUN_SENTENCE_IN_LANGUAGE, LANGUAGE

###PROMPT
Please translate the following
↳ sentence from {LANGUAGE} into
English, ensuring that the
↳ translation contains a pun. I
↳ will
provide you with a sentence in {
↳ LANGUAGE} using the format
[LANGUAGE]: "{
↳ PUN_SENTENCE_IN_LANGUAGE}".
↳ Please respond using
the format below:
[English]:

```

- SM or VLM

### Instruction

#### Translation Task: English Pun to French

You will receive a sentence in English that includes a pun—a word with multiple meanings or a homophone that makes the sentence humorous or tricky. For the purpose of this task, we will clarify one specific meaning of the pun word.

#### Your Objective:

Translate the English sentence into French, using the provided meaning of the pun to resolve any ambiguity. This means your translation should clearly reflect only the specified meaning and not the pun itself.

#### Process:

1. Review the three machine-generated French translations we provide.
2. Choose the translation that seems most accurate based on the given meaning.
3. If necessary, make adjustments to the chosen translation to ensure accuracy and naturalness in French.

Please ensure that the essence of the sentence is preserved while the pun is translated according to the given context, maintaining the integrity of the intended message.

#### Pun Sentence

Can honeybee abuse lead to a sting operation?

#### Pun Word

sting

#### Word Meaning

the sharp pain caused by the venomous injection of an insect, such as a bee or wasp.

#### (Suggestions) Machine-Generated Translations

1. Est-ce que nuire aux abeilles pourrait entraîner leur défense et leur piqûre?
2. Est-ce que l'abus d'abeilles à miel peut mener à une piqûre?
3. L'abus d'abeilles peut-il conduire à une opération de piqûre?

#### Your Translation

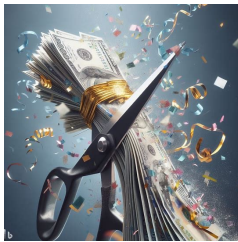
Your Translation

Submit

Figure 5: A screenshot of the human annotation interface for pun-aware text translation.



**Pun Sentence: He and his partner made knives, and they shared a cut.**



**cut**

to divide something with a sharp object



**Fr:** Lui et son partenaire ont fabriqué des couteaux, et ils se sont engagés dans un processus de **division ou de séparation** de matériaux avec des objets pointus.

**De:** Er und sein Partner produzierten Messer und beschäftigten sich mit dem Prozess des **Teilens oder Trennens** von Materialien mit scharfen Objekten.

**Kr:** 그와 그의 파트너는 칼을 만들었고, 그들은 날카로운 물체로 재료를 분할하거나 분리하는 과정에 참여했습니다.

**cut**

share of the profits from a business



**Fr:** Lui et son partenaire ont produit des couteaux et ils ont chacun reçu une **part des bénéfices** gagnés de cette entreprise.

**De:** Er und sein Partner stellten Messer her und sie erhielten jeweils einen **Anteil an den Gewinnen** aus diesem Vorhaben.

**Kr:** 그와 그의 파트너는 칼을 생산했고 그들 각각이 이 과정에서 벌어들인 이익의 일부를 받았습니다.

**Pun Sentence: The science teachers broke up because there was no chemistry between them.**



**chemistry**

science studying matter and its interactions



**Fr:** Les professeurs de sciences se sont séparés parce qu'il n'y avait aucune collaboration ou interaction sur le sujet de la **chimie** entre eux.

**De:** Die Naturwissenschaftslehrer trennten sich, weil es zwischen ihnen keine Zusammenarbeit oder Interaktion im Fach **Chemie** gab.

**Kr:** 그들 사이에 화학에 관한 협업이나 상호 작용이 없어서 과학 교사들이 헤어졌습니다.

**chemistry**

sense of attraction between two people



**Fr:** es professeurs de sciences se sont séparés parce qu'il n'y avait pas d'**attraction romantique ou de sentiment** de connexion entre eux.

**De:** Die Naturwissenschaftslehrer trennten sich, weil es keine **romantische Anziehung oder Gefühl** der Verbundenheit zwischen ihnen gab.

