Déjà Vu of Strange Stickers! Enhancing Out-of-Distribution Robustness in Sticker Retrieval via Cross-Modal Intent Alignment

Anonymous ACL submission

Abstract

The rapid evolution of digital communication has amplified the demand for sticker retrieval systems that can match vivid stickers as carriers to satisfy the user's expressive needs. However, real-world sticker retrieval faces significant out-of-distribution (OOD) challenges from unseen queries and stickers, due to the diverse user expression habits and sticker visual representations. The OOD issues often result in the retrieval of irrelevant or inappropriate stickers, negatively impacting the user experience. Inspired by symbolic interactionism in cognition, this paper proposes XAlign-SR to improve OOD robustness in sticker retrieval by aligning abstract expressive intent between queries and stickers across different modalities. We construct two OOD sticker retrieval benchmarks that simulate real-world OOD queries and sticker scenarios. Both online and offline experiments demonstrate that our approach significantly outperforms prevailing baselines [1].

1 Introduction

The popularization of instant messaging and social media platforms has cemented stickers as indispensable tools for digital communication. During online chat, stickers serve as vivid visual elements that enhance conversational dynamics by transcending linguistic and cultural barriers (Tang and Hew, 2019). Effective sticker retrieval systems, which match user queries with stickers that align with their intended expression from a massive repository, are crucial for meeting users' evolving expressive needs in real-world applications.

The out-of-distribution challenge in sticker retrieval. As one of the most popular instant messaging platforms (Datareportal, 2024; Tencent, 2023), WeChat is a representative yet challenging application scenario of sticker retrieval. During our investigation of sticker retrieval in WeChat, we found that

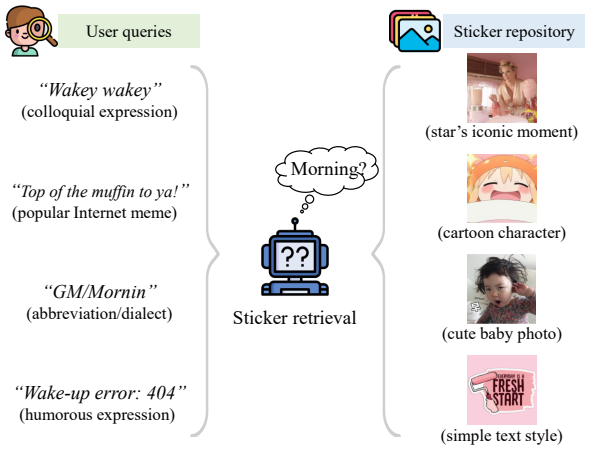

Figure 1: The sticker retrieval is challenged by the diverse query expressions and various sticker presentation formats, even under the same expressive intent.

there are specific characteristics of the query side and the sticker side as shown in Figure 1: (i) *User queries are expressed in various ways.* Queries are usually short and casually expressed, leading to significant query diversity. This variation arises from personal expression, cultural background, and even typing habits. For instance, user queries for 'good morning" may appear in over 28 variations, including synonyms, abbreviations, typos, and incomplete phrases, "morrrning," "sunshine," "gm," "fresh day," or even playful phrases like "wakey wakey." (ii) *Stickers with the same expression intent can vary significantly.* This variation arises due to differences in artistic design, cultural influences, and content. For instance, stickers expressing happiness involve over 169 characters and 18 styles, ranging from simple emoji, artistic text to laughing animated characters, or even meme-inspired humorous images. Additionally, the evolution of internet culture constantly introduces novel stickers that users may like, creating an ever-expanding and diverse sticker repository.

These characteristics collectively present out-of-distribution (OOD) robustness challenges in sticker retrieval during training, (i) *OOD queries* refer to unseen query expressions; and (ii) *OOD stickers*

[1]Our code is available at https://anonymous.4open.science/r/Sticker-OOD-DF54.

refer to unseen or newly popular sticker contents and styles. The inherent differences between query and sticker modalities, combined with the diverse expressions of queries and stickers, make the relevance pattern difficult to learn naturally, resulting in retrieving plenty of low-quality, irrelevant stickers in practical sticker retrieval. We believe that improving OOD robustness in sticker retrieval is essential for deploying an effective and reliable system, ultimately enhancing the user experience.

Using OOD retrieval solutions for sticker retrieval. Yet little effort has been directed toward addressing OOD challenges in sticker retrieval. The most related work in this direction has focused on OOD issues in text retrieval (Thakur et al., 2021; Yu et al., 2022; Fang et al., 2024), where the key idea is to enhance a model's fine-grained matching ability between queries and documents, allowing it to ignore spurious correlations and focus on truly relevant features. However, such approaches are not directly applicable to sticker retrieval, where relevance depends on capturing abstract expressive intent rather than fine-grained textual matching. As queries and stickers with the same intent can take diverse forms, strict fine-grained matching may introduce noise. Our experiments show that applying state-of-the-art OOD documents (Jeronymo et al., 2023; Yu et al., 2022) or OOD queries (Zhuang and Zuccon, 2021) methods from text retrieval yields only marginal gains in sticker retrieval (Section 6.1), as they struggle to capture the global, abstract intent alignment between queries and stickers.

Our approach: Cross-modal intent alignment. In this paper, we propose a novel training method for sticker retrieval called XAlign-SR. Our goal is to mitigate the challenge of OOD queries and stickers by identifying the core expressive intent behind various query and sticker expressions, and aligning their cross-modal expressive intent during training. This approach is inspired by the cognitive science concept of symbolic interactionism (Blumer, 1986), which posits that *humans abstract expressive intent by interactively aligning cross-modal expressive symbols (such as queries and stickers) in the brain*.

XAlign-SR trains a text intent encoder for queries and sticker text, and an image intent encoder for sticker images, following three key steps: (i) Text-focused intent understanding, uses chain-of-thought (CoT) reasoning to capture query and sticker intent, generating diverse expressions for contrastive learning. (ii) Image-focused intent understanding, jointly trains the image and text encoders to align visual expressions of similar intents across stickers. (iii) Cross-modal intent alignment, enhances relevance by aligning query and sticker representations across modalities. Model training combines unsupervised and semi-supervised learning to maximize data efficiency.

Experiments and contributions. We introduce the first two OOD benchmarks for sticker retrieval, $\text{WeChat}_{\text{OOD}}$ and $\text{Sticker820K}_{\text{OOD}}$. Experiments demonstrate that our method outperforms state-of-the-art sticker retrieval models as well as OOD-robust retrieval baselines. Online tests on WeChat further validate its practical efficacy, showing a 13% increase in user preference scores. To the best of our knowledge, we are the first to address the underexplored challenge of OOD sticker retrieval and introduce corresponding benchmarks.

