Character Retrieval in Manga via Semantic Query Refinement

Seo Won (Sean) Yi
sean.yi@mail.utoronto.ca
University of Toronto

Anna Rui Xu Yang
anna.xuyang@mail.utoronto.ca
University of Toronto

Abstract

Character-based manga retrieval presents unique challenges due to stylized visual conventions, multi-character compositions, and the need to distinguish specific entities within complex panels. While vision-language models like CLIP have demonstrated cross-domain transfer capabilities, their effectiveness for manga character search remains unexplored. This work systematically evaluates strategies for integrating visual and semantic information in character-based manga retrieval using frozen CLIP encoders and LLM-generated character descriptions. We compare single-modality baselines against multimodal fusion approaches, examining both late fusion reranking—a well-established technique that combines modalities at the scoring stage—and query-level semantic integration through our proposed Query-Conditioned Feedback Retrieval (QCFR). Using 630 manga pages across 28 titles with 100 character-focused queries, we analyze how query formulation (full page versus cropped character), dataset configuration (full page versus character-isolated), and semantic integration strategy affect performance. Our experiments reveal that query-level integration consistently outperforms score-level fusion, with QCFR achieving Recall@50 = 0.855 and mAP@50 = 0.631 compared to image-first late fusion's Recall@50 = 0.838 and mAP@50 = 0.618. Both multimodal approaches substantially surpass single-modality baselines, demonstrating that frozen CLIP models can effectively transfer to highly stylized manga imagery when semantic information guides retrieval, and that the timing of semantic integration significantly impacts retrieval quality. [1]

1 Introduction

1.1 Manga as Global Entertainment

Japanese comics, known as manga, have evolved into a global entertainment medium. The industry has seen rapid growth, with international readership expanding through digital platforms and official translations. This popularity stems from the medium's diversity, spanning genres such as action, romance, psychological thriller, and slice-of-life.

With thousands of titles available and new chapters released weekly, readers face significant challenges in finding content that matches their preferences. Unlike text-based media where keyword search is effective, manga discovery requires understanding visual aesthetics, character design, artistic style, and narrative tone [13].

1.2 Manga as a Complex Visual Medium

Manga presents unique challenges for content retrieval due to its complex visual structure. As a narrative form, it combines textual and visual storytelling through panel layouts, artistic style, and visual iconography. A single manga page may contain multiple panels with distinct compositions, multiple characters with detailed designs, and domain-specific visual conventions.

This complexity creates retrieval challenges at multiple levels. Individual artists develop distinct visual styles that vary substantially across creators and works. Character designs involve specific attributes (hair styles, eye shapes, clothing details) rendered in highly stylized rather than photorealistic ways. Critically, panels frequently contain multiple characters with distinct visual attributes, requiring systems that can distinguish specific entities rather than treating the entire panel as a single unit. Additionally, manga employs stylized visual conventions (screentones, speed lines, symbolic backgrounds) that differ significantly from photographic imagery.

1.3 The Search Problem

Reader preferences in manga are often attribute-focused. Consumers seek content based on specific visual characteristics: character aesthetics, artistic style (e.g., detailed vs. minimalist, realistic vs. chibi), visual tone, or compositional preferences. These preferences parallel problems in domains like fashion e-commerce, where users search by garment attributes; however, manga's stylized multi-entity nature makes the challenge more complex.

The core retrieval question is: *How can we find manga pages with similar visual attributes to a given reference, without requiring manual annotation?* While the broader goal is attribute-focused retrieval across multiple dimensions, we focus specifically on character design as an initial exploration.

Traditional recommendation systems based on collaborative filtering or metadata tags (genres, demographics, publication year) fail to capture these fine-grained visual preferences. A reader who enjoys the character designs in one manga may seek similar aesthetics elsewhere, but existing discovery tools cannot bridge this gap based on visual content alone.

1.4 Limitations of Supervised Learning

Supervised approaches to this problem face significant barriers. Training models for attribute-based manga retrieval would require large-scale annotated datasets containing detailed character-level attribute labels for every character in thousands of panels. This presents multiple challenges:

Annotation cost. The scale required makes manual annotation impractical. Each page may contain multiple characters across multiple panels, and attributes must be labeled consistently across different poses, angles, and artistic renderings.

Combinatorial complexity. With multiple attributes per character and multiple characters per panel, the space of possible attribute combinations grows exponentially.

[1] Code available at: https://github.com/xuyangan/semantic-manga-retrieval

Domain specificity. Manga's unique visual conventions mean that models trained on photographic imagery may not transfer effectively. While datasets like Manga109 [4] exist for detection tasks, they do not provide the attribute-level annotations needed for retrieval, and creating such annotations at scale is unfeasible.

These challenges render supervised learning an impractical solution for scalable manga retrieval.

1.5 Pretrained Vision-Language Models and CLIP

The limitations of supervised approaches have driven interest in leveraging pretrained vision-language models for retrieval. Contrastive Language-Image Pre-training (CLIP) [7], a model pretrained on image-text pairs, provides a shared embedding space where images and text describing similar concepts lie close together. This enables retrieval via similarity search without task-specific training.