**Kr:** 그들 사이에 로맨틱한 끌림이나 연결감이 없었기 때문에 과학 선생님이 헤어졌습니다.

**Pun Sentence: Diabetics should not be allowed to have sweet dreams.**



**sweet**

having a sugary taste, like candy or sugar



**Fr:** Les diabétiques ne devraient pas être autorisés à rêver de nourritures **sucrées** car ce n'est pas sain pour eux.

**De:** Diabetiker sollten nicht davon träumen dürfen, **zuckerhaltige** Lebensmittel zu essen, weil es nicht gesund für sie ist.

**Kr:** 당뇨병 환자들은 **설탕이 많은** 음식에 대해 꿈꾸는 것이 건강에 좋지 않으므로 그것이 허용되어서는 안됩니다.

**sweet**

pleasant or agreeable



**Fr:** Les diabétiques ne devraient pas pouvoir avoir des rêves **agréables ou plaisants**.

**De:** Diabetiker sollten keine **angenehmen oder erfreulichen** Träume haben können.

**Kr:** 당뇨병 환자들이 **즐거운 또는 기쁜** 꿈을 꾸지 못해야 합니다.

Figure 6: Example annotations of homographic puns in  UNPIE benchmark.

**Pun Sentence:** I tried to record an album in a reptile shop, but there was a terrible **gecko**.



**gecko**

small, usually nocturnal, carnivorous lizard



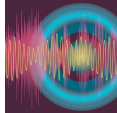
**Fr:** J'ai essayé d'enregistrer un album dans un magasin de reptiles, mais il y avait un horrible **gecko**.

**De:** Ich habe versucht, ein Album in einem Reptilienladen aufzunehmen, aber es gab eine schreckliche **Gecko**.

**Kr:** 나는 파충류 가게에서 앨범을 녹음하려고 했지만, 끔찍한 **도마뱀**이 있었습니다.

**echo**

a sound caused by the reflection of sound waves



**Fr:** J'ai essayé d'enregistrer un album dans une boutique de reptiles, mais il y avait un terrible **écho**.

**De:** Ich habe versucht, ein Album in einem Reptiliengeschäft aufzunehmen, aber es gab ein schreckliches **Echo**.

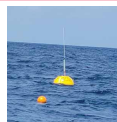
**Kr:** 나는 파충류 상점에서 앨범을 녹음하려고 했지만, 끔찍한 **에코**가 있었습니다.

**Pun Sentence:** She became a lifeguard at the beach and kept the **buoys** in line.



**buoy**

a floating device drifting with ocean currents



**Fr:** Elle est devenue sauveteuse à la plage et a maintenu les **bouées** en ligne.

**De:** Sie wurde Rettungsschwimmerin am Strand und hielt die **Bojen** in einer Linie.

**Kr:** 그녀는 해변에서 구조요원이 되어 **부표**를 일렬로 유지했습니다.

**boy**

a young male of human



**Fr:** Elle est devenue sauveteuse à la plage et a gardé les **garçons** en ligne.

**De:** Sie wurde eine Rettungsschwimmerin am Strand und hielt die **Jungen** in Schach.

**Kr:** 그녀는 해변에서 구조요원이 되어 **소년**들을 통제했습니다.

**Pun Sentence:** How do farmers make crop circles? With a **protractor**.



**protractor**

a semicircular tool used for measuring angles



**Fr:** Comment les agriculteurs font-ils des cercles de culture ? Avec un **rapporteur**.

**De:** Wie machen Bauern Kornkreise? Mit einem **Winkelmesser**.

**Kr:** 농부들은 어떻게 농작물 원을 만드는가? **각도계**를 이용하여.

**tractor**

a motor vehicle used on farms for hauling equipment



**Fr:** Comment les agriculteurs font-ils des cercles de cultures ? Avec un **tracteur**.

**De:** Wie machen Bauern Kornkreise? Mit einem **Traktor**.

**Kr:** 농부들은 어떻게 농작물 원을 만드나요? **트랙터**로.

Figure 7: Example annotations of heterographic puns in 🍌 UNPIE benchmark.