Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

X-STAR: Cross-lingual Scene Text-aware Image Retrieval

Dikshant Sharma
Indian Institute of Technology Jodhpur
Prateek Prateek
Indian Institute of Technology Jodhpur
Anand Mishra
mishra@iitj.ac.in

Indian Institute of Technology Jodhpur

Research Article

Keywords:

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11752904/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

X-STAR: Cross-Lingual Scene Text-Aware Image Retrieval

Dikshant Sharma, Prateek, Anand Mishra

[1]Indian Institute of Technology Jodhpur, Jodhpur, 342030, Rajasthan, India
https://github.com/vl2g/X-STAR.

*Corresponding author(s). E-mail(s): mishra@iitj.ac.in;

Abstract

Scene text in local languages are shared, searched and consumed across linguistic boundaries. However, most existing scene text-aware retrieval methods overwhelmingly assume a monolingual setting where both scene text and query are in the same language (typically English). In this work, we ask a fundamental question: Can we retrieve images containing scene text in one language using queries in another? Toward answering this, we introduce X-STAR – the first dataset and benchmark for Cross-Lingual Scene Text-Aware Image Retrieval. The dataset contains 6,449 images containing scene text spanning 13 languages and 83,837 queries across the same set of languages, enabling both monolingual and cross-lingual retrieval (i.e., scene text in one language and queries in another). We further propose CoLITE-CLIP – a new architecture that combines visual-textual co-attention over CLIP embeddings with a ColBERT-inspired late-interaction module for fine-grained matching, enabling the model to handle cross-lingual retrieval by reducing multilingual queries and scene text to a shared monolingual space through machine translation, thereby improving retrieval quality. Extensive experiments show that CoLITE-CLIP substantially outperforms many strong baselines in both cross-lingual and monolingual settings, e.g., in X-STAR it improves the base CLIP model by 22.3 points in R@1 (from 24.2% to 46.5%) and outperforms the most competitive baselines. Beyond the cross-lingual setting, CoLITE-CLIP also achieves competitive performance on the monolingual TextCaps benchmark, despite relying on a substantially more compact backbone than several competing approaches. Our findings further show that, in the absence of machine translation, current state-of-the-art models still struggle substantially on cross-lingual scene text-aware retrieval. Our data and code are publicly available at https://github.com/vl2g/X-STAR.

Keywords: Cross-lingual retrieval, Scene text, Image retrieval, Multilingual vision-language models, Late interaction, Low-resource languages

1 Introduction

With the increasing global use of visual media, images containing scene text in local languages are shared, searched, and consumed across linguistic boundaries. However, most existing scene text-aware retrieval systems [1–3] assume monolingual setting where typically, scene text is assumed to be in English, and query is also performed using English. This creates a significant barrier for users who wish to retrieve information across languages. For example, a signboard or a natural scene containing Hindi[1] in an image may contain crucial semantic cues, but a user querying in English or another language may probably ignore visual text content on it. We, therefore, introduce a novel task namely Cross-Lingual Scene

[1]A language from Global South that has approximately 600 million total speakers worldwide.



Fig. 1 Motivation for this work. Queries in different languages refer to the same image containing Hindi scene text (भारतीय वायु सेना), transliterated as "Bharatiya Vayu Sena" and translated as "Indian Air Force." Can we make such images searchable using queries in any language?

Text-Aware Retrieval (X-STAR in short)[2] that aims to bridge this linguistic barrier by jointly understanding the visual scene text, its language context, and the query semantics across multiple languages. As illustrated in Figure 1, it enables users to search images regardless of the language visible in them, thereby enabling multilingual accessibility, cross-cultural information exchange, and inclusive AI systems that truly reflect the linguistic diversity. Please refer to Figure 2 which illustrates the motivation and the goal of our problem.

Early works on image retrieval [4–6] often ignored scene text while retrieving images. One of the earliest works [7] that treated textual cues in images as indispensable signals for image retrieval appeared over a decade ago. Later with the advancements in scene text recognition [8–10], particularly in English, authors have extensively used scene text for image retrieval [1, 2]. Early works in this space have mainly focused on keyword-based retrieval using OCR-extracted tokens [7, 11], for example, searching for images containing the word "*restaurant*" within a large image collection. With subsequent progress in visio-lingual understanding, retrieval methods



Fig. 2 (a) **Monolingual scene text-aware retrieval [1, 2, 7, 12, 13]:** Queries and scene text are both in the same language, typically English. Such retrieval often ignores scene text in other languages. (b) **X-STAR (Ours):** We address cross-lingual scene text-aware retrieval, where the scene text and query can appear in any of the supported languages: English, Chinese, Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Odia, Punjabi, Tamil, and Telugu.

began to combine both textual and visual cues [2, 12], for instance, querying for "*People eating pizza in Joe's restaurant.*", which requires integrating the detected scene text ("Pizza, Joe's") with the visual appearance (people, eating, restaurant). Some recent work has also enabled for searching over the Chinese scene text [13, 14]. However, existing datasets are largely monolingual, limiting progress in cross-lingual scene text-aware retrieval. To address this, we introduce X-STAR, the first dataset built specifically for this task.

There are two plausible approaches to address X-STAR: (i) Represent the query and OCRed text in their original languages and perform semantic matching in a shared multilingual embedding space. Further, visual cues can also be seamlessly integrated to such embedding space using models such as SigLIP 2 [15] and Multilingual CLIP [16]. While multilingual representations are steadily improving; especially for low-resource languages, this approach often falls short in practice. Cross-lingual semantic alignment for short texts is inherently challenging and difficult to fine-tune due to the lack of paired data in low-resource languages. The problem is further exacerbated by the noisy OCR outputs common in these languages. (ii) Reduce the task to a monolingual setting by translating both the query and the OCRed text into

[2]X in X-STAR follows the common NLP naming convention, where it denotes "cross" to signify cross-lingual or multilingual settings.

English, followed by retrieval using strong English retrieval models. This approach benefits from (a) mature and robust English retrieval systems, (b) better semantic consistency in a single language space, and (c) reduced cross-lingual alignment complexity. Moreover, translation can partially normalize OCR noise, leading to improved downstream matching. Although this approach is not without limitations, as it relies heavily on the performance of OCR and machine translation systems, the absence of existing models and large-scale paired training data for this task makes it the most practical choice for developing our framework. To this end, inspired by the seminal work ColBERT [17], we adopt a ColBERT-style late interaction mechanism on top of CLIP-based representations to enable fine-grained token-level matching between translated query and translated OCRed scene text. This design is particularly suitable for X-STAR, as it mitigates OCR noise and short-text brittleness by allowing robust, localized alignment instead of relying on a single global embedding.

Contributions: We make the following contributions in this paper: (i) We introduce X-STAR — a novel task and dataset for Cross-Lingual Scene Text-Aware Retrieval. Although scene text-aware retrieval has previously been extensively studied, to the best of our knowledge, this is the first work to extend it to a cross-lingual setting. (ii) We propose CoLITE-CLIP, which fuses OCR lines and image patches through bidirectional co-attention and scores them with ColBERT-style late interaction. To the best of our knowledge, it is the first model to apply such OCR-aware late interaction to cross-lingual retrieval over natural scene images. (iii) We conduct extensive experiments demonstrating that our method significantly outperforms competitive baselines in scenarios where both the scene text and the query may appear in any of 13 languages including English, Chinese, Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Odia, Punjabi, Tamil, Telugu. Moreover, CoLITE-CLIP achieves strong and competitive performance on the monolingual TextCaps benchmark.

2 Related Work

Scene Text-aware Retrieval: In the last decade, several works have emerged in which scene text is shown as an important cue for broader scene understanding tasks leading to benchmarks such as OCR-VQA [20], TextVQA [21], TextKVQA [22], ST-VQA [3], TextCaps [18]. Earlier research has also confirmed that the text in the scene contains important semantic information for image retrieval [7, 11, 23]. The modeling paradigm has also evolved from traditional architectures to large-scale pretrained models. The large-scale pretrained vision language models, such as CLIP and its variants, e.g., mCLIP [16], metaCLIP [24], openCLIP [25], and SigLIP [26], have emerged as de facto models for sentence-to-image retrieval. However, for scene text–aware retrieval, these models face two primary limitations. First, they are not text-aware, i.e., they fail to explicitly utilize the textual content embedded in the scene. This becomes even worse in low-resource languages. This requires integration of visual signals and OCR into the embeddings. Second, their reliance on global embeddings of images and text often limits fine-grained matching, especially when the query refers to a small localized region within the image. To address a similar challenge in the information retrieval domain, ColBERT [17] was proposed, which employs a late interaction mechanism to maintain multiple contextualized token-level representations instead of a single global embedding. Late interaction has since been adopted for image–text and visual document retrieval: FILIP [27] uses token-wise maximum similarity between image patches and words, ColPali [28] applies ColBERT-style multi-vector late interaction to document page images, and ColMate [29] combines OCR-based pretraining with visual late interaction for multimodal document retrieval. Building on these ideas, we propose CoLITE-CLIP, which integrates visual-textual co-attention between OCR lines and image patches over CLIP embeddings and employs late interaction for fine-grained retrieval in multilingual natural scenes.

Cross-Lingual Information Retrieval (CLIR): CLIR has gained significant attention in recent years [30–32]. Most CLIR methods align multilingual queries and documents within the textual modality. In contrast, our setting requires matching multilingual queries to scene images containing multilingual text, introducing the additional challenges of visual grounding and multilingual scene-text understanding.

Table 1 Comparison of X-STAR with existing scene-text-aware image retrieval datasets. X-STAR uniquely provides multilingual street-view images with cross-lingual sentence-level queries spanning 13 languages.

Dataset	#Imgs	#Langs	#Queries	Multilingual	Query Type
TextCaps [18]	28,408	1 (en)	142,040	✗	Sentence
CTC (StacMR) [2]	10,683	1 (en)	53,415	✗	Sentence
SVT [19]	249	1 (en)	427	✗	Word
Sports-10K [7]	10,000	1 (en)	10	✗	Sentence
TV-series-1M [7]	1M	1 (en)	20	✗	Sentence
IIIT STR [7]	10,000	1 (en)	50	✗	Word
DL-CSVTR [13]	2,070	1 (zh)	89	✗	Word
X-STAR (ours)	**6,449**	**13**	**83,837**	✓	**Sentence**

3 X-STAR Dataset

Scene text-aware image retrieval using English queries and images containing English scene text has been extensively explored [2, 7, 18], with several benchmark datasets available for this task. In contrast, to the best of our knowledge, no existing dataset addresses the cross-lingual setting, where the query is in one language (say, language-A) and the scene text in the image is in another (say, language-B). To bridge this gap, we introduce X-STAR (Cross-lingual Scene Text-Aware Retrieval), a new benchmark comprising 6,449 images with scene text in 13 popular languages from South Asia, including English, Chinese, Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Odia, Punjabi, Tamil, Telugu, and 83,837 multilingual queries across the same set of languages. Some sample images from the dataset are shown in Figure 3. Note that these languages (even if we exclude English) are used by approximately one-third of the global population. We compare X-STAR with other scene text-aware image retrieval datasets in Table 1. We observe that multi-linguality and cross-lingual query support make our dataset unique. X-STAR will be available for download from our project website. We now describe the dataset collection, annotation, and its statistics below.

3.1 Data Collection

We obtain images containing multilingual scene text by combining two publicly available datasets, namely the Bharat Scene Text Dataset [33] and the RCTW-17 Dataset [34]. The Bharat Scene Text Dataset is a highly challenging multilingual benchmark that includes scene text images in 11 widely used Indian languages, namely Assamese, Bengali, Gujarati, Hindi, Odia, Marathi, Kannada, Malayalam, Punjabi, Tamil, and Telugu along with English. The RCTW-17 dataset, on the other hand, contains scene text in Chinese. Together, these datasets yield 6,449 images in all. Each of these images also comes with *manual annotations* of the text they contain. To adapt this dataset, for the cross-lingual scene text-aware retrieval task, we further generate corresponding text queries using a Large Language Model (LLM), as described next.

3.2 Query Generation and Multilingual Expansion

Since obtaining queries from language experts across 13 languages requires a significant budget, we first translate all the ground truth scene text into English using state-of-the-art machine translators i.e. IndicTrans [35] for translating Indian languages and Google Translate [36] for translating Chinese. We then use this translated text in English and the image to prompt a state-of-the-art LLM to generate a query that uses the scene text as a cue. The prompt used to generate the queries with the LLM is shown in the *LLM Prompt for Query Generation* box. To be specific, we use GPT-4V [37] as LLM to obtain the queries. Once queries are obtained, they are translated back to the respective languages. In addition, we collect **user-written queries** for one language (i.e. Hindi) by hiring three language experts and instructing them to write scene-text-aware queries. This process yields 3,642 additional



GroundTruth:
శ్రీపోతన చైతన్య వేదిక, విశ్వనాధ భవనం, గూడూరు, పాలకుర్తి మండలం, జనగామ జిల్లా 506 252
Scene Text Language: Telugu

English: Sripothana Chaitanya Vedika Telugu signboard
Assamese: শ্ৰীপোথান চৈতন্য বেদিকা তেলেগু ছাইনব'ৰ্ড
Bengali: শ্রীপোথান চৈতন্য বেদিকা তেলেগু সাইনবোর্ড
Gujarati: શ્રીપોથના ચૈતન્ય વેદિકા તેલુગુ સાઇનબોર્ડ
Hindi: श्रीपोथना चैतन्य वेदिका तेलुगु साइनबोर्ड
Kannada: ಶ್ರೀಪೋತನ ಚೈತನ್ಯ ವೇದಿಕಾ ತೆಲುಗು ಫಲಕ
Malayalam: ശ്രീപോതന ചൈതന്യ വേദിക തെലുങ്ക് സൈൻബോർഡ്
Marathi: श्रीपोथना चैतन्य वेदिका तेलगू फलक
Odia: ଶ୍ରୀପୋଥାନ ଚୈତନ୍ୟ ବେଦିକା ତେଲୁଗୁ ସାଇନବୋର୍ଡ
Punjabi: ਸ਼੍ਰੀਪੋਥਾਨਾ ਚੈਤੰਨਿਆ ਵੇਦਿਕਾ ਤੇਲਗੂ ਸਾਈਨ ਬੋਰਡ
Tamil: ஸ்ரீபோத்தனா சைதன்யா வேதிகா தெலுங்கு அடையாளப் பலகை
Telugu: శ్రీపోతన చైతన్య వేదిక తెలుగు సైన్బోర్డ్
Chinese: 斯里波塔纳·柴坦尼亚·维迪卡 泰卢固语招牌



Ground Truth:
悟空烧烤, Wokoo BBQ, 小食汇, 用心烤好每一串, 餐粥烧烤小海鲜生烫牛杂, 特色美食
Scene Text Language: Chinese

English: Wokoo BBQ storefront sign red facade
Assamese: ৱোকু বি. বি. কিউ. ষ্ট ৰ ফ্ৰন্টত ৰঙা মুখমণ্ডলত চিহ্ন দিয়া হৈছে
Bengali: ওকু বি. বি. কিউ স্টোরফ্রন্ট সাইন লাল সম্মুখভাগ
Gujarati: વોકૂ બી. બી. ક્યૂ સ્ટોરફ્રન્ટ સાઇન લાલ રવેશ
Hindi: वोकू बी. बी. क्यू. स्टोरफ्रंट साइन लाल मुखौटा
Kannada: ವೊಕ್ಕೂ ಬಿಬಿಕ್ಯೂ ಅಂಗಡಿಯ ಮುಂಭಾಗವು ಕೆಂಪು ಮುಂಭಾಗವನ್ನು ಸೂಚಿಸುತ್ತದೆ
Malayalam: വോക്കൂ ബിബിക്യു സ്റ്റോർ ഫ്രണ്ട് സൈൻ റെഡ് ഫെയിസ്
Marathi: वोकू बी. बी. क्यू. स्टोअरफ्रंटचा दर्शनी भाग लाल रंगाचा
Odia: ୱୋକୁ ବି.ବି.କ୍ୟୁ. ଷ୍ଟୋରଫ୍ରଣ୍ଟ ସାଇନ୍ ଲାଲ ସାମ୍ନା
Punjabi: ਵੋਕੂ ਬੀ. ਬੀ. ਕਿਊ. ਸਟੋਰਫ੍ਰੰਟ'ਤੇ ਲਾਲ ਰੰਗ ਦਾ ਨਿਸ਼ਾਨ
Tamil: வோக்கூ பி. பி. கே கடையின் முன்புறத்தில் சிவப்பு முகப்பில் கையெழுத்திடுங்கள்
Telugu: వోకూ బి. బి. క్యూ స్టోర్ ఫ్రంట్ సైన్ ఎరుపు ముఖభాగం
Chinese: 沃库烧烤店面标志红色门面

Fig. 3 Two sample images from X-STAR with their associated queries in 13 languages. The top image contains scene text in Telugu (a language spoken in South India), whereas the bottom image contains Chinese scene text.

queries, which we use separately to evaluate the proposed model.

Since the LLM is prompted with manually annotated ground-truth OCR tokens along with the corresponding images, it generates semantically grounded, scene-text-aware queries. Although translation into 13 languages may introduce minor grammatical errors or artifacts, our data-quality evaluation (Section 3.4) confirms that the generated multilingual queries remain high quality.

3.3 Dataset Statistics

Our corpus contains 6,449 street-view images. We have 83,837 natural language queries using 13 languages, making the dataset rigorous for testing cross-lingual search. The average query length for X-STAR is 6 words. In addition, most of these scene images are rich in scene text. As reported in Appendix B.4, across all splits, about 90% of the queries include at least one scene-text term, and roughly 66% contain two or more. The average OCR token on this dataset is 6, with 68.3% non-English scene text words. We randomly split the dataset into train, test and validation, containing (4,515 images, 58,695 queries) in the train set, (645

LLM Prompt for Query Generation

You must output exactly **ONE** English search query on a single line (no quotes or labels).

Goal: Retrieve the image by prioritizing distinctive scene text.

Inputs: *Image* and ground truth scene-text tokens: {scene_text_en}

Instructions:

1. Use 5–9 words as a concise noun phrase.
2. Begin with the most distinctive scene-text n-gram whenever possible (e.g., a proper noun, brand, place, or event).
3. Add at most one or two visual or contextual anchors, such as storefront sign, roadside billboard, night street, or indoor counter.
4. Prefer uniqueness over completeness and ignore minor objects.
5. Use nouns and adjectives only; avoid verbs and stopwords.
6. Include at most one brand or place name.
7. Avoid punctuation, except one comma when helpful.
8. When reliable scene text is unavailable, use a distinctive visual anchor together with the object or scene category.
9. Avoid leading phrases such as "photo of," "image showing," or "scene of."

Return only the query line.

images, 8,385 queries) in the validation set, and the remaining (1,289 images, 16,757 queries) in test set. The annotation and validation protocols, along with further dataset statistics, are provided in Appendix B.

3.4 Data Quality Evaluation

Although the scene texts in our dataset are manually annotated, we rely on GPT-4V and off-the-shelf translation systems to generate natural language queries and their translation into 12 languages. Consequently, ensuring the quality and reliability of this automatically generated content is crucial. To rigorously assess data quality, we conduct both user study and an unsupervised automatic evaluations, as described below.

3.4.1 User Study

We conducted a user study on 500 randomly sampled image–query pairs from our dataset. The evaluation was carried out across 10 languages, with quality assessments performed by hired native speakers of the respective languages.[3] Annotators were shown image and query in their respective language and asked the following five questions: (i) *Is the query relevant to the image?* (Options: a. Relevant, b. Partially Relevant, c. Not relevant). (ii) *Is the query grammatically correct?* (Options: a. Correct, b. Partially correct, c. Incorrect). (iii) *Does the query contain any OCR token?* (Options: a. Yes, b. No.) (iv) Does the query contain visual cues? (Options: a. Yes, b. No.) (v) Rate the query in terms of realism? (Options: a. 1: worst, b. 2: poor, c. 3: average, d. 4: good, e. 5: best).

The evaluation indicates that 90.5% of the queries are either partially or fully relevant to the image. Further, 52% of the queries are rated grammatically correct and 37.2% contain only minor imperfections, reflecting natural variability in real-world user queries. User study also confirms that 94.6% of the images contain relevant scene text and 88.6% of the samples have a strong visual cues compatible with the query. The average score for perceived realism of the queries is 4.36/5, which emphasizes the overall high quality of the dataset.

3.4.2 Unsupervised Automatic Quality Evaluation

In addition to the user study, we also perform an unsupervised automatic evaluation of translation quality using cyclic self-consistency. Specifically,

[3] We were unable to recruit evaluators for two languages, namely Chinese and Malayalam.

we measure self-consistency by (a) translating English to any of the non-English language and then translating back to English and (b) translating non-English to English and then translating back to the original language. We obtain average METEOR score 0.57 and an average BERTScore of 0.92 between original query and translated back via another non-English language. These results confirm high translation quality, with back-translated queries remaining consistent with the original English queries and largely preserving their meaning.

4 CoLITE-CLIP

Given a pool of N images containing scene text in any language of a set of languages Ω, and a natural language query expressed in one of these languages, our goal is to retrieve the most relevant image at the top of the ranked list. In our cross-lingual experiments, $\Omega = \{$*English, Chinese, Assamese, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Odia, Punjabi, Tamil, Telugu*$\}$. To address this problem, we leverage the state-of-the-art OCR engines and machine translators for respective languages and take the problem to the space of English by translating all the scene text and queries into English. Thanks to the recent advancements in machine translation literature [35, 38], these translations are reasonably good in quality for the purpose of cross-lingual scene-text aware retrieval. Given these, our major technical contribution comes from the proposed approach – CoLITE-CLIP (Co-attentive Late Interaction for Text-Image Embedding over CLIP), that enhances the CLIP [39] representations by co-attending visual signals with OCR-recognized text (or translated text) from the images. Further, inspired by the seminal information retrieval framework ColBERT [17], CoLITE-CLIP learns a multi-vector representation for scene-text images, enabling fine-grained late interaction with the query through a MaxSim operation. This design captures both visual and textual nuances of scene-text images, leading to more precise cross-lingual retrieval. An overview of the proposed framework is illustrated in Figure 4.

The proposed framework has four components: (i) Pretrained Vision-Language CLIP backbone that yields patch-level image embeddings and token-level text embeddings in a shared space; (ii) OCR Embedding Module that extracts scene text and re-encodes each token with the CLIP text encoder; (iii) Multi-Block Bidirectional Cross-Attention Fusion that exchanges information between the image patches and the OCR tokens to produce a scene-text–aware token bank; and (iv) Late-Interaction Retrieval Head that scores query-image pairs using ColBERT-style MaxSim over the fused token bank, optimized with a symmetric, temperature-scaled InfoNCE loss using in-batch negatives.

We pretrain CoLITE-CLIP on TextCaps and subsequently, fine-tune on X-STAR, enabling robust multilingual generalization.

4.1 Pretrained Vision-Language CLIP Backbone

We adopt the CLIP ViT-B/32 backbone [39] as our image encoder and extract patch-level features from its final transformer block and project them into CLIP's shared 512-dimensional image–text embedding space. For a mini-batch $\mathbf{I} = \{i_1, \ldots, i_B\}$, the encoder returns per-token image representations $V = \Phi_v^{\text{post}}(\mathbf{I}) \in \mathbb{R}^{B\times P\times D}$, where $D = 512$ and P is the number of visual tokens per image. With ViT-B/32 at 224×224 resolution, $P = 7\times7 + 1 = 50$ (49 patch tokens plus one `[CLS]` token). We retain all P tokens per image as the visual token bank $\mathbf{V}$ and *do not* apply CLIP's default global pooling at this stage; the global head is used only to define the shared embedding space, while downstream modules consume $\mathbf{V}$ token-wise.

For a mini-batch of B images, each image is paired with aligned caption (i.e., positive query). We stack all captions into Q queries and pad or truncate each query to a fixed length $T_{\max}$, obtaining token embeddings $E \in \mathbb{R}^{Q\times T_{\max}\times D}$ with $D = 512$, together with a validity mask $\mathbf{Z} \in \{0,1\}^{Q\times T_{\max}}$ that marks non-padding tokens.

4.2 OCR Embedding Module

We incorporate scene text alongside visual tokens. For Indic images in X-STAR, we use an Indic OCR engine, namely IndicPhotoOCR [40] that outputs *line-level* outputs. For Chinese images sourced via RCTW-17 [34], we adopt the dataset provided line-level OCR tokens. For TextCaps, OCR is provided by Rosetta [41] at the *word* level; we convert

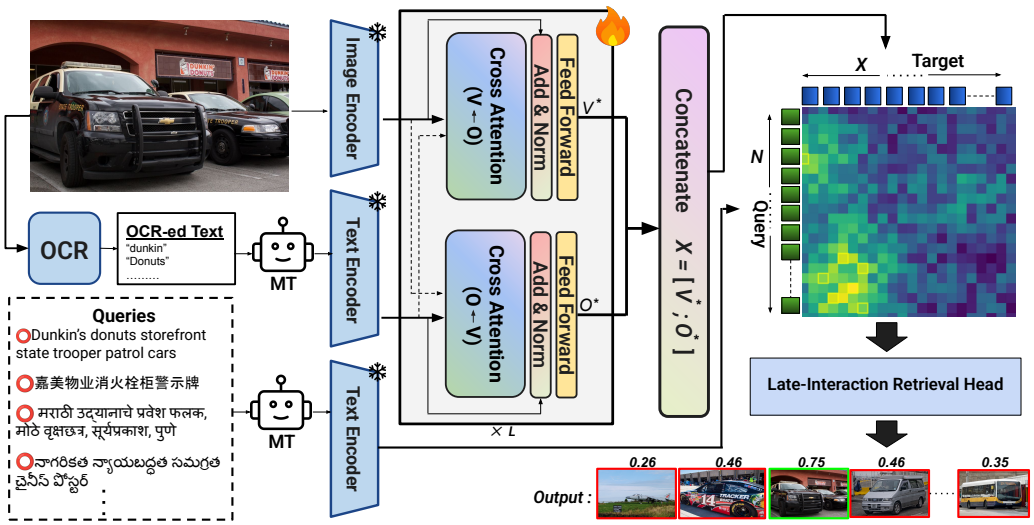


Fig. 4 Overview of CoLITE-CLIP. (i) *Pretrained CLIP backbone (frozen):* the image encoder outputs patch tokens and the text encoder encodes text into token embeddings. (ii) *OCR and MT:* scene text is extracted from the image; when it is not in English, it is machine-translated and then encoded by the CLIP text encoder. (iii) *Bi-directional co-attention fusion:* a stack of L cross-attention blocks (V↔O) with Add & Norm and a feed-forward network fuses the visual and OCR streams into an OCR-aware image representation X. (iv) *Late-interaction head (MaxSim):* the query (translated to English if needed) is encoded by the same CLIP text encoder and compared with X via token-level MaxSim to rank the images; the top-scoring images are returned.

words to line segments using a centroid-based grouping [42] with slope-aware fitting, yielding line strings suitable for downstream fusion.

Let Ψ denote the OCR aggregator. For an image i, $\Psi(i)$ returns a sequence of line strings $\mathcal{S}_i = \{\ell_{i,1}, \ldots, \ell_{i,N_o^{(i)}}\}$. Each line is encoded once with the *same* CLIP text encoder used for queries, and we keep the *post-transformer, projected* representation (512-d in CLIP's joint space). This gives $O_i = \Phi_t^{\text{post}}(\mathcal{S}_i) \in \mathbb{R}^{N_o^{(i)} \times D}$, where $N_o^{(i)}$ is the number of OCR lines in image i (capped at N_o) and each row is a single D-dimensional embedding for one OCR *line* (we do not store per-token (≤ 77) text positions for OCR).

4.3 Multi-Block Bidirectional Cross-Attention Fusion

We fuse the visual patch tokens V with the OCR tokens O using a bi-directional co-attention block. Each direction applies: (i) LayerNorm, (ii) cross-attention where one stream attends to the other, and (iii) a residual FFN update. This design injects text cues into visual patches and grounds OCR tokens in image evidence, following a standard Pre-LayerNorm Transformer pattern with GELU+dropout MLPs . Let V^* and O^* denote the visual and OCR tokens after the last of the L blocks; we concatenate them into the token bank

$$X = [\,V^*; O^*\,] \in \mathbb{R}^{B \times M \times D}, \quad M = P + N_o. \quad (1)$$

The final token bank X is ℓ_2-normalized and scored using a ColBERT-style MaxSim late-interaction layer as described below.

4.4 Late-Interaction Retrieval Head

Given the fused OCR-aware image token bank $X = [V^*; O^*] \in \mathbb{R}^{B \times M \times D}$ (with $M = P + N_o$) and a caption token sequence $q = \{q_t\}_{t=1}^{T}$, we compute a ColBERT-style *MaxSim* score by matching each caption token to its best image token and summing the matches. All token vectors are ℓ_2-normalized, so dot products equal cosine similarities.

$$s(q, X) = \sum_{t=1}^{T} \max_{m \in \{1,\ldots,M\}} \langle q_t, X_m \rangle. \quad (2)$$

In a mini-batch of size B, we build a score matrix $S \in \mathbb{R}^{B \times B}$ with entries $S_{ij} = s\big(q^{(i)}, X^{(j)}\big)$.

4.5 Training

With temperature τ, we optimize a *symmetric* InfoNCE objective (i.e., text→image and image→text) given by the following expression:

$$\mathcal{L}_{\text{InfoNCE}}(S;\tau) = \tfrac{1}{2}\big[\text{CE}(S/\tau,\, y) + \text{CE}(S^{\top}/\tau,\, y)\big], \tag{3}$$

where CE denotes the row-wise cross-entropy loss and $y_i = i$ ($i = 1, \ldots, B$), i.e., the matching query–image pairs lie on the diagonal of S.

Further, the loss is averaged over the number of positive captions per image (C), where $S^{(k)}$ is the score matrix computed using the k-th caption of every image in the batch:

$$\mathcal{L} = \frac{1}{C}\sum_{k=1}^{C} \mathcal{L}_{\text{InfoNCE}}\big(S^{(k)};\tau\big). \tag{4}$$

5 Experiments and Results

5.1 Datasets Used and Metrics

We evaluated CoLITE-CLIP on two datasets: our newly introduced X-STAR (see Section 3) and TextCaps [18]. Unlike TextCaps, which is monolingual scene text retrieval benchmark with English-only queries and scene text, X-STAR is a multilingual dataset that supports queries in 13 languages and contains scene text spanning these languages. For X-STAR, we report retrieval performance in the test split comprising 1,289 images and 16,757 queries. For TextCaps, we use the validation split and perform retrieval using the provided images and their associated captions.

We report recall at K (denoted R@k) with k=1,5 and 10 to evaluate the proposed method.

5.2 Implementation Details

All models are trained on a single NVIDIA RTX A6000 (≈48 GB). All images are resized to 224×224 by the CLIP ViT-/32. For pretraining on TextCaps, we use a batch size of 20 for 40 epochs, AdamW optimizer, a peak learning rate of 1×10^{-3} with a 5,000-step linear warmup and cosine decay, dropout = 0.1, and temperature $\tau = 0.02$. The OCR tokens are capped at N_o=30 per image. The fusion head uses 11 blocks, each with 4 attention heads and a feed-forward layer of width $2D$; the complete training configuration is provided in Appendix A.

5.3 Baselines and Prior Approaches

We consider a wide range of baselines and prior approaches to assess the efficacy of the proposed model. Following Table 2, we group them into two categories according to how they handle the language gap, described below; within each category, the table further separates methods by the source of information used for retrieval (OCR-only, vision-only, and vision+OCR).

5.3.1 Pure Cross-lingual Approaches

These approaches do not translate the scene text or the query. Instead, they directly perform semantic matching in a shared multilingual representation space. This category includes the multilingual text encoders LaBSE [30] and Multi-e5-large [43], which match queries against the OCR text, and the multilingual vision-language models mCLIP [16], multilingual SigLIP [26], and SigLIP2 [15].

5.3.2 Cross-lingual Reduced to Monolingual Approaches

Here, the query and the scene text are first translated into English using an off-the-shelf machine translation system, and retrieval is then performed by an English retrieval method. This category includes (i) text-based retrievers over the translated OCR text, namely BM25 (OCR text) [44], MiniLM [45], and LaBSE and Multi-e5-large applied after translation; (ii) English vision-language models, namely CLIP [39], MetaCLIP [24], OpenCLIP [25], SigLIP [26], SigLIP2 [15], EVA-CLIP [46], and ELIP [47]; and (iii) image–text and scene-text-aware retrieval methods originally designed for English queries and English scene text, namely SCAN [48], VSRN [49], SGRAF [50], IMRAM [51], GSMN [52], HREM [53], STAR-Net [2], LightningDOT (w/ST+MSM) [54], ST-SM [55], HOPID [56], and MSTAR [57]. The proposed CoLITE-CLIP also belongs to this category. It should be noted that methods categorized as *Pure Cross-lingual Approaches* can also be evaluated under this setting by first translating the inputs to English and then performing retrieval,

allowing us to study whether explicit translation provides any advantage over direct multilingual semantic matching.

5.4 Results and Discussion

We report the retrieval performance of CoLITE-CLIP on X-STAR and TextCaps in Table 2, and compare it with the baselines and prior approaches described in Section 5.3.

5.4.1 Results on TextCaps

On TextCaps, our method significantly outperforms traditional monolingual retrieval approaches. The best previously reported result on this benchmark is R@1 = 54.7 from a recent method [55], which our model surpasses by 8.1 points (62.8% vs. 54.7% R@1). Notably, CoLITE-CLIP achieves this performance with a comparatively compact backbone relative to several competing approaches, indicating that the improvement does not simply result from increased model scale. Methods based on strong multilingual representations, such as Multilingual SigLIP and SigLIP, achieve slightly higher performance when evaluated with English queries on the TextCaps dataset, obtaining an R@1 that is 2.7 and 5.4 points higher than ours. However, these approaches fall considerably short when dealing with multilingual queries and scene text in X-STAR. This observation suggests that despite recent advances in multilingual representations, cross-lingual semantic matching, particularly for low-resource languages, remains challenging.

5.4.2 Results on X-STAR

The advantage of our approach becomes much clearer on the more challenging X-STAR benchmark. Our method translates both the query and the scene text into English using an off-the-shelf machine translation system and performs retrieval using a late multi-vector interaction mechanism instead of a single global representation over CLIP features. This design significantly outperforms all other publicly available approaches on X-STAR. In particular, our method improves the vanilla CLIP baseline from 24.2% to 46.5% in terms of R@1. More interestingly, it surpasses all pure cross-lingual approaches by a large margin.

These results highlight two key insights. First, reducing the problem to a monolingual setting via translation allows us to operate in a unified semantic space, and partially normalize OCR noise before retrieval. Second, the use of multi-vector late interaction enables fine-grained matching between query and scene text tokens, which is particularly beneficial when the relevant information is sparse, noisy, or localized within the image text. In contrast, global embedding approaches often dilute such fine-grained signals. Together, these design choices make our approach more robust for multilingual scene text retrieval in real-world settings.

5.5 Ablations and Analysis

5.5.1 Query Language-wise Results

We show query language-wise retrieval results on X-STAR in Table 3. We observe that our model CoLITE-CLIP substantially outperforms our base model vanilla CLIP in *all* languages. The performance on English queries is best, and others are slightly inferior; this may reflect limitations of current machine-translation systems, which can produce literal or inconsistent translations of named entities and thereby misalign queries with the relevant images. Despite this, R@10 remains largely consistent across all languages.

5.5.2 Impact of OCR Quality

The quality of OCR is an important factor influencing the overall cross-modal retrieval performance. As shown in Table 2, the R@1 score of CoLITE-CLIP improves by nearly 18% (absolute) when ground-truth OCR is used instead of automatically recognized text. This highlights that OCR errors significantly affect the downstream retrieval performance. Because OCR for several languages in X-STAR remains underdeveloped, better recognition accuracy is expected to further improve retrieval performance.

5.5.3 Retrieval Using User-Written Queries

We evaluated the cross-lingual retrieval performance using 3,642 user-written queries. Our model CoLITE-CLIP achieves 30.1% 54.3%, 64.4% Recall at 1, 5 and 10, respectively. This

Table 2 Retrieval results on our newly introduced dataset, namely X-STAR and TextCaps. We report mean Recall@K (%) over all test queries; for cross-lingual retrieval, results are averaged over test queries from 13 languages. Methods are grouped by the primary source of information used for retrieval: OCR-only, vision-only, and vision+OCR. CoLITE-CLIP outperforms all prior approaches, both pure cross-lingual and cross-lingual reduced to monolingual, on the more challenging X-STAR dataset, and remains competitive on the monolingual TextCaps dataset.

Method	X-STAR			TextCaps		
	R@1	R@5	R@10	R@1	R@5	R@10
1. Only OCR-based Retrieval						
(a) Pure Cross-lingual Approaches						
LaBSE [30]	23.3	36.1	42.9	–	–	–
Multi-e5-large [43]	19.9	32.5	38.7	–	–	–
(b) Cross-lingual reduced to Monolingual Approaches						
LaBSE + MT [30]	25.6	39.2	45.9	23.1	37.0	42.7
Multi-e5-large [43]	26.7	40.3	45.8	17.8	32.8	40.3
BM25 [44]	43.0	57.1	61.0	39.2	51.8	56.4
MiniLM-L6-v2 [45]	30.3	44.2	48.6	29.8	43.5	48.8
2. Only Vision-based Retrieval						
(a) Pure Cross-lingual Approaches						
M-CLIP-B/16 [16]	9.7	22.5	30.0	39.3	60.0	69.1
M-SigLIP-B/16 [26]	12.7	23.9	30.2	65.5	80.6	85.6
SigLIP2-B/32 [15]	7.4	14.1	18.2	56.6	74.7	81.0
(b) Cross-lingual reduced to Monolingual Approaches						
SCAN [48]	3.1	8.7	12.8	46.8	64.0	69.8
VSRN (f.t.) [49]	9.4	26.2	36.4	44.9	62.7	68.7
SGRAF [50]	7.2	20.0	27.8	42.9	58.4	63.9
IMRAM [51]	3.6	9.2	14.9	47.9	65.6	71.4
GSMN [52]	4.5	12.5	18.5	48.1	66.3	72.3
HREM [53]	6.0	17.8	26.6	51.7	71.0	77.5
CLIP-B/32 [39]	24.2	42.1	51.3	46.4	66.1	73.8
MetaCLIP-B/32 [24]	23.9	42.7	51.0	49.4	68.7	76.2
OpenCLIP-B/32 [25]	22.4	41.1	48.9	46.3	65.6	73.4
SigLIP-B/16 [26]	39.6	62.6	71.4	**68.2**	**83.0**	**87.8**
SigLIP2-B/32 [15]	30.0	51.4	60.1	56.6	74.7	81.0
EVA02-CLIP-B/16 [46]	24.9	45.3	54.7	51.7	70.9	78.4
ELIP-C [47]	31.5	48.5	56.7	–	–	–
3. Vision + OCR-based Retrieval						
(b) Cross-lingual reduced to Monolingual Approaches						
STARNet (f.t.) [2]	8.9	24.4	34.7	19.8	40.1	51.6
LightningDOT [54]	–	–	–	40.5	63.0	71.1
ST-SM [55]	–	–	–	54.7	73.4	79.3
HOPID [56]	–	–	–	49.5	70.7	78.6
MSTAR [57]	21.6	34.3	40.4	–	–	–
CoLITE-CLIP (Ours)	**46.5**	**65.8**	**73.0**	62.8	81.2	86.7
CoLITE-CLIP (w/ GT OCR)	64.2	83.7	88.2	–	–	–

shows that our model that is trained on LLM-generated queries generalizes well with user-written queries as well.

5.5.4 Ablation Study

The proposed CoLITE-CLIP has two innovative modules, namely OCR-visual co-attention and



Fig. 5 Top-five retrieval results of CoLITE-CLIP for queries in five languages: (a) Telugu, (b) Bengali, (c) Gujarati, (d) Hindi, and (e) Chinese; the English translation of each query is shown below it. Retrieved images are ranked from left to right; the correct image is framed in green and incorrect ones in red. The correct image is retrieved at rank 1 for four of the five queries, including the cross-lingual case in (a), where a Telugu query retrieves an image whose scene text is in Chinese; in (d), it is retrieved at rank 2. Incorrect candidates are often visually similar scenes (e.g., other railway platforms in (b)), which highlights the importance of matching the scene text. (Best viewed in color.)

Late Interaction. To study their utility over the base CLIP model, we perform an ablation and report the result in Table 4. We observe that the co-attention module improves R@1 of the vanilla

Table 3 Language-wise retrieval results on X-STAR (%). Query-language-wise comparison of CoLITE-CLIP with the CLIP baseline. CoLITE-CLIP consistently outperforms CLIP across all languages. Performance on English is slightly higher, suggesting some information loss during machine translation. Note that both the query and the scene text in images may appear in any of these languages.

Method	**CoLITE-CLIP**			**CLIP baseline**		
	R@1	**R@5**	**R@10**	**R@1**	**R@5**	**R@10**
English	52.3	69.3	75.7	26.5	44.8	54.5
Assamese	44.2	62.1	70.1	23.3	41.2	51.1
Bengali	45.1	65.0	72.9	23.8	41.5	50.4
Gujarati	47.0	68.0	74.6	24.1	42.4	52.3
Hindi	46.8	67.0	72.9	24.7	42.3	51.8
Kannada	45.8	64.7	72.3	23.7	42.4	51.6
Malayalam	49.0	67.6	74.9	25.2	43.2	51.8
Marathi	43.0	63.5	70.4	23.4	40.1	48.9
Odia	46.9	65.7	73.6	24.1	40.8	50.7
Punjabi	46.4	66.2	73.1	23.7	41.8	51.5
Tamil	43.8	65.1	73.2	23.6	41.6	50.2
Telugu	46.4	65.7	73.2	24.3	42.7	50.7
Chinese	47.5	65.7	72.1	24.4	42.1	50.8
Average	**46.5**	**65.8**	**73.0**	24.2	42.1	51.3

Table 4 Component-wise ablation on X-STAR (%). Co-Attn. denotes OCR–visual co-attention and Late Int. denotes late interaction. The top row corresponds to the base CLIP model, whereas the bottom row corresponds to our full model.

Model Variant	**Co-Attn.**	**Late Int.**	**R@1**	**R@5**	**R@10**
CLIP baseline	✗	✗	24.2	42.1	51.3
CLIP + Co-Attention	✓	✗	37.7	62.1	71.2
CoLITE-CLIP	✓	✓	**46.5**	**65.8**	**73.0**

CLIP model by 13.5 points (from 24.2% to 37.7% R@1), attributed to its seamless integration of OCR and visual embeddings. Further, ColBERT-style late interaction further improves it by 8.8 points (to 46.5% R@1), indicating the importance of multi-vector representation and late interaction for fine-grained retrieval.

5.5.5 Qualitative Results

With a Telugu query (Figure 5), our model retrieves the image that both matches the scene context and contains the target scene text in Chinese, which demonstrates cross-lingual, scene-text–aware retrieval.

5.6 Limitations

The proposed model has significantly improved the base model in multiple scenarios, including the cross-lingual retrieval setting. We identify two limitations of the current work: (i) Despite improvements in machine translation, they are not yet flawless and introduce some literal or inconsistent translations of named entities and script normalization differences, which misalign queries with a real image (ii) Scene text recognition for low-resource languages is far from being solved. We leave addressing these two shortcomings as future work.

6 Conclusion

This work is the first to systematically study the cross-lingual scene text-aware retrieval task. We introduced a semi-automatically curated large-scale dataset namely X-STAR, specifically designed for this purpose, along with a strong baseline CoLITE-CLIP. The novelty of our proposed approach lies in its two innovative modules, i.e., OCR-visual co-attention and late-interaction that empower the base CLIP model to perform scene text-aware retrieval more effectively. Experimental results demonstrate that our model

significantly outperforms existing baselines and successfully retrieves images even when the query language differs from the scene text. We firmly believe that this work will open up a promising new direction for the cross-lingual cross-modal information retrieval community.

Acknowledgments

This work is supported by the Ministry of Electronics and Information Technology, Govt. of India, under the NLTM-Bhashini Project and a sponsored project by TIH iHub Drishti Foundation.

Declarations

- **Funding:**
- **Conflict of interest/Competing interests:** The authors declare no competing interests.
- **Ethics approval and consent to participate:** Not Applicable
- **Consent for publication:** Not applicable.
- **Data availability:** The X-STAR dataset is publicly available at https://github.com/vl2g/X-STAR.
- **Materials availability:** Not applicable.
- **Code availability:** The code is publicly available at https://github.com/vl2g/X-STAR.
- **Author contribution:** Dikshant led the implementation of the work, conducted the experiments, and prepared the initial draft. Prateek contributed by implementing and running selected baselines. Anand supervised the overall project and contributed to refining and improving the manuscript.

References

1. Wang, H., Bai, X., Yang, M., Zhu, S., Wang, J., Liu, W.: Scene text retrieval via joint text detection and similarity learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4558–4567 (2021)
2. Mafla, A., Rezende, R.S., Gomez, L., Larlus, D., Karatzas, D.: StacMR: Scene-text aware cross-modal retrieval. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 2220–2230 (2021)
3. Biten, A.F., Tito, R., Mafla, A., Gomez, L., Rusinol, M., Valveny, E., Jawahar, C., Karatzas, D.: Scene text visual question answering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4291–4301 (2019)
4. Jégou, H., Douze, M., Schmid, C., Pérez, P.: Aggregating local descriptors into a compact image representation. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3304–3311 (2010). IEEE
5. Sivic, J., Zisserman, A.: Video Google: A text retrieval approach to object matching in videos. In: Proceedings Ninth IEEE International Conference on Computer Vision, pp. 1470–1477 (2003). IEEE
6. Nister, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 2, pp. 2161–2168 (2006). IEEE
7. Mishra, A., Alahari, K., Jawahar, C.: Image retrieval using textual cues. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3040–3047 (2013)
8. Shi, B., Bai, X., Yao, C.: An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. IEEE transactions on pattern analysis and machine intelligence **39**(11), 2298–2304 (2016)
9. Zhou, X., Yao, C., Wen, H., Wang, Y., Zhou, S., He, W., Liang, J.: EAST: an efficient and accurate scene text detector. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5551–5560 (2017)
10. Liao, M., Wan, Z., Yao, C., Chen, K., Bai, X.: Real-time scene text detection with differentiable binarization. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, pp. 11474–11481 (2020)
11. Bissacco, A., Cummins, M., Netzer, Y., Neven, H.: PhotoOCR: Reading text in uncontrolled conditions. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 785–792 (2013)
12. Cheng, M., Sun, Y., Wang, L., Zhu, X., Yao, K., Chen, J., Song, G., Han, J., Liu, J., Ding, E., *et al.*: ViSTA: Vision and scene

text aggregation for cross-modal retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5184–5193 (2022)
13. Li, G., Shen, H., Zhou, Y.: Beyond cropped regions: New benchmark and corresponding baseline for Chinese scene text retrieval in diverse layouts. arXiv preprint arXiv:2506.04999 (2025)
14. Wang, X., Liu, Y., Shen, C., Ng, C.C., Luo, C., Jin, L., Chan, C.S., Hengel, A.v.d., Wang, L.: On the general value of evidence, and bilingual scene-text visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10126–10135 (2020)
15. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025)
16. Chen, G., Hou, L., Chen, Y., Dai, W., Shang, L., Jiang, X., Liu, Q., Pan, J., Wang, W.: mCLIP: Multilingual CLIP via cross-lingual transfer. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 13028–13043 (2023)
17. Khattab, O., Zaharia, M.: ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 39–48 (2020)
18. Sidorov, O., Hu, R., Rohrbach, M., Singh, A.: TextCaps: a dataset for image captioning with reading comprehension. In: European Conference on Computer Vision, pp. 742–758 (2020). Springer
19. Shahab, A., Shafait, F., Dengel, A.: ICDAR 2011 robust reading competition challenge 2: Reading text in scene images. In: 2011 International Conference on Document Analysis and Recognition, pp. 1491–1496 (2011). IEEE
20. Mishra, A., Shekhar, S., Singh, A.K., Chakraborty, A.: OCR-VQA: Visual question answering by reading text in images. In: 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 947–952 (2019). https://doi.org/10.1109/ICDAR.2019.00156
21. Singh, A., Natarjan, V., Shah, M., Jiang, Y., Chen, X., Parikh, D., Rohrbach, M.: Towards VQA models that can read. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8317–8326 (2019)
22. Singh, A.K., Mishra, A., Shekhar, S., Chakraborty, A.: From strings to things: Knowledge-enabled VQA model that can read and reason. In: ICCV (2019)
23. Jaderberg, M., Vedaldi, A., Zisserman, A.: Deep features for text spotting. In: European Conference on Computer Vision, pp. 512–528 (2014). Springer
24. Xu, H., Xie, S., Tan, X.E., Huang, P., Howes, R., Sharma, V., Li, S., Ghosh, G., Zettlemoyer, L., Feichtenhofer, C.: Demystifying CLIP data. In: The Twelfth International Conference on Learning Representations,(ICLR) 2024 (2024)
25. Ilharco, G., Wortsman, M., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., et al.: OpenCLIP. Zenodo (2021)
26. Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 11975–11986 (2023)
27. Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., Xu, C.: FILIP: Fine-grained interactive language-image pre-training. In: International Conference on Learning Representations (ICLR) (2022)
28. Faysse, M., Sibille, H., Wu, T., Omrani, B., Viaud, G., Hudelot, C., Colombo, P.: ColPali: Efficient document retrieval with vision language models. In: International Conference on Learning Representations (ICLR) (2025)
29. Masry, A., Thakkar, M., Bechard, P., Madhusudhan, S.T., Awal, R., Mishra, S., Suresh, A.K., Daruru, S., Hoque, E., Gella, S., Scholak, T., Rajeswar, S.: ColMate: Contrastive late interaction and masked text for multimodal document retrieval. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, pp. 2071–2080. Association

for Computational Linguistics, Suzhou, China (2025)
30. Feng, F., Yang, Y., Cer, D., Arivazhagan, N., Wang, W.: Language-agnostic BERT sentence embedding. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 878–891 (2022)
31. Artetxe, M., Schwenk, H.: Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. Transactions of the association for computational linguistics **7**, 597–610 (2019)
32. Litschko, R., Vulić, I., Ponzetto, S.P., Glavaš, G.: On cross-lingual retrieval with multilingual text encoders. Information Retrieval Journal **25**(2), 149–183 (2022)
33. De, A., Penamakuri, A.S., Yadav, R., Rathore, A., Shah, H., Sharma, D., Agarwal, S., Kumar, P., Mishra, A.: Bharat Scene Text: A novel comprehensive dataset and benchmark for Indian language scene text understanding. arXiv preprint arXiv:2511.23071 (2025)
34. Shi, B., Yao, C., Liao, M., Yang, M., Xu, P., Cui, L., Belongie, S., Lu, S., Bai, X.: ICDAR2017 competition on reading Chinese text in the wild (RCTW-17). In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 1, pp. 1429–1434 (2017). IEEE
35. Gala, J., Chitale, P.A., Raghavan, A.K., Gumma, V., Doddapaneni, S., M, A.K., Nawale, J.A., Sujatha, A., Puduppully, R., Raghavan, V., Kumar, P., Khapra, M.M., Dabre, R., Kunchukuttan, A.: IndicTrans2: Towards high-quality and accessible machine translation models for all 22 scheduled Indian languages. Transactions on Machine Learning Research (2023)
36. Google: Google Translate. Accessed: 2025-10-10 (2025). https://translate.google.com
37. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: GPT-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
38. Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al.: Google's neural machine translation system: Bridging the gap between human and machine translation. arXiv preprint arXiv:1609.08144 (2016)
39. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., *et al.*: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763 (2021). PmLR
40. De, A., et al.: IndicPhotoOCR: A comprehensive toolkit for Indian language scene text understanding. https://github.com/Bhashini-IITJ/IndicPhotoOCR/ (2024)
41. Borisyuk, F., Gordo, A., Sivakumar, V.: Rosetta: Large scale system for text detection and recognition in images. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 71–79 (2018)
42. Hu, R., Singh, A., Darrell, T., Rohrbach, M.: Iterative answer prediction with pointer-augmented multimodal transformers for TextVQA. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9992–10002 (2020)
43. Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., Wei, F.: Multilingual E5 text embeddings: A technical report. arXiv preprint arXiv:2402.05672 (2024)
44. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: BM25 and beyond. Foundations and Trends in Information Retrieval **3**(4), 333–389 (2009)
45. Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., Zhou, M.: MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. Advances in neural information processing systems **33**, 5776–5788 (2020)
46. Sun, Q., Fang, Y., Wu, L., Wang, X., Cao, Y.: EVA-CLIP: improved training techniques for CLIP at scale. CoRR **abs/2303.15389** (2023) https://doi.org/10.48550/ARXIV.2303.15389 2303.15389
47. Zhan, G., Liu, Y., Han, K., Xie, W., Zisserman, A.: ELIP: Enhanced visual-language foundation models for image retrieval. In: 2025 International Conference on Content-Based Multimedia Indexing (CBMI), pp. 1–8

(2025). IEEE
48. Lee, K.-H., Chen, X., Hua, G., Hu, H., He, X.: Stacked cross attention for image-text matching. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 201–216 (2018)
49. Li, K., Zhang, Y., Li, K., Li, Y., Fu, Y.: Visual semantic reasoning for image-text matching. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4654–4662 (2019)
50. Diao, H., Zhang, Y., Ma, L., Lu, H.: Similarity reasoning and filtration for image-text matching. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 1218–1226 (2021)
51. Chen, H., Ding, G., Liu, X., Lin, Z., Liu, J., Han, J.: IMRAM: Iterative matching with recurrent attention memory for cross-modal image-text retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12655–12663 (2020)
52. Liu, C., Mao, Z., Zhang, T., Xie, H., Wang, B., Zhang, Y.: Graph structured network for image-text matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10921–10930 (2020)
53. Fu, Z., Mao, Z., Song, Y., Zhang, Y.: Learning semantic relationship among instances for image-text matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15159–15168 (2023)
54. Miyawaki, S., Hasegawa, T., Nishida, K., Kato, T., Suzuki, J.: Scene-text aware image and text retrieval with dual-encoder. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pp. 422–433 (2022)
55. Cheng, S., Zhang, F., Zhang, X., Sun, Z.: Scene-text aware cross-modal retrieval based on semantic matching. Multimedia Systems **30**(5), 284 (2024)
56. Zhao, Z., Li, L., Zhang, J., Sun, Y., Sheng, X., Yin, H., Jiang, S.: Heterogeneous prompt-guided entity inferring and distilling for scene-text aware cross-modal retrieval. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 39, pp. 10537–10545 (2025)
57. Yin, L., Xie, X., Li, Z., Bai, X., Liu, Y.: MSTAR: Box-free multi-query scene text retrieval with attention recycling. Advances in Neural Information Processing Systems **38**, 57856–57881 (2026)

Appendix A Implementation

A.1 Training Configuration

We train the ColBERT-style late-interaction fusion module on precomputed CLIP ViT-B/32 patch embeddings and token-level query embeddings from the frozen CLIP text encoder. Table A1 lists the configuration. The loss is the symmetric InfoNCE over in-batch image–caption pairs (Eq. (3)). For TextCaps, the loss is applied independently to each of the five captions per image and then averaged; for X-STAR, training uses one English query per image.

Table A1 Training configuration of CoLITE-CLIP.

Setting	Value
Embedding dimension D	512
Attention heads H	4
Fusion blocks L	11
Max. OCR tokens per image N_o	30 (line-level)
Batch size	20
Optimizer	AdamW
Peak learning rate	10^{-3}
Weight decay	10^{-3}
LR schedule	Cosine (5k warm-up)
Dropout	0.1
FFN expansion	2
Training epochs	40
Random seed	42

For each image, we load ViT-B/32 patch tokens $V \in \mathbb{R}^{P\times D}$ and line-level OCR embeddings $O \in \mathbb{R}^{N_o\times D}$; missing OCR tokens are replaced by a learned null embedding. The fusion blocks perform bidirectional cross-attention between V and O and return a fused token bank $X = [V; O] \in \mathbb{R}^{M\times D}$ with $M = P + N_o$, which is ℓ_2-normalized per token. Late interaction is implemented via MaxSim over this bank.

During training, we enable TF32 matrix multiplications on Ampere GPUs and use a sharded ColBERT index with memory-mapped arrays and an LRU cache to stream text tokens from disk. In-batch MaxSim scores are computed in chunks of 20 to keep memory usage bounded. Each reported result is obtained from a single training run, while hyperparameters were varied during development without changing the seed.

A.2 Computing Environment

Experiments were conducted on a system with an AMD Ryzen Threadripper PRO 3975WX CPU, 251 GiB RAM, and NVIDIA RTX A6000 GPUs with approximately 48 GB memory each. We used Ubuntu 20.04.6 LTS, Python 3.10.19, PyTorch 2.4.1 with CUDA 12.1, cuDNN 9.1, and torchvision 0.19.1.

Appendix B Dataset Details

B.1 User Validation of Queries

We evaluate the reliability of our multilingual scene-text data by constructing a user-validation pipeline for 13 languages: Assamese, Bengali, Gujarati, Hindi, Odia, Kannada, Malayalam, Marathi, Punjabi, Tamil, Telugu, English, and Chinese. The final dataset includes 6,449 images representing both indoor and outdoor scenes with naturally occurring text. For each image, we generated language-specific queries and corresponding annotations using our annotation interface.

To assess the quality of the generated queries, we sampled 50 images for each of ten languages, namely the Indian languages other than Malayalam (English, Chinese, and Malayalam were not evaluated), resulting in a total of 500 images. We then deployed a web-based validation tool. The interface showed annotators the image, the corresponding query, and a set of standardized evaluation criteria. Annotators were asked to judge:

- *Image–query relevance:* whether the query correctly matches the image content;
- *Grammaticality:* whether the query is linguistically correct in the target language;
- *Scene-text awareness:* whether the query appropriately refers to visible text in the image;
- *Visual feature dependence:* whether the query requires understanding non-textual visual cues;
- *Realism:* a 1–5 score indicating whether the query plausibly reflects real-world search behavior.

Each annotator's responses were automatically saved in a CSV file. We aggregated the judgments across the 50 samples per language and computed average annotation quality for each evaluation dimension, thereby quantifying the reliability of

the dataset. These user-verified scores provide an empirical measure of annotation accuracy and linguistic fidelity across the 13 languages.

B.2 Validation Results

Out of the 500 samples, 70.2% queries were judged to be fully correct, 25.4% queries were rated as partially correct, and Only 4.4% were considered incorrect. Taken together, these results indicate that 95.6% of the generated queries exhibit at least partial alignment with the image content. Regarding grammatical quality 52.0% queries were rated as grammatically correct, 37.2% contained minor grammatical issues that could be improved but did not hinder understanding. Only 10.8% queries were considered grammatically incorrect suggesting that nearly 90% of the generated queries are either grammatically sound or require only minimal refinement. To evaluate whether the queries make effective use of the textual information provided in the image, we measured scene-text awareness. We found that 94.6% of the queries used information from the scene text, while only 5.4% of the queries did not make use of scene text information. To evaluate whether the queries require the interpretation of the image's content, we found that for 88.6% of the queries, it was necessary for the annotators to look at the image's content in order to be able to answer the query. Only 11.4% of the queries could be answered without reference to the image's content.

Finally, we asked annotators to rate the realism of the generated queries on a five-point scale, where higher ratings correspond to more realistic search queries. The realism of the generated queries was rated with an average of 4.36 out of 5. In addition, with regard to the score distribution, it is noticeable that while 45.6% of the queries were rated with a score of 5, 45.4%, was rated with a score of 4. Only a small percentage of queries were rated with lower scores, i.e., 8.2% with a score of 3, 0.6% with a score of 2, and 0.2% with a score of 1. In any case, it is noticeable that a total of 91% of the queries were rated with a realism score of 4 and above, which implies that the generated queries are very similar to real search queries.

B.3 Translation Quality

To assess the quality of translations from English into ten target languages, we conducted a human evaluation. For each language, 50 queries were randomly selected and evaluated by native speakers with expertise in the respective languages. The translations were rated on a scale from 1 to 10, where higher scores indicate better semantic equivalence with the source text.

Across all ten languages, the translations achieved an average score of 8.07 out of 10, indicating a strong semantic correspondence between the source and translated queries. Translation quality was rated through a similar web-based interface that showed the source English query, its translation, and a 1–10 rating scale.

B.4 Scene-Text Usage in Queries

To further quantify how strongly the generated queries rely on visible text, we measure the proportion of queries that contain at least one and at least two scene-text terms across the train, validation, and test splits. As shown in Table B1, a large majority of queries in all splits explicitly incorporate scene text, confirming that the dataset construction pipeline consistently produces scene-text-grounded retrieval queries.

Table B1 Scene-text usage in generated queries across dataset splits: percentage (%) of queries containing at least one ($\geq$1) and at least two ($\geq$2) scene-text terms.

Split	#Images	$\geq$1 term	$\geq$2 terms
Train	4,515	90.25	66.42
Validation	645	90.54	68.53
Test	1,289	89.91	66.41

Appendix C Ablations

In Section 5, we ablate the architectural components of the proposed fusion module. Here, we additionally study the impact of two key hyperparameters, the fusion depth L and the OCR token budget N_o, while keeping the rest of the configuration fixed to the default setting on the TextCaps dataset. Table C1 reports both studies.

C.1 Effect of Fusion Depth

We vary the number of fusion blocks L while keeping $D = 512$, $H = 4$, $N_o = 30$, and the

Table C1 Hyperparameter ablations on the TextCaps validation split: (a) fusion depth L and (b) OCR token budget N_o. The default setting is marked with †.

Setting	R@1	R@5	R@10
(a) Fusion depth L			
$L = 3$	51.5	73.5	86.0
$L = 11^\dagger$	**62.8**	**81.2**	**86.7**
$L = 13$	60.9	79.7	85.6
(b) OCR token budget N_o			
$N_o = 20$	61.4	80.0	85.8
$N_o = 30^\dagger$	**62.8**	**81.2**	**86.7**
$N_o = 40$	61.8	80.5	86.3

optimizer configuration unchanged. As shown in Table C1(a), very shallow models underfit, while deeper stacks provide diminishing returns beyond our default setting.

C.2 Effect of OCR Token Budget

Next, we vary the maximum number of line-level OCR tokens N_o per image while keeping all other hyperparameters fixed. Table C1(b) shows that using too few OCR tokens harms performance, whereas larger values provide only marginal gains while increasing memory and runtime.