CLIP's success in domains like fashion e-commerce raises a question: can frozen CLIP models transfer to manga? Manga differs substantially from CLIP's training distribution—it is stylized rather than photographic, multi-entity rather than single-object focused, and governed by artistic conventions rather than physical realism.

1.6 Research Questions

This work investigates character-based manga retrieval using CLIP with LLM-generated character descriptions. We ask:

(1) Can frozen CLIP encoders combined with character descriptions generated by Large Language Model (LLM) enable effective character-based manga retrieval?
(2) How does the integration strategy for semantic information (late fusion reranking versus query refinement) affect retrieval performance when targeting specific characters in multi-character panels?
(3) Does removing background context through character isolation improve retrieval compared to using full manga pages?

Scope: We focus on character design as the primary retrieval attribute. We use frozen pretrained models without fine-tuning or domain adaptation. This allows us to isolate the effect of the semantic integration strategy and dataset configuration.

1.7 Contributions

This work contributes the following:

(1) **Cross-domain manga retrieval evaluation**: We demonstrate the feasibility of character-based manga retrieval using frozen CLIP and LLM-generated descriptions without domain-specific training.
(2) **Query-Conditioned Feedback Retrieval (QCFR)**: We propose a two-pass retrieval method that uses semantic information to refine the query embedding before re-retrieval, allowing semantic intent to influence which pages are retrieved rather than only their ranking.
(3) **Comparison of integration strategies**: We compare query refinement (QCFR) against late fusion reranking to isolate the effect of *when* semantic information is applied.
(4) **Character isolation analysis**: We evaluate whether removing background context through automatic character detection improves retrieval compared to full-page representations.

2 Related Work

2.1 Manga Analysis

Computational manga analysis has historically focused on element detection tasks, including panel detection [5], text and balloon detection [4], character detection [4], and scene narration [10]. These approaches typically rely on supervised training, with the most prominent methods utilizing the Manga109 dataset [4]. Manga109 consists of over 21,000 images with annotations for panels, text blocks, characters (face and body), and text-speaker associations.

However, Manga109 has limitations for our retrieval task. First, it provides object-level annotations (bounding boxes) but lacks the fine-grained attribute descriptions (e.g., hair style, clothing details) needed for attribute-based retrieval. Second, supervised models trained on 109 manga titles may not generalize to thousands of other manga with divergent artistic styles. Third, creating similar annotations for attribute-based retrieval at scale would be prohibitively expensive, negating the benefits of an annotation-free pipeline.

These limitations motivate our approach using frozen pretrained models: rather than requiring extensive labeled data, we leverage pretrained vision-language models and LLM-generated descriptions to enable retrieval without domain-specific annotation or fine-tuning. To our knowledge, no prior work has investigated attribute-based retrieval in manga, nor has compositional image retrieval using frozen pretrained models been evaluated in this domain.

2.2 Compositional Image Retrieval

Compositional image retrieval (CIR) addresses the task of finding target images given a reference image and a natural language modification. Recent surveys [14] provide comprehensive overviews of the field.

Supervised Methods. Vo et al. [15] introduced the Text-Image Residual Gating (TIRG) framework, which uses gating mechanisms to combine image features with text-derived residuals. This approach requires expensive triplet annotations (reference image, text modification, target image). Similarly, Chen et al. [1] proposed attention mechanisms to focus on specific image regions relevant to the text modification. While effective, these supervised approaches face the same annotation barriers discussed above.

Training-Free Methods. Recent work has explored training-free alternatives using pretrained vision-language models without task-specific fine-tuning. Saito et al. [12] introduced Pic2Word, which maps reference images to learnable pseudo-word tokens in CLIP's text embedding space. Karthik et al. [2] proposed CIReVL, which operates in language space: captioning the reference image, using an LLM to rewrite the caption incorporating the modification, and retrieving based on the edited caption.

These training-free methods have demonstrated success in fashion and general image domains, but their applicability to manga remains unexplored. Our work extends this line of research by

evaluating frozen CLIP models for manga character retrieval and comparing different strategies for integrating semantic information.

2.3 Vision-Language Models

CLIP [7] has emerged as the foundation for modern visual tasks requiring no task-specific training. Trained via contrastive learning on 400M image-text pairs, it creates a shared multimodal embedding space. Beyond retrieval, CLIP's embedding space supports semantic arithmetic and attribute-based manipulation. Patashnik et al. [6] demonstrated that meaningful attribute directions exist in CLIP's latent space, enabling text-guided image editing through embedding manipulation.

Prior studies examining pretrained vision-language models in non-photographic domains have reported mixed results. Research on sketches [11], cartoons, and artwork shows that CLIP's performance can degrade due to domain shift: the model's training distribution, dominated by photographic images, may not adequately represent stylized visual conventions.

Manga represents an extreme case of stylization, combining abstract character designs, symbolic visual elements (e.g., speed lines, emotion indicators), and dense multi-panel compositions with multiple co-occurring characters. The success of character-centric retrieval methods in manga would demonstrate robustness to significant domain shift and validate the cross-domain transfer capability of vision-language models to highly stylized visual domains beyond their training distribution.