2 Related Work

Cross-modal retrieval. Cross-modal retrieval refers to retrieving relevant data across different modalities (e.g., text, image, video) based on a query from one modality (Wang et al., 2025). Text-to-image retrieval bridges the semantic gap between text queries and images (Ray et al., 2024; Datta et al., 2008). Existing approaches focus on object recognition (Zhang et al., 2024), object relationship reasoning (Pham et al., 2024), and enhancing detail discernment via denoising or adversarial learning (Sarafianos et al., 2019; Tan et al., 2021; Long et al., 2025). However, in sticker retrieval, sticker images often abstractly convey certain concepts or emotions (Tang and Hew, 2019). Traditional image retrieval methods with detailed object recognition and analysis could be susceptible to the highly diverse ways stickers express meaning.

Sticker retrieval. The sticker retrieval model has become a crucial component of instant messaging applications, as it enables users to find stickers that match their expressive needs through queries (Zhao et al., 2023). For example, Zhao et al. (2023) first adopted CLIP (Radford et al., 2021) to capture the visual and textual features of stickers; and Int-RA (Liang et al., 2024) comprises a relation-aware method to retrieve stickers. Existing studies primarily evaluate retrieval performance under the independently and identically distributed (IID) scenario. In this paper, we focus on a more realistic and challenging OOD scenario—enhancing the ro-

bustness of sticker retrieval under OOD data.

OOD robustness of retrieval model. Due to the inherent flaws of deep neural networks, neural retrieval models have been shown to exhibit vulnerability when dealing with unseen data (Thakur et al., 2021; Liu et al., 2023; Song et al., 2024). Studies have found that despite their remarkable IID performance, neural retrieval models could fall short in OOD robustness (Thakur et al., 2021; Chen et al., 2023; Petroni et al., 2021). Existing studies primarily enhance OOD robustness through data augmentation (Bonifacio et al., 2022; Zhuang and Zuccon, 2022), distributionally robust optimization (Yu et al., 2022), or domain-invariant projection (Xin et al., 2022). Essentially, these methods guide retrieval models to focus on fine-grained matching signals, thereby reducing the impact of OOD content variations. Sticker retrieval presents a typical scenario for OOD challenges. However, existing OOD enhancement methods are not directly applicable, as OOD sticker retrieval primarily focuses on the abstract expressive intent of queries and stickers rather than fine-grained matching signals.

3 Problem Statement

Task description. Given a textual query q, the aim of sticker retrieval is to return a ranked list $\mathcal{R}$ of top-K relevant stickers from a large sticker repository $\mathcal{S} = \{s_1, s_2, \ldots, s_N\}$ with a total of N stickers, prioritizing the more relevant the closer to the top. In general, each sticker s consists of text t and image v, where the text includes the caption and style category, etc. The sticker retrieval model f should produce a relevance score $\mathrm{Rel}\,(q, s)$ of the query q for each sticker s in $\mathcal{S}$, based on the sticker text t and image v. Then a truncated ranked list $\mathcal{R}$ is recalled by selecting the top-K stickers with the highest relevance scores, where $K \ll N$.

OOD robustness in sticker retrieval. In retrieval tasks, the OOD robustness refers to a retrieval model's ability to generalize and maintain ranking performance when encountering unseen data (Thakur et al., 2021; Liu et al., 2024, 2023). Formally, given a retrieval model $f_{\mathcal{D}_{\text{train}}}$ trained on the original training data $\mathcal{D}_{\text{train}}$, its OOD robustness is derived from its ranking performance R_M under unseen test data $\mathcal{D}^*_{\text{test}}$ with metric M:

$$\text{Robustness}_{\text{OOD}} = R_M\left(f_{\mathcal{D}_{\text{train}}}; \mathcal{D}^*_{\text{test}}\right). \quad (1)$$

In sticker retrieval, due to the diverse and abstract expression of queries and stickers, the unseen test data $\mathcal{D}^*_{\text{test}}$ is inherently divided into two types: (i) *OOD queries* refer to unseen query variations for models of the same expression intent, arising from typos, language habits, or memes et al; and (ii) *OOD stickers* refer to unseen expressions of the same intent, including strange or newly emerging content, style et al. In this paper, we propose a method to simultaneously address the challenges posed by OOD queries and stickers, while evaluating robustness in two OOD scenarios separately.

Benchmark construction. We construct two benchmarks for OOD evaluation, i.e., WeChat$_{\text{OOD}}$ and Sticker820K$_{\text{OOD}}$, based on sticker retrieval datasets, WeChat (WeChat, 2024) and Sticker820K (Zhao et al., 2023), respectively. The detailed information is shown in Appendix A.1.

Discussion. The OOD challenges in sticker retrieval differ from traditional text/image retrieval (Guo et al., 2022; Zhao et al., 2022; Wang et al., 2019; Cao et al., 2022). Traditional retrieval focuses on matching key details to meet informational needs, with OOD issues often mitigated by keyword/object matching (Thakur et al., 2021). In contrast, sticker retrieval involves diverse visual factors and casual, expressive context, increasing OOD challenges. It prioritizes capturing query intent over literal content, as visual elements of stickers often can not directly serve as a reliable relevance criterion. To address this, methods must effectively capture the abstract intent behind both the user's query and the stickers' expressive nature.

4 Our Method

We introduce XAlign-SR, enhancing sticker retrieval robustness under OOD queries and stickers.

4.1 Overview

The metadata of stickers includes both image v and text t, so we use separate encoders to process them independently. Additionally, the text encoder also handles queries. As illustrated in Figure 2, the XAlign-SR framework contains three key components: (i) Text-focused intent encoder, which captures the expressive intent behind the query and the sticker text separately. (ii) Image-focused intent encoder, which identifies the visual expressive intent of stickers. (iii) Cross-modal intent alignment, which achieves relevance matching between queries and stickers by interactively aligning their intents. Besides, we introduce unsupervised warm-up for text & image encoders, and semi-supervised training for cross-modal intent alignment.

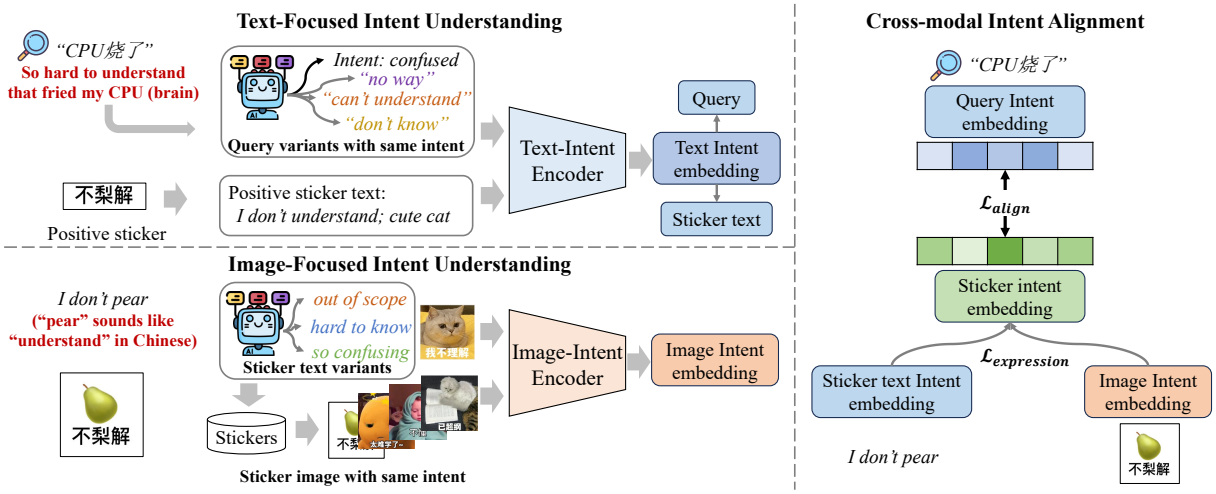

Figure 2: The framework of XAlign-SR that understands and aligns the expressive intent of queries and stickers.

4.2 Text-Focused Intent Understanding

The goal of this stage is to address OOD issues arising from text modalities by accurately understanding the expressed intent of both the query and sticker text. We start by warming up the text-intent encoder (see 4.5) and then further refine it to enhance its intent identification capabilities. Specifically, we first expand the short query and then train the text encoder using a contrastive loss.

CoT-based query expansion. As mentioned earlier, queries in sticker retrieval tend to be shorter and more casual compared to those in typical retrieval tasks. To bridge this gap, we leverage large language models (LLMs) to interpret the underlying expressive intent behind user queries and enrich the query content with alternative expressions.

Given a query q, which may represent either a specific expression or a direct intent: (i) We first prompt the LLM to analyze the query. If the query contains a specific expression, the LLM extracts its underlying intent i_q. Queries that directly express intent proceed to the next step. (ii) Based on the identified intent i_q, we adopt a chain-of-thought (CoT) approach (Wang et al., 2023; Wei et al., 2022) to generate alternative expressions of the query in a step-by-step manner. At each step, the LLM generates a new query expression depending on query intent and previously generated expressions. After k steps we obtain the alternative expressions $\{e_1, e_2, ..., e_k\}$ that preserve intent while varying in linguistic form. (iii) Finally, the original query is then augmented with the identified intent and generated expressions, forming an enriched query $q' = [q; i_q; e_1; ...; e_k]$. The prompt templates used in this process are detailed in Appendix A.2.1.

Text-intent encoder. An effective text encoder should cluster different textual expressions, including queries and sticker text, that share the same intent in semantic space. For each expanded query and its corresponding positive sticker pair (q', s^+), where sticker $s^+ = (t^+, v^+)$ contains sticker text and image: (i) We first use LLMs to generate n variant textual expressions $\{t_1^+, ..., t_n^+\}$ for the original sticker text t^+ (prompt templates are provided in Appendix A.2.2); (ii) Then, we sample m random sticker texts $\{t_1^-, ..., t_m^-\}$ as negative examples; and (iii) Finally, we train the text encoder to capture the expressive intent of queries and sticker text within the output text embeddings using contrastive loss:

$$\mathcal{L}_{\text{text}} = -\log \frac{\sum_{i \in n} \exp(\mathbf{q}^\top \mathbf{t}_i^+)}{\sum_{i \in n} \exp(\mathbf{q}^\top \mathbf{t}_i^+) + \sum_{j \in m} \exp(\mathbf{q}^\top \mathbf{t}_j^-)}, \tag{2}$$

where $\mathbf{q}$ and $\mathbf{t}$ are intent embeddings of query and sticker text, respectively, generated by text encoder.

4.3 Image-focused Intent Understanding

The aim of this stage is to train an image-intent encoder to mitigate OOD issues arising from diverse visual expressions in sticker images. We begin by warming up the image-intent encoder (see 4.5) and then enhance its intent identification capabilities.

The image-intent encoder should cluster diverse visual representations that share the same intent in semantic space. Given an expanded query and its corresponding positive sticker pair (q', s^+): (i) We match the query-sticker text pair in the training set using the n variant textual sticker expressions $\{t_1^+, ..., t_n^+\}$ generated above; (ii) For each gener-

ated expression t_i^+, we compute the cosine similarity using the text-intent encoder and take the image embedding of the most similar query-sticker text pair as the positive examples $\{\mathbf{v}_1^+, ..., \mathbf{v}_n^+\}$. Additionally, we randomly sample m sticker images $\{\mathbf{v}_1^-, ..., \mathbf{v}_m^-\}$ from the sticker repository as negative examples; and (iii) Finally, we train image encoder to capture the expressive intent of sticker images by optimizing the output image-based intent embeddings with contrastive loss:

$$\mathcal{L}_{\text{img}} = -\log \frac{\sum_{i \in n} \exp(\mathbf{v}^\top \mathbf{v}_i^+)}{\sum_{i \in n} \exp(\mathbf{v}^\top \mathbf{v}_i^+) + \sum_{j \in m} \exp(\mathbf{v}^\top \mathbf{v}_j^-)}, \tag{3}$$

where $\mathbf{v}$ denotes the intent embedding of the sticker image, generated by the image encoder.

4.4 Cross-Model Intent Alignment

Having learned the respective intents of text and images, the next challenge is how to effectively match the expressed intent between the query and the sticker to achieve meaningful relevance interaction. Specifically, we propose cross-modal intent alignment, where query intent embeddings are aligned with sticker intent embeddings that convey the same intent during training. This approach enables the model to learn query and sticker intent embeddings that not only recognize intent but also effectively match it across modalities.

Given an expanded query and its corresponding positive sticker pair (q', s^+), the cross-modal intent alignment is performed as follows: (i) *Extracting sticker embeddings*: We compute the sticker text embedding $\mathbf{t}_s$ and the sticker image-based intent embedding $\mathbf{v}_s$ using the text and image encoders:

$$\mathbf{t}_s = f_T(t), \quad \mathbf{v}_s = f_I(v) \tag{4}$$

where f_T and f_I represent the text and image encoders, respectively. (ii) *Integrating sticker representations:* We employ a Transformer (Vaswani et al., 2017) where sticker image-based intent embeddings $\mathbf{v}_s$ serve as values (V) and keys (K), while sticker text-based intent embeddings $\mathbf{t}_s$ serve as queries (Q), producing an integrated sticker intent embedding $\mathbf{h}_s$:

$$\mathbf{h}_s = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V = \text{softmax}\left(\frac{\mathbf{t}_s \mathbf{v}_s^\top}{\sqrt{d}}\right)\mathbf{v}_s, \tag{5}$$

where d is the embedding dimension and $\sqrt{d}$ is scaling factor. (iii) *Sticker expression Loss*: To ensure consistency between the text-based intent embedding and the integrated sticker intent embedding, we minimize the sticker expression loss:

$$\mathcal{L}_{\text{expression}} = \|\mathbf{t}_s - \mathbf{h}_s\|. \tag{6}$$

(iv) *Cross-modal alignment learning*: To align the expression intent of the query and sticker while distinguishing them from unrelated stickers, we employ cross-modal intent alignment loss with randomly sampled negative stickers:

$$\mathcal{L}_{\text{align}} = -\log \frac{\exp(\text{sim}(\mathbf{q}, \mathbf{h}_s))}{\sum_{s \in \mathcal{B}} \exp(\text{sim}(\mathbf{q}, \mathbf{h}_s^-))}, \tag{7}$$

where $\mathcal{B}$ is the training batch, $\mathbf{h}_s^-$ is negative sticker embedding, and $\text{sim}(\cdot)$ is the cosine similarity.

4.5 Training

The training process consists of three phases: unsupervised warm-up, semi-supervised joint intent learning, and intent alignment learning.

Unsupervised warm up. For the text and image intent encoders, we begin with an unsupervised training method to initialize both encoders simultaneously, ensuring they possess initial text-image matching capabilities. For each sticker s, we randomly sample one type of sticker metadata (i.e., captions, OCR text, IP tags) t'_s as query, and calculate the initialize loss with its image v_s:

$$\mathcal{L}_{\text{init}} = -\log \frac{\exp(\text{sim}(\mathbf{t'}_s, \mathbf{v}_s))}{\sum_{s \in \mathcal{B}_w} \exp(\text{sim}(\mathbf{t'}_s, \mathbf{v}_s))} \tag{8}$$

where $\mathcal{B}_w$ represents the warm-up batch and $\mathbf{t'}_s$ is the embedding of the sampled sticker metadata.

Semi-supervised joint intent learning. After initializing the encoders, we propose jointly optimizing the intent understanding capabilities of the text-intent encoder (Equation 2) and the image-intent encoder (Equation 3). We dynamically balance the optimization process between the two encoders using adaptive weighting:

$$\mathcal{L}_{\text{joint}} = \alpha \mathcal{L}_{\text{text}} + (1 - \alpha)\mathcal{L}_{\text{img}} \tag{9}$$

where $\alpha = \frac{\mathcal{L}_{\text{text}}}{\mathcal{L}_{\text{text}} + \mathcal{L}_{\text{img}}}$ is the adaptive weight. During training, we generate pseudo-labels using LLM and text similarity matching, enabling semi-supervised learning for the training data.

Intent alignment learning. Once the text and image encoders have acquired intent understanding capabilities, we perform cross-modal intent alignment, combining the sticker expression loss (Equation 6) and cross-modal intent alignment loss (Equation 7) for overall training:

$$\mathcal{L}_{\text{total}} = \gamma \mathcal{L}_{\text{expression}} + \mathcal{L}_{\text{align}} \tag{10}$$

where γ is a hyperparameter that controls the weight given to the sticker expression.

5 Experimental Setting

5.1 Evaluation

Evaluation scenario. For each benchmark dataset, we evaluate the retrieval performance across three scenarios, i.e., the *original scenario*, *OOD queries*, and *OOD stickers*, which correspond to the original retrieval effectiveness and the OOD robustness when faced with unseen queries and stickers. The model is trained under the training set that is IID with the original scenario.

Evaluation metrics. In retrieval tasks, OOD robustness is measured by the retrieval performance under OOD data (Thakur et al., 2021; Liu et al., 2023). In this paper, we adopt two metrics for each scenario: (i) Mean reciprocal rank (MRR@K) measures how high the relevant stickers rank within the top-K result (Ma et al., 2021; Metilda et al.); and (ii) Recall@K measures the proportion of relevant stickers that exist in the top-K result (Ma et al., 2022; Guo et al., 2022). For ease of observation, the above metrics are presented as percentages.

5.2 Baseline methods

Baselines. We compare our method with several representative approaches, including regular sticker retrieval models, retrieval methods tackling OOD queries and OOD documents (stickers, in this paper), respectively, from text & image retrieval:

- For regular sticker retrieval models, we adopt: BM25 (Robertson and Walker, 1994), StickerCILP (Zhao et al., 2023), StickerLLM (Zhao et al., 2023), and Int-RA (Liang et al., 2024).
- For retrieval methods tackling OOD queries, we adopt: DRTA (Zhuang and Zuccon, 2021), DST (Tasawong et al., 2023), PlugIR (Lee et al., 2024).
- For retrieval methods tackling OOD stickers, we adopt: Inpars (Jeronymo et al., 2023), COCO-DR (Yu et al., 2022), and DAR (Long et al., 2025).

Variants of XAlign-SR. We also implement three variants of XAlign-SR for ablation studies to validate the effectiveness of different components: (i) XAlign-SR$_{\text{-Text}}$, which removes the text-focused intent understanding and only uses the original query-sticker pair to train the text-intent encoder in Section 4.2; (ii) XAlign-SR$_{\text{-Image}}$, which removes the image-focused intent understanding and only uses the original query-sticker pair to train the image-intent encoder in Section 4.2; and (iii) XAlign-SR$_{\text{-Align}}$, which removes the cross–model intent alignment step.

5.3 Implementation details

For the backbone model, following (Zhao et al., 2023; Metilda et al.), to facilitate comparison, both XAlign-SR and the baseline methods use Chinese-CLIP (Yang et al., 2022) as the text & image encoder. For the original scenario, we randomly extract 80% of the annotated query-sticker pairs for training and reserve the remaining 20% for testing. For sticker text, we directly contact the text fields of the sticker with the format of *"Caption:, Emotion:, Style:, IP:, OCR:"*. For the LLM applied in this paper, employ the *gpt-4-turbo* API provided by OpenAI (OpenAI, 2024). For XAlign-SR, we set the query variant number $k = 8$, the sticker text variant number $n = 7$, and the negative example number $m = 56$. During training, the loss hyperparameters α and γ are set to 0.6 and 0.01, respectively. We train the model with a batch size of 64, maximum sequence length of 256, and learning rate of 1e-5. We repeated our experiment 3 times on 4 $\times$ Tesla V100 32G to get the average results.

6 Experimental Result

6.1 Main result

Table 1 shows the comparison of XAlign-SR and baselines across original (IID) scenario and OOD scenarios, including OOD queries and stickers.

Comparison in original scenario. From the performance in the original scenario, we can observe that (i) The models' retrieval performance on WeChat$_{\text{OOD}}$ is generally lower than on Sticker820K$_{\text{OOD}}$, as WeChat$_{\text{OOD}}$ queries come from real online users, introducing more randomness. This also highlights that practical sticker retrieval is a particularly challenging scenario; (ii) The methods tailored for sticker retrieval (like StickerCILP and Int-RA) perform relatively well in the original scenario, and training approaches that enhance OOD robustness can slightly improve performance on top of them. This reveals that sticker retrieval itself is a highly dynamic scenario, where data augmentation and robustness optimization can help alleviate noise and uncertainty in the test set to some extent; and (iii) XAlign-SR performs best among all the methods because it understands the expressive intent of the user query and matches stickers that demonstrate the corresponding intent, thus excelling in sticker retrieval scenarios.

Comparison in OOD scenarios. When we look at retrieval performance in OOD scenarios, we can

Dataset & method	Original Scenario			OOD Queries			OOD Stickers		
WeChat$_{\text{OOD}}$	MRR	Recall		MRR	Recall		MRR	Recall	
	@10	@5	@10	@10	@5	@10	@10	@5	@10
BM25	19.1	20.5	23.3	8.3	9.6	12.3	16.7	17.9	20.9
StickerCILP	23.5	25.8	29.3	11.2	13.0	16.5	8.2	9.1	11.2
StickerLLM	25.7	27.3	31.2	16.5	18.6	21.3	12.5	14.1	17.9
Int-RA	32.8	34.8	39.9	19.3	20.7	24.6	16.9	18.0	23.3
DRTA	26.3	28.8	32.2	22.6	24.3	31.2	16.4	17.6	22.3
DST	26.5	28.3	32.6	23.2	24.9	30.1	16.2	17.8	22.8
PlugIR	28.6	31.9	36.8	25.2	27.0	32.9	18.7	20.6	24.3
DAR	24.8	26.2	30.6	18.2	19.9	23.4	19.9	21.8	25.6
COCO-DR	25.0	26.9	31.5	20.3	22.1	25.6	23.1	23.5	26.9
InPars	27.5	29.3	32.0	21.7	22.9	26.8	24.6	24.9	28.7
XAlign-SR (Ours)	**35.6***	**39.2***	**43.1***	**28.8***	**32.3***	**36.2***	**27.0***	**31.2***	**35.4***
Sticker820K$_{\text{OOD}}$	MRR	Recall		MRR	Recall		MRR	Recall	
	@10	@5	@10	@10	@5	@10	@10	@5	@10
BM25	38.5	46.7	52.6	20.5	24.8	30.2	33.8	41.7	46.9
StickerCILP	52.1	67.1	72.8	29.2	41.6	49.8	26.3	30.2	39.8
StickerLLM	59.8	73.6	78.4	35.8	44.9	52.1	32.8	36.9	42.3
Int-RA	65.9	82.3	86.3	39.5	47.3	59.0	37.0	42.1	47.7
DRTA	60.8	75.2	81.3	49.8	64.0	67.3	38.4	53.0	60.2
DST	61.3	76.0	82.9	50.3	64.2	68.9	37.8	52.4	59.9
PlugIR	63.2	79.3	85.2	54.8	69.9	74.3	42.9	56.7	64.3
DAR	58.9	71.1	76.9	38.6	47.3	52.3	50.3	59.3	67.2
COCO-DR	60.3	74.3	79.6	41.2	52.8	57.9	53.1	66.3	70.2
InPars	63.0	78.6	84.6	43.8	56.7	62.1	55.6	69.8	74.3
XAlign-SR (Ours)	**70.1***	**89.6***	**92.1***	**60.3***	**77.9***	**82.3***	**58.7***	**74.8***	**80.1***

Table 1: Retrieval performance of XAlign-SR and the baselines across the original, OOD queries, and OOD stickers scenarios on two benchmark datasets; $*$ indicates significant improvements over the best baseline ($p \leq 0.05$).

find that (i) BM25, which has demonstrated strong robustness in OOD text retrieval (Thakur et al., 2021; Petroni et al., 2021), loses its advantage in the OOD sticker retrieval scenario. This suggests that precise text matching is not suitable for capturing the abstract intent-based relevance in stickers; (ii) StickerCILP/LLM, and Int-RA experience a significant drop in OOD scenarios because they are designed for IID setting, making it challenging for them to grasp the underlying expressive intent behind diverse, unseen queries and stickers.

From the performance of baselines tailored for enhancing OOD robustness, we can observe that (i) BM25, which demonstrates strong robustness in OOD text retrieval (Thakur et al., 2021; Petroni et al., 2021), loses its advantage in the OOD sticker retrieval scenario. This suggests that precise text matching is not suitable for capturing the abstract intent-based relevance in stickers; (ii) StickerCILP, StickerLLM, and Int-RA experience a significant drop in performance in OOD scenarios because they only consider IID test data, making it challenging for them to grasp the underlying expressive intent behind diverse, unseen queries and stickers; (iii) Methods designed for OOD queries and stickers improve robustness in their respective scenarios, but perform poorly in the other scenario. This suggests that in sticker retrieval, focusing solely on either the query or sticker modality is insufficient for the model to adapt to various unseen OOD scenarios; and (iv) Leveraging LLMs to address OOD challenges (such as PlugIR and InPars) performs better than similar methods, as the strong reasoning capabilities of LLMs can understand the expressive intent behind unseen queries and stickers.

XAlign-SR outperforms baselines, demonstrating that (i) Existing methods for OOD retrieval are influenced by expressive noise in sticker retrieval

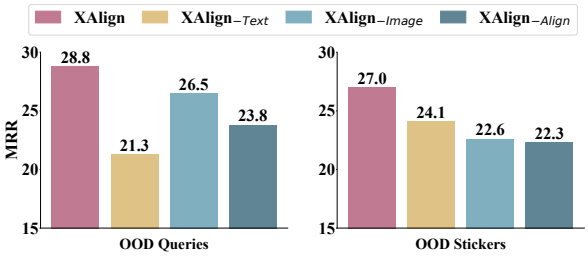

Figure 3: The ablation study of XAlign-SR (denoted as XAlign) in OOD queries and OOD stickers scenarios.

due to their focus on fine-grained matching signals. XAlign-SR, through text/image-focused intent understanding, abstracts the expressive intent behind queries and stickers, enabling intent-level relevance matching; (ii) Classical text/image retrieval methods struggle to bridge the significant expressive gap between the query and sticker modalities, making it difficult to handle OOD scenarios in sticker retrieval. However, XAlign-SR addresses this by aligning cross-modal expressive intents, effectively bridging the differences in the various forms of queries and stickers; and (iii) Jointly optimizing OOD query and sticker robustness allows the model to gain benefits in cross-modal learning, resulting in a model that is robust to both.

6.2 Ablation study

We compare XAlign-SR with three variants: XAlign-SR$_{\text{-Text}}$, XAlign-SR$_{\text{-Image}}$, and XAlign-SR$_{\text{-Align}}$ to validate the effectiveness of different components. The MRR performance in the original scenario of XAlign-SR, XAlign-SR$_{\text{-Text}}$, XAlign-SR$_{\text{-Image}}$, and XAlign-SR$_{\text{-Align}}$ is 35.6, 29.8, 32.9, and 27.5, respectively. The performance in OOD queries and stickers of XAlign-SR with its variants is shown in Figure 3. We report the MRR under WeChat$_{\text{OOD}}$, with similar observations on the other dataset and metrics. We find that: (i) After removing intent understanding for the text and image modalities, the OOD performance for the corresponding query and sticker modalities significantly drops. The decline is more pronounced for the text modality, as it contains more information; and (ii) The cross-modal intent alignment component can further enhance OOD performance by building upon text/image intent understanding, achieving joint gains for OOD queries and stickers.

6.3 Online test

To further validate the performance of XAlign-SR in real-world scenarios, we conducted an online test of our method in the sticker search system of WeChat. The ever-evolving sticker repository

Method	MRR	Rele.	Prefer.
Online system	24.3	82.7	28.1
Int-RA	17.3	65.9	17.3
PlugIR	19.9	74.3	21.9
XAlign-SR	**26.6**	**88.3**	**32.7**

Table 2: The online test between XAlign-SR, the online system, and representative baselines.

and the inherent randomness of user queries make the sticker search system a naturally OOD environment compared to the experimental training set. We selected 10 volunteer users and asked them to perform as many diverse daily queries as possible, resulting in 151 valid searches. We use user clicks as ground truth and select a subset of sticker repository for comparison. Then, we ask users to score the overall relevance (from 0 to 100) of the displayed results and their preference for different model results (with the sum totaling 100). The result is shown in Table 2, we find that XAlign-SR outperforms representative baselines in both retrieval accuracy and user satisfaction, even surpassing online systems. This is because it can mitigate the impact of OOD data, returning stickers that better align with the user's expressive intent.

Case study. Example outputs from different methods are provided in Appendix A.4. Through these examples, we observe that when faced with user queries that exhibit strong individuality, baseline methods and the online system may stick to literal meanings and lack familiarity with niche memes. In contrast, XAlign-SR excels at identifying the abstract intent behind query and stickers, effectively aligning the user's expressive intent with the corresponding meme and its derivative sticker content.

7 Conclusion

This paper focused on the critical challenge of OOD robustness in sticker retrieval, driven by two key observations: (i) the vast diversity of user query expressions and (ii) the stylistic heterogeneity of stickers sharing identical intents. We propose XAlign-SR, a cognitive-inspired framework that aligns cross-modal expressive intent to perform robust relevance judgment under unseen data. We contrast two benchmarks for OOD sticker retrieval, including OOD queries and OOD stickers, to enable systematic evaluation of OOD robustness. Offline and online experiments demonstrate XAlign-SR outperforms baselines and online systems in term of OOD scenarios. This validates intent-level alignment as essential for practical sticker retrieval.

8 Limitations

Our work has several limitations to address in future research. (i) First, this paper investigates the OOD issue in sticker retrieval within the context of the first-stage retrieval process. Sticker retrieval operates as a pipeline system, with the first-stage retrieval serving as its foundation and being the most directly impacted by OOD data. While this study focuses on the first-stage retrieval, the reranking phase is also a promising area for future work, where incorporating user characteristics into sticker retrieval could further mitigate OOD challenges; (ii) Then, due to computational resource constraints and for ease of comparison, we adopted the representative cross-modal retrieval model CLIP as our backbone. In internal testing, our approach maintained its advantage even when utilizing advanced close-source multimodal models. In future work, we plan to validate our method on a broader range of cross-modal models, such as BLIP (Li et al., 2022), and even explore its effectiveness on multimodal LLMs (Lu et al., 2024); and (iii) Finally, in this paper, we categorize the OOD scenarios in sticker retrieval into two broad types: OOD queries and OOD stickers. However, in real-world scenarios, OOD phenomena can be more nuanced and diverse. For instance, OOD queries may include searches related to emerging topics, query variations, or multilingual queries, while OOD stickers may involve newly introduced IPs, novel meme formats, different content sources, or animated stickers. In future work, we plan to develop more targeted solutions to address specific OOD challenges within these categories.

9 Ethics Statement

We approach ethics with great care. In this paper, all the models we use are open-source. For datasets, we construct benchmarks based on the open-source dataset, invite volunteers with industry experience to label. We pay our volunteers a salary that is in line with the local pay scale. and ensured that all data in the baseline were desensitized. Additionally, the methods we propose aim to enhance the OOD robustness of sticker retrieval and do not encourage or induce the model to produce any harmful information or leakage of user data.

References

Herbert Blumer. 1986. *Symbolic interactionism: Perspective and method*. Univ of California Press.

Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. Inpars: Data augmentation for information retrieval using large language models. *arXiv preprint arXiv:2202.05144*.

Min Cao, Shiping Li, Juntao Li, Liqiang Nie, and Min Zhang. 2022. Image-text retrieval: A survey on recent research and development. *arXiv preprint arXiv:2203.14713*.

CCIR. 2024. The 30th china conference on information retrieval. https://www.cips-ir.org.cn/CCIR2024/. Accessed: 2024-10-09.

Xuanang Chen, Ben He, Kai Hui, Le Sun, and Yingfei Sun. 2023. Dealing with textual noise for robust and effective bert re-ranking. *IPM*, 60:103135.

Datareportal. 2024. Digital 2024: Global overview report. Accessed: 2025-02-14.

Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. 2008. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys (Csur)*, 40(2):1–60.

Yan Fang, Qingyao Ai, Jingtao Zhan, Yiqun Liu, Xiaolong Wu, and Zhao Cao. 2024. Combining multiple supervision for robust zero-shot dense retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17994–18002.

Jiafeng Guo, Yinqiong Cai, Yixing Fan, Fei Sun, Ruqing Zhang, and Xueqi Cheng. 2022. Semantic models for the first-stage retrieval: A comprehensive review. *TOIS*, 40(4):1–42.

Vitor Jeronymo, Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, Roberto Lotufo, Jakub Zavrel, and Rodrigo Nogueira. 2023. Inpars-v2: Large language models as efficient dataset generators for information retrieval. *arXiv preprint arXiv:2301.01820*.

Saehyung Lee, Sangwon Yu, Junsung Park, Jihun Yi, and Sungroh Yoon. 2024. Interactive text-to-image retrieval with large language models: A plug-and-play approach. *arXiv preprint arXiv:2406.03411*.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR.

Bin Liang, Bingbing Wang, Zhixin Bai, Qiwei Lang, Mingwei Sun, Kaiheng Hou, Lanjun Zhou, Ruifeng Xu, and Kam-Fai Wong. 2024. Reply with sticker: New dataset and model for sticker retrieval. *arXiv preprint arXiv:2403.05427*.

Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Wei Chen, and Xueqi Cheng. 2023. On the robustness of generative retrieval models. In *Gen-IR@SIGIR*.

Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2024. Robust neural information retrieval: An adversarial and out-of-distribution perspective. *arXiv preprint arXiv:2407.06992*.

Zijun Long, Kangheng Liang, Gerardo Aragon-Camarasa, Richard Mccreadie, and Paul Henderson. 2025. Zero-shot interactive text-to-image retrieval via diffusion-augmented representations. *arXiv preprint arXiv:2501.15379*.

Zhuqiang Lu, Zhenfei Yin, Mengwei He, Zhihui Wang, Zicheng Liu, Zhiyong Wang, and Kun Hu. 2024. B-vllm: A vision large language model with balanced spatio-temporal tokens. *arXiv preprint arXiv:2412.09919*.

Xinyu Ma, Jiafeng Guo, Ruqing Zhang, Yixing Fan, Xiang Ji, and Xueqi Cheng. 2021. Prop: Pre-training with representative words prediction for ad-hoc retrieval. In *WSDM*, pages 283–291.

Xinyu Ma, Ruqing Zhang, Jiafeng Guo, Yixing Fan, and Xueqi Cheng. 2022. A contrastive pre-training approach to discriminative autoencoder for dense retrieval. In *CIKM*, pages 4314–4318.

Chee Heng Er Metilda, Jiayin Wang, Zhiqiang Guo, Weizhi Ma, and Min Zhang. Persrv: Personalized sticker retrieval with vision-language model. In *THE WEB CONFERENCE 2025*.

OpenAI. 2024. Openai api. https://openai.com/api/.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. 2021. Kilt: a benchmark for knowledge intensive language tasks. In *NAACL*, pages 2523–2544.

Khoi Pham, Chuong Huynh, Ser-Nam Lim, and Abhinav Shrivastava. 2024. Composing object relations and attributes for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14354–14363.

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.

Arijit Ray, Filip Radenovic, Abhimanyu Dubey, Bryan Plummer, Ranjay Krishna, and Kate Saenko. 2024. Cola: A benchmark for compositional text-to-image retrieval. *Advances in Neural Information Processing Systems*, 36.

Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94*, pages 232–241. Springer.

Nikolaos Sarafianos, Xiang Xu, and Ioannis A Kakadiaris. 2019. Adversarial representation learning for text-to-image matching. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5814–5824.

Ge Song, Kai Huang, Hanwen Su, Fengyi Song, and Ming Yang. 2024. Deep ranking distribution preserving hashing for robust multi-label cross-modal retrieval. *IEEE Transactions on Multimedia*.

Hongchen Tan, Xiuping Liu, Baocai Yin, and Xin Li. 2021. Cross-modal semantic matching generative adversarial networks for text-to-image synthesis. *IEEE Transactions on Multimedia*, 24:832–845.

Ying Tang and Khe Foon Hew. 2019. Emoticon, emoji, and sticker use in computer-mediated communication: A review of theories and research findings. *International journal of communication*, 13:27.

Panuthep Tasawong, Wuttikorn Ponwitayarat, Peerat Limkonchotiwat, Can Udomcharoenchaikit, Ekapol Chuangsuwanich, and Sarana Nutanong. 2023. Typo-robust representation learning for dense retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1106–1115.

Tencent. 2023. Tencent announces 2023 third quarter results. Accessed: 2025-02-14.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *NIPS*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS*, volume 30.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634.

Tianshi Wang, Fengling Li, Lei Zhu, Jingjing Li, Zheng Zhang, and Heng Tao Shen. 2025. Cross-modal retrieval: a systematic review of methods and future directions. *Proceedings of the IEEE*.

Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. 2019. Camp: Cross-modal adaptive message passing for text-image retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5764–5773.

WeChat. 2024. 2024 national information retrieval challenge cup (ccir cup). https://algo.weixin.qq.com/. Accessed: 2024-10-09.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Ji Xin, Chenyan Xiong, Ashwin Srinivasan, Ankita Sharma, Damien Jose, and Paul Bennett. 2022. Zero-shot dense retrieval with momentum adversarial domain invariant representations. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4008–4020.

An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. 2022. Chinese clip: Contrastive vision-language pretraining in chinese. *arXiv preprint arXiv:2211.01335*.

Yue Yu, Chenyan Xiong, Si Sun, Chao Zhang, and Arnold Overwijk. 2022. Coco-dr: Combating distribution shifts in zero-shot dense retrieval with contrastive and distributionally robust learning. *arXiv preprint arXiv:2210.15212*.

Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. In *ICLR*.

Bo-Jian Zhang, Guang-Hai Liu, Zuo-Yong Li, and Shu-Xiang Song. 2024. Locating target regions for image retrieval in an unsupervised manner. *IEEE Transactions on Neural Networks and Learning Systems*.

Sijie Zhao, Yixiao Ge, Zhongang Qi, Lin Song, Xiaohan Ding, Zehua Xie, and Ying Shan. 2023. Sticker820k: Empowering interactive retrieval with stickers. *arXiv preprint arXiv:2306.06870*.

Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. 2022. Dense text retrieval based on pretrained language models: A survey. *arXiv preprint arXiv:2211.14876*.

Shengyao Zhuang and Guido Zuccon. 2021. Dealing with typos for bert-based passage retrieval and ranking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2836–2842.

Shengyao Zhuang and Guido Zuccon. 2022. Characterbert and self-teaching for improving the robustness of dense retrievers on queries with typos. In *SIGIR*.

A Appendix

A.1 Benchmark information

A.1.1 Sticker retrieval datasets

We collect two original sticker retrieval datasets: (i) WeChat (WeChat, 2024) is a public dataset from the sticker retrieval challenge of CCIR Cup 2024 (CCIR, 2024), with 500K stickers spanning 73 different styles and 6,250 intellectual properties (IPs, i.e., characters) from WeChat; (ii) Sticker820K (Zhao et al., 2023) is another public sticker retrieval dataset constructed by other researchers containing around 820K stickers across 8 different styles. For WeChat dataset, the metadata of a sticker includes image, caption, emotion, style, IP and OCR (text recognized in the image). Here, the sticker text t comprises the caption, emotion, style, IP, and OCR and the sticker image is denoted as v. The fields of Sticker820K dataset are similar to WeChat dataset, but lack the IP field. The data examples of WeChat are shown in Figure 4.

A.1.2 Benchmark construction

Since these datasets do not have direct OOD data part, we further construct OOD benchmarks for evaluation by dividing query and sticker data. We construct the benchmarks of WeChat$_{\text{OOD}}$ and Sticker820K$_{\text{OOD}}$ for WeChat and Sticker820K datasets, respectively. The overall statistics are shown in Table 3.

- For OOD queries, we randomly sample 5% of labeled queries from each of the two datasets to generate their variants as OOD queries. Specifically, we invited practitioners with experience working with sticker search to write five variants for each sampled query. These variants simulate differences arising from typos, verbal expression habits, and cultural contexts, but are guaranteed to be relevant to the original stickers. We exclude the stickers associated with the sampled queries from the training set and use them along with the query variants as the test set for the OOD queries.
- For OOD stickers, we divide about 10% of the labeled stickers from each of the two datasets and ensure that types of style and IP of the divided stickers no longer exist in the original dataset. For Sticker820K$_{\text{OOD}}$, since the dataset does not contain an IP field in its metadata, we perform segmentation based only on style. The sampled stickers, along with their associated queries, are excluded from the training set and used as the test set for OOD evaluation.

Scenario	Data	WeChat$_{\text{OOD}}$	Sticker820K$_{\text{OOD}}$
Original	# Queries	11,018	50,235
	# Stickers	450K	800K
	# Pairs	11,519	50,302
OOD Queries	# Queries	2,465	12,530
	# Stickers	500K	800K
OOD Stickers	# Queries	1,243	5,024
	# Stickers	500K	800K

Table 3: The data statistic of WeChat$_{\text{OOD}}$ and Sticker820K$_{\text{OOD}}$. # Pairs denotes query-sticker pairs.

A.2 LLM prompts

A.2.1 Prompt for CoT-based query expansion

The guiding prompt for recognizing the expressed intent is: "`Given a user query for searching stickers, {query}, it may either represent an expressive form of intent or a direct intent expression. For example, "overjoyed to the point of flying" is an expressive form, whereas "happy" is a direct intent expression. First, determine whether the query is an expressive form of intent or a direct intent expression. (i) If it is an expressive form of intent, analyze its underlying core intent and output it directly; (ii) If it is a direct intent expression, output it as is.`"

The prompt for guiding the step-by-step generation of the different expressions of the query is: "`Given a user query for searching stickers, along with its expressive intent, your task is to generate a new variant of the sticker query that aligns with the given intent while ensuring it is distinct from both the original query and any previously generated variants. The new variant should maintain the intended meaning but introduce differentiation wherever possible.`
`Query: {query}`
`Intent: {intent}`
`Pervious variants: {variants}`"

A.2.2 Prompt for generating variant sticker text expressions

The prompt for generating variant sticker text expressions is: "`Given a user query for searching stickers and the associated sticker text (including captions, text within the sticker, characters, and sticker style), the task is to identify the abstract expressive intent behind the query and the sticker text. Based`

User query	Relevant sticker					
	Image	Caption	OCR	IP	Emotion	Style
你醒啦? (Are you awake?)		可爱猫醒了吗 (Is the cute cat awake?)	醒了吗 (You're up?)	动物: 猫 (Animal: Cat)	可爱 (Cute)	萌宠 拍摄 (Cute pets Shooting)
我好孤独 (I'm so lonely)		我的世界只剩下孤独 (The only thing left in my world is loneliness)	我的世界只剩下孤独 (The only thing left in my world is loneliness)	兔斯基: 饿疯兔 (Bugs bunny: Hungry crazy bunny)	日常 (Daily)	绘制表情 卡通形象 简笔画 (Drawing sticker Cartoon Sketches)
梦里啥都有 (It's all in the dream)		波吉开始做梦 (Boogie starts dreaming)	开始做梦 (Start dreaming)	国王排名: 波吉 (King's eanking: Boogie)	日常 (Daily)	动漫人物 绘制彩图 (Anime characters Drawing color)
嗨皮 (Happy (homonym))		有亿点开心 (I'm only a little (billion) happy)	有1点开心 (A little happy)	明星: 成龙(Star: Jackie Chan)	搞笑 (Funny)	真人 男 拍摄 (Real Man Shooting)

Figure 4: Data examples of query and sticker metadata in Wechat. Note that Sticker820K does not have the IP field.

on this understanding, a new variation of the sticker text should be directly generated that aligns with the given expressive intent while ensuring it is distinct from both the original text and previously generated variations. The generated variation should maintain meaningful differentiation while staying true to the intended expression.
Query: {query}
Relevant sticker text: {sticker text}
Pervious variants: {variants}"

A.3 Baseline details

We compare our method with several representative approaches, including regular sticker retrieval models, retrieval methods tackling OOD queries and OOD documents (stickers, in this paper), respectively, from text & image retrieval:

- For regular sticker retrieval models, we adopt: (i) BM25 (Robertson and Walker, 1994) is a classical probabilistic retrieval model that shows effectiveness on OOD retrieval tasks (Thakur et al., 2021). We take sticker text as the document; (ii) StickerCILP (Zhao et al., 2023) directly fine–tunes CLIP model (Radford et al., 2021; Yang et al., 2022) to capture sticker features and text features; (iii) StickerLLM (Zhao et al., 2023) is similar with StickerCILP but uses ChatGLM-6B (Zeng et al.) as the query encoder; and (iv) Int-RA (Liang et al., 2024) matches stickers by understanding the common-sense requirements of the query.
- For retrieval methods that enhance robustness to OOD queries, we adopt: (i) DRTA (Zhuang and Zuccon, 2021) is from text retrieval and uses contrastive learning to bridge the gap between the query and the possible variants; (ii) DST (Tasawong et al., 2023) is from text retrieval and aligns the ranking list between query and variants in a self-training manner; and (iii) PlugIR (Lee et al., 2024) is from image retrieval and leverages LLM reasoning to understand and refine unseen queries.
- For retrieval methods that enhance robustness to OOD stickers, we adopt: (i) Inpars (Jeronymo et al., 2023) is from text retrieval and generates pseudo query for OOD stickers with LLMs; (ii) COCO-DR (Yu et al., 2022) is from text retrieval and uses distributionally robust optimization to learn domain features of unseen stickers; and (iii) DAR (Long et al., 2025) is from image retrieval and uses a diffusion model to generate unseen stickers to assist with retrieval.

A.4 Case study

Figure 5 illustrates an example where a user searches for a popular internet homophonic meme sticker. In this case, the literal meaning of the query is *"blue, thin mushroom"*, but it actually

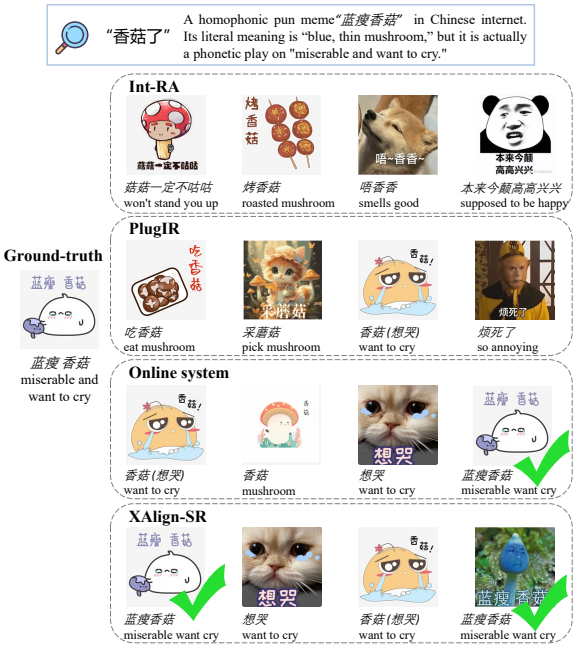

Figure 5: A case study on queries and stickers with homophonic puns memes to express feeling sad and wanting to cry cross baselines and XAlign-SR.

represents a widely recognized meme expressing sadness and the urge to cry. From the results, we observe that baseline methods typically rely on the literal meaning, retrieving multiple stickers related to mushrooms while also being misled by seemingly relevant but ultimately unrelated stickers. Online systems can partially capture the intended meaning of *"feeling sad and wanting to cry"*, but they fail to accurately retrieve stickers associated with the specific meme. In contrast, XAlign-SR effectively understands the user's intent and the expressive meaning embedded within the meme stickers, enabling precise retrieval.

A.5 Necessary Statements

The experiments were conducted on 4 × NVIDIA Tesla V100 32G GPUs. Training an XAlign-SR model takes approximately 12 hours, while evaluation requires about 1.5 hours.

We invited ethical labelers and volunteers and ensured that they received higher than the standard local hourly rate. For online test volunteers, we invited users who were proficient in using the sticker search function and ensured that all information seen by all users was fully desensitized. For the baseline, we ensured that all information was desensitized and did not reveal user or system information.