2.4 Multimodal Fusion Strategies

The question of when and how to combine multimodal signals has been extensively studied. Wang et al. [16] provide a survey of fusion strategies in cross-modal retrieval, categorizing approaches into early fusion (combining features before encoding), late fusion (combining scores after retrieval), and hybrid methods. Their analysis shows that no single strategy dominates universally, with performance varying significantly across domains and tasks.

Part of our work examines the effectiveness of late fusion reranking in the context of character-centric manga retrieval. While late fusion has become a standard approach in many multimodal retrieval systems, its application to manga search remains underexplored. To our knowledge, this is the first evaluation of late fusion strategies specifically for character-based manga retrieval, comparing image-first and text-first initialization approaches across different hyperparameter configurations. Additionally, we propose Query-Conditioned Feedback Retrieval (QCFR), which integrates semantic information at the query representation level rather than only at the scoring stage, enabling semantic intent to influence candidate generation itself.

2.5 Relevance Feedback and Query Refinement

Relevance feedback is a classical paradigm in information retrieval for refining queries based on user intent. One of the earliest and most influential methods is the Rocchio algorithm [8], which updates a query vector by moving it closer to the centroid of relevant documents and farther from non-relevant ones. This formulation established the principle that modifying the query embedding itself—rather than only adjusting document scores—can significantly impact retrieval quality.

Building on this idea, pseudo-relevance feedback (PRF) methods assume that the top-ranked results from an initial retrieval are relevant and use them to automatically refine the query [3]. PRF has been widely adopted in text retrieval systems and later extended to multimedia retrieval through feature averaging, metric learning, and interactive feedback mechanisms [9]. However, many of these approaches either rely on explicit user feedback, supervised training, or operate within a single modality, limiting their applicability in fully automatic and training-free settings.

Recent vision-language retrieval methods implicitly revisit query refinement under pretrained multimodal embedding spaces. Approaches such as CIReVL [2] modify queries by rewriting captions with large language models, while others incorporate semantic constraints through late fusion reranking. These methods typically inject semantic information after candidate generation, treating it as a post-hoc adjustment rather than modifying the retrieval query itself.

Our proposed Query-Conditioned Feedback Retrieval (QCFR) draws inspiration from classical relevance feedback while adapting it to a training-free, multimodal setting. Instead of relying on labeled relevance judgments, QCFR uses LLM-generated semantic descriptions to construct query-conditioned semantic directions within the frozen CLIP embedding space. By updating the query representation prior to retrieval, QCFR allows semantic intent to influence candidate generation directly, rather than solely affecting score fusion or reranking. This bridges established query refinement principles with modern vision-language models in a highly stylized domain where supervised annotations are scarce.

3 Dataset

3.1 Dataset Construction

Manga Selection and Sampling. We compiled a dataset of 630 manga pages drawn from 28 distinct manga titles. For five titles, we selected 50 pages each, including 20 pages containing a target character and 30 pages in which the character does not appear. These same-style negative examples share identical artistic style but exclude the target character, encouraging retrieval models to move beyond style-based matching. The remaining titles provide additional diversity in artistic style, genre, and character design. Pages were sampled across different narrative arcs to capture variation in character appearance while preserving identity consistency.

Query Selection. For each of the five curated titles, one character was designated as the target. The 20 pages in which the character appears serve both as query images and ground truth, yielding 100 total queries. To support baseline analysis, we additionally generated character-specific cropped images for each page, which are used as alternative image query variants.

3.2 Ground Truth

A key challenge in manga retrieval evaluation is the absence of existing labeled datasets. We leverage a domain-specific assumption

of manga: within-manga consistency. Pages from the same manga title share character designs, artistic style, and narrative context.

Ground truth definition. For a query targeting a specific character from manga M, relevant pages are defined as only those from M containing the target character. Pages from other manga titles are considered non-relevant, even if they contain visually similar characters. This evaluation framework tests whether systems can identify character identity within a consistent artistic style, rather than simply matching visual similarity across different styles.

Manual annotation. For each of the 5 query characters, we manually annotated which of the 50 corresponding pages contain that character, providing binary relevance labels. This yields 20 positive examples (pages with the character) and 30 negative examples (same-style pages without the character) per query character.

Figure 1 shows a t-SNE visualization of our dataset with pages labeled by manga title. The clear separation between different manga demonstrates distinct artistic styles across titles. This clustering confirms that visual features alone can distinguish between manga, highlighting the challenge of character-centric retrieval within visually cohesive styles—systems must identify specific characters rather than simply recognizing artistic style.

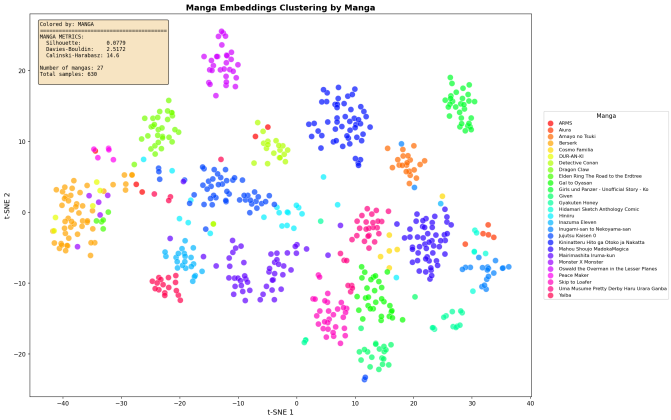


Figure 1: t-SNE visualization of manga pages colored by title. Clear clustering by manga demonstrates distinct artistic styles across the dataset.

4 Proposed Approach

We use OpenCLIP with the ViT-L/14 architecture and `laion2b_s32b_b82k` pretrained weights, producing 768-dimensional embeddings for both images and text. We use Claude Sonnet 4.5 to transform user queries into structured descriptive schemas.

4.1 Offline Embedding Representation

Before retrieval, we precompute representations for all manga pages (or panels) in the database. For each page i, we construct:

- A single CLIP image embedding $\mathbf{v}_i \in \mathbb{R}^{768}$ computed from the full page image.
- A set of LLM-generated, attribute-only character descriptions extracted from the page image. Descriptions are generated as separate lines (one line per character when applicable), using a fixed schema ordering.
- CLIP text embeddings $\{\mathbf{t}_{i,k}\}_{k=1}^{K_i} \subset \mathbb{R}^{768}$, one embedding per description line, where K_i varies across pages depending on the number of valid description lines produced and retained.

Each text line is intended to fit within CLIP's maximum text sequence length (77 tokens). In practice, we explicitly measure token counts using the CLIP tokenizer and record truncation statistics; lines exceeding the limit are truncated by the tokenizer at encoding time.

All embeddings are ℓ_2-normalized. Similarity between embeddings is computed using cosine similarity, which for normalized vectors reduces to the dot product:

$$\text{sim}(\mathbf{a}, \mathbf{b}) = \mathbf{a}^\top \mathbf{b}, \qquad \|\mathbf{a}\| = \|\mathbf{b}\| = 1. \tag{1}$$

This aligns with CLIP's contrastive objective and enables efficient maximum inner-product search in FAISS after normalization.

4.2 Experimental Settings

To disentangle the contribution of visual and semantic cues, we evaluate a set of baselines and ablations that systematically remove or isolate different sources of information. All experiments use the same OpenCLIP backbone; differences arise solely from how representations are constructed and queried.

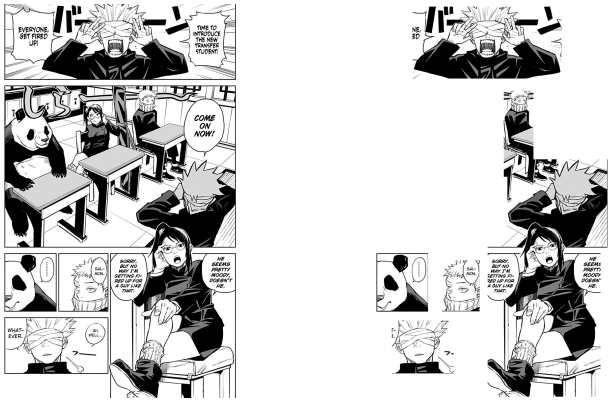
4.2.1 ***Dataset Configurations.*** We consider three dataset configurations that progressively restrict the available visual signal and emphasize character-specific information:

(1) **Baseline (Full Pages).** CLIP image embeddings are computed from complete manga pages. This setting preserves all visual information, including character appearance, background details, panel layout, and overall artistic style, and serves as the strongest image-only baseline.
(2) **Character Isolation.** We apply an external character detection model (Magi [10]) to obtain bounding boxes for characters on each page. All pixels outside the detected character regions are masked prior to encoding, suppressing background content, panel structure, and stylistic context. This configuration isolates character appearance and evaluates whether retrieval can succeed without relying on page-level composition or manga-wide stylistic cues.
(3) **Text Embeddings.** Pages are represented exclusively by LLM-generated, attribute-only character descriptions encoded using the CLIP text encoder. This setting removes visual input entirely and provides a purely semantic representation based on structured character attributes.

Figure 2 illustrates the difference between baseline full pages and character isolation.

4.2.2 ***Query Variants.*** We evaluate two query formulations that differ in the amount of visual context provided at query time:

(1) **Full Page Queries.** The query consists of a complete manga page containing the target character, accompanied by a natural-language instruction that specifies which character to attend to (e.g., location-based cues such as "the character



(a) Full page (b) Background masked

Figure 2: Dataset configuration examples. (a) A full page image. (b) A character isolated image.

in the top panel"). This setting reflects realistic user queries but requires the model to disambiguate the target character from surrounding context.

(2) **Cropped Character Queries.** The query consists of a tightly cropped image containing only the target character. This formulation provides a maximally specific visual signal.

As illustrated in Figure 3, we compare two image query formulations: (a) a full manga page and (b) a tightly cropped character image. This comparison isolates the effect of visual context at query time and evaluates whether removing background, panel structure, and surrounding characters improves retrieval performance for character-specific queries.

For all remaining comparisons, we use the full-page image as the visual query and pair it with a natural-language user instruction specifying the target character. This instruction is processed by the LLM to generate an attribute-only descriptive query, which is then encoded using the CLIP text encoder and used as the semantic query representation.

These query variants are processed using the query construction pipeline described in the following subsection.

4.3 Query Processing

Given a query image and a natural language instruction, we construct both visual and semantic query representations as follows:

(1) A vision-capable LLM produces a structured, attribute-only description of the target character, constrained to the CLIP text encoder's maximum sequence length.
(2) The description is encoded using the CLIP text encoder to obtain a semantic query embedding $\mathbf{t}_q \in \mathbb{R}^{768}$.
(3) The query image is encoded using the CLIP image encoder to obtain an image query embedding $\mathbf{q} \in \mathbb{R}^{768}$.
(4) Both $\mathbf{q}$ and $\mathbf{t}_q$ are ℓ_2-normalized prior to retrieval.

4.4 Semantic Integration Strategies

We compare two strategies for incorporating semantic information into the retrieval pipeline. The strategies differ only in *when* semantic signals are applied: either after retrieval at the score level (Late Fusion), or prior to a second retrieval pass via query refinement (QCFR).

4.4.1 ***Late Fusion (Score-Level).*** Late fusion utilizes a two-stage approach where one modality acts as the primary retrieval signal and the other serves as a post-retrieval reranker. For instance, in image-first late fusion, an initial candidate set is retrieved based solely on visual similarity by querying the index with the image embedding $\mathbf{q}$. Subsequently, semantic information from text embeddings are applied exclusively to reorder this fixed candidate set.

LLM-Based Text Representations. For each page i in the database, we generate multiple character-level textual descriptions using an LLM conditioned on the page image. Each description is independently encoded using the CLIP text encoder, yielding a set of text embeddings $\{\mathbf{t}_{i,1}, \ldots, \mathbf{t}_{i,K_i}\}$ associated with page i. This representation allows a single page to be described by multiple semantic descriptors corresponding to different characters or character attributes.

To aggregate multiple textual descriptors per page, we define a semantic compatibility score via max pooling:

$$s_i = \max_k \langle \mathbf{t}_q, \mathbf{t}_{i,k} \rangle, \tag{2}$$

where $\mathbf{t}_q$ denotes the semantic query embedding derived from the LLM-generated query description.

Image-First Late Fusion. In the image-first variant, candidate pages are retrieved by querying the image index with $\mathbf{q}$. For each retrieved page i, visual and semantic similarity are combined using a convex combination:

$$\text{score}^{\text{img}}_{\text{late}}(i) = \alpha \langle \mathbf{q}, \mathbf{v}_i \rangle + (1 - \alpha) s_i, \tag{3}$$

where $\mathbf{v}_i$ denotes the image embedding of page i and $\alpha \in [0, 1]$ controls the relative contribution of visual and semantic similarity. Candidates are reranked according to this fused score.

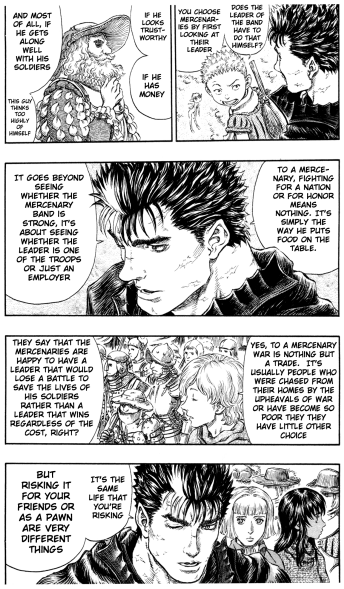
Text-First Late Fusion. In the text-first variant, retrieval is initialized using semantic similarity by querying the text index with $\mathbf{t}_q$. Text-level results are deduplicated at the page level by retaining the maximum similarity per page, yielding the same semantic score s_i. For each resulting page i, the corresponding image embedding is retrieved and combined as:

$$\text{score}^{\text{text}}_{\text{late}}(i) = \alpha \langle \mathbf{q}, \mathbf{v}_i \rangle + (1 - \alpha) s_i. \tag{4}$$

Pages are subsequently ranked by this combined score.

Discussion. In both late fusion variants, semantic information influences only the final ordering of candidates. The candidate set itself is fixed by the initial retrieval stage, and pages not retrieved in the first stage cannot be introduced through late fusion.

4.4.2 ***Query-Conditioned Feedback Retrieval (Proposed).*** In contrast to late fusion, we propose *Query-Conditioned Feedback Retrieval (QCFR)*, a two-pass retrieval strategy in which semantic information is used to refine the query representation prior to a



(a) Full manga page used as the image query.

(b) Tightly cropped image of the target character extracted from the full page.

I want to find a manga with a character similar to the one-eyed male character in the bottom panel

(c) Example user instruction indicating the character of interest within the page.

Male, young adult, athletic build, medium-length spiky dark hair swept upward, defined angular facial features with strong jawline, one eye closed in wink expression, wearing dark high-collared garment, bold high-contrast visual style with heavy black ink work

(d) Attribute-only character description generated by the LLM from the user instruction.

Figure 3: Query formulations and semantic processing pipeline. (a) A full manga page paired with a natural-language instruction specifying the target character. (b) A cropped character image used as a maximally specific visual query. (c) An example user instruction describing the character of interest. (d) The corresponding attribute-only character description generated by the LLM and subsequently encoded for semantic retrieval.

second retrieval pass. This allows semantic intent to influence not only ranking but also which pages are retrieved.

As in late fusion, we begin by encoding the query image and generating an LLM-based semantic description, yielding embeddings $\mathbf{q}$ and $\mathbf{t}_q$. An initial candidate pool is formed by querying both the image index with $\mathbf{q}$ and the text index with $\mathbf{t}_q$. For each candidate page i, we compute a hybrid compatibility score:

$$s_i = \alpha \langle \mathbf{q}, \mathbf{v}_i \rangle + (1 - \alpha) \max_k \langle \mathbf{t}_q, \mathbf{t}_{i,k} \rangle. \tag{5}$$

The highest-scoring pages are treated as pseudo-positive examples, while the lowest-scoring pages are treated as pseudo-negative examples. Let $\mathcal{P}$ and $\mathcal{N}$ denote the corresponding sets. Using the hybrid scores s_i, we compute feedback centroids in the image embedding space:

$$\mathbf{c}^+ = \text{norm}\left(\sum_{i \in \mathcal{P}} w_i^+ \mathbf{v}_i \right), \qquad \mathbf{c}^- = \text{norm}\left(\sum_{i \in \mathcal{N}} w_i^- \mathbf{v}_i \right). \tag{6}$$

The weights are derived from the hybrid scores and normalized within each set. For pseudo-positive pages, weights are computed using a softmax over scores:

$$w_i^+ = \frac{\exp(s_i)}{\sum_{j \in \mathcal{P}} \exp(s_j)}, \tag{7}$$

assigning greater influence to pages that align strongly with both the visual query and semantic intent.

For pseudo-negative pages, weights are computed using a softmax over negated scores:

$$w_i^- = \frac{\exp(-s_i)}{\sum_{j \in \mathcal{N}} \exp(-s_j)}, \tag{8}$$

emphasizing pages that are most incompatible with the query.

The query embedding is then refined via a Rocchio-style update:

$$\mathbf{q}' = \text{norm}\big(a\,\mathbf{q} + b\,\mathbf{c}^+ - c\,\mathbf{c}^- + d\,\mathbf{t}_q \big), \tag{9}$$

where a, b, c, d are fixed coefficients controlling the influence of the original query, positive feedback, negative feedback, and semantic intent.

Finally, a second nearest-neighbor search is performed over the full image index using the refined query $\mathbf{q}'$:

$$\text{score}_{\text{QCFR}}(i) = \langle \mathbf{q}', \mathbf{v}_i \rangle. \tag{10}$$

By refining the query representation and re-running retrieval, QCFR is not restricted to reordering an existing candidate set. Pages absent from the initial retrieval may be introduced in the final results, allowing semantic intent to reshape the retrieved region of the embedding space rather than merely adjust ranking scores.

5 Evaluation Protocol

5.1 Evaluation Metrics

We evaluate retrieval performance using standard information retrieval metrics:

- **Recall@K**: The proportion of relevant pages retrieved among the 20 ground truth pages within the top-K results.

We report Recall@10, Recall@20, Recall@30, Recall@40, and Recall@50. Since each query has 20 relevant pages, the maximum achievable Recall@10 is 0.5.

- **Mean Average Precision@K (MAP@K)**: The mean of average precision scores computed across all queries, where average precision is defined as the mean of precision values at each rank position where a relevant item is retrieved. The same K values as in Recall@K are used.

5.2 Evaluation Protocol

For each query:

(1) Provide the query image (containing the target character) and an LLM-generated character description.
(2) Retrieve pages from the database (500 pages total across all manga).
(3) Rank retrieved results by similarity score.
(4) Compare rankings against ground truth annotations to compute metrics.

Results are averaged across all queries.

5.3 Baseline and Ablation Studies

To understand the contribution of each component, we evaluate:

- **Image-only baseline**: CLIP image embeddings with cosine similarity.
- **Image-only on isolated characters**: Use character detection from Magi [10] to detect character bounding boxes, mask the background, and query using embeddings of these isolated character images.
- **Image-only on padded queries**: Pad query images to square aspect ratio and query using image embeddings.
- **Text-only baseline**: Character description embeddings only.
- **Score-level fusion (Image First)**: Image retrieval followed by reranking using text embeddings.
- **Score-level fusion (Text First)**: Text retrieval followed by reranking using image embeddings.
- **Representation-level fusion**: Query-conditioned embedding editing before retrieval.

6 Results and Discussions

6.1 Baselines

We present the retrieval performance across all queries and manga in Table 1 for Recall@k and Table 2 for mAP@k. Method names are abbreviated as follows:

- **Full Page, Full Page**: Full page query images searched against a database of full manga page embeddings.
- **Full Page, Masked Page**: Full page query images searched against character-isolated embeddings (background masked).
- **Cropped, Full Page**: Cropped character query images searched against full manga page embeddings.
- **Cropped, Masked Page**: Cropped character query images searched against character-isolated embeddings.
- **Text-to-Text**: LLM-generated character descriptions searched against text embeddings of all characters in the database.

Table 1: Recall@k performance across all baseline methods (mean values).

Method (Query, Index)	@10	@20	@30	R@40	@50
Full Page, Full Page	0.338	0.548	0.671	**0.751**	**0.804**
Full Page, Masked Page	0.316	0.487	0.587	0.651	0.700
Cropped, Full Page	0.362	0.572	**0.682**	0.750	0.798
Cropped, Masked Page	**0.380**	**0.581**	0.677	0.740	0.783
Text-to-Text	0.172	0.270	0.338	0.400	0.456

Table 2: mAP@k performance across all baseline methods (mean values).

Method (Query, Index)	@10	@20	@30	@40	@50
Full Page, Full Page	0.307	0.449	0.512	0.546	0.564
Full Page, Masked Page	0.278	0.384	0.428	0.451	0.465
Cropped, Full Page	0.328	0.480	0.539	0.566	**0.583**
Cropped, Masked Page	**0.346**	**0.491**	**0.541**	**0.567**	0.582
Text-only (LLM to LLM)	0.124	0.164	0.186	0.200	0.213

Figure 4 presents a visual comparison of these Recall@k and mAP@k metrics across all baseline configurations. The visualization illustrates the relative performance of each method as the number of retrieved results (k) increases.

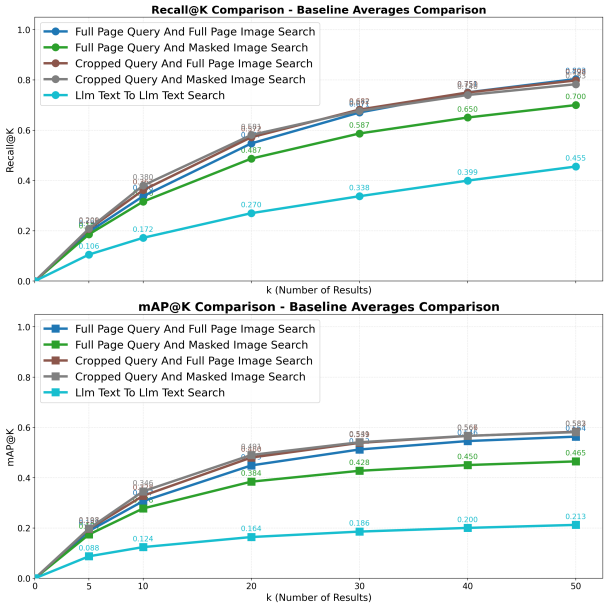


Figure 4: Recall@k and mAP@k performance comparison across baseline methods.

Discussion. Both metrics show consistent improvement as k increases. Full page and cropped query search perform nearly on par, with cropped queries showing slight advantages at lower k values. Character-isolated (masked) methods perform moderately worse than full page search, suggesting that contextual visual information aids retrieval. Text-only search shows significantly lower

performance, highlighting the importance of visual features for character-centric retrieval.

6.2 Late Fusion Reranking with Image-First and Text-First Retrieval

We evaluate two Late Fusion Reranking strategies that differ in their initial retrieval stage:

(1) **Image-First Reranking**: Retrieve candidates using image similarity, then rerank using text embeddings.
(2) **Text-First Reranking**: Retrieve candidates using text similarity, then rerank using image embeddings.

Both methods use full page queries rather than cropped character queries. While cropped queries achieve higher precision and comparable recall performance in image-only retrieval, (Table 1 and Table 2), full page queries allow users to specify additional contextual attributes beyond individual characters, providing greater flexibility in real-world retrieval scenarios.

6.2.1 ***Hyperparameter Search***. We perform a grid search over two key hyperparameters:

- $\alpha \in [0.1, 1.0]$: Controls the weight between image and text similarity scores in the fusion ***Equation 2*** and ***Equation 3***.
- $M \in \{50, 100, 200\}$: Number of initial candidates retrieved before reranking.

Some results for α are omitted as they consistently yielded lower performance across all metrics. Tables 3 and 4 present results for the remaining α values at $M = 100$, and Tables 5 and 6 show the effect of varying M at the optimal $\alpha = 0.8$.

Table 3: Recall@k performance for score-level fusion methods across different α values at $M = 100$.

Method	@10	@20	@30	@40	@50
Reranking (Image First)					
$\alpha = 0.7$	0.359	0.582	**0.711**	0.788	0.828
$\alpha = 0.8$	**0.365**	**0.590**	0.710	**0.791**	**0.838**
$\alpha = 0.9$	0.363	0.576	0.704	0.781	0.827
Reranking (Text First)					
$\alpha = 0.7$	0.344	0.500	0.562	0.580	0.590
$\alpha = 0.8$	0.347	0.503	0.563	0.582	0.591
$\alpha = 0.9$	0.346	0.497	0.558	0.583	0.591

6.2.2 ***Comparison with Full-Image Query Baselines***. Figure 5 visualizes the performance differences between single-modality baselines (image-only and text-only retrieval) and late fusion reranking methods that combine both modalities after an initial retrieval stage.

While $M = 100$ yielded the best performance, we also include results for $M = 50$ to enable direct comparison with the baseline methods, which retrieve at most 50 entries for evaluation at $k \leq 50$.

Discussion. The main findings are:

- Both reranking strategies demonstrate substantial improvements over their respective single-modality baselines across all k values.

Table 4: mAP@k performance for score-level fusion methods across different α values $M = 100$.

Method	@10	@20	@30	@40	@50
Reranking (Image First)					
$\alpha = 0.7$	0.329	0.488	0.558	0.592	0.606
$\alpha = 0.8$	**0.338**	**0.501**	**0.567**	**0.602**	**0.618**
$\alpha = 0.9$	0.334	0.487	0.555	0.589	0.605
Reranking (Text First)					
$\alpha = 0.7$	0.314	0.423	0.456	0.462	0.465
$\alpha = 0.8$	0.319	0.427	0.459	0.466	0.468
$\alpha = 0.9$	0.319	0.424	0.456	0.465	0.468

Table 5: Recall@k performance for score-level fusion with different M values at $\alpha = 0.8$ (mean values).

Method	@10	@20	@30	@40	@50
Reranking (Image First)					
$M = 50$	**0.366**	**0.591**	0.709	0.776	0.804
$M = 100$	0.365	0.590	**0.710**	**0.791**	**0.838**
$M = 200$	0.365	**0.591**	**0.710**	0.789	0.837
Reranking (Text First)					
$M = 50$	0.313	0.414	0.432	0.435	0.435
$M = 100$	0.347	0.503	0.563	0.582	0.591
$M = 200$	0.364	0.556	0.657	0.699	0.724

Table 6: mAP@k performance for score-level fusion with different M values at $\alpha = 0.8$.

Method	@10	@20	@30	@40	@50
Reranking (Image First)					
$M = 50$	**0.338**	0.501	0.565	0.595	0.604
$M = 100$	**0.338**	0.501	**0.567**	**0.602**	**0.618**
$M = 200$	**0.338**	**0.502**	**0.567**	0.601	**0.618**
Reranking (Text First)					
$M = 50$	0.286	0.353	0.361	0.363	0.363
$M = 100$	0.319	0.427	0.459	0.466	0.468
$M = 200$	0.336	0.474	0.528	0.546	0.555

- Image-first reranking achieves the highest overall performance, suggesting that initializing with visually similar candidates and refining with semantic information provides the most effective retrieval strategy for character-centric manga search.
- The optimal hyperparameters are $\alpha = 0.8$ and $M \in \{100, 200\}$. The high α value indicates that image embeddings contribute more strongly to the final ranking than text embeddings, reflecting the importance of visual features for character recognition in manga retrieval.

6.3 Query-Conditioned Feedback Retrieval (QCFR)

We evaluate the proposed Query-Conditioned Feedback Retrieval (QCFR) framework and compare it against late fusion reranking strategies. QCFR differs fundamentally from late fusion in that

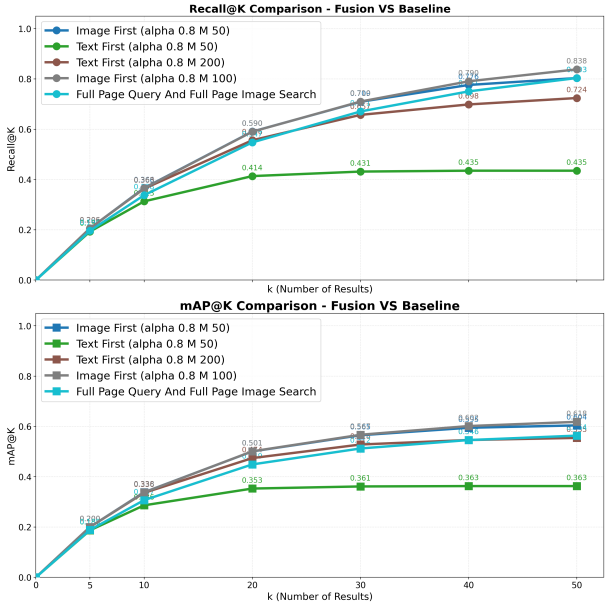


Figure 5: Recall@k and mAP@k performance comparison between single-modality baselines and late fusion reranking methods. Both image-first and text-first reranking strategies demonstrate substantial improvements over the image-only baseline, with image-first reranking achieving the highest performance in most of the k values.

semantic information is used to refine the query representation prior to a second retrieval pass, rather than only reordering a fixed candidate set.

All QCFR experiments use full-page image queries paired with LLM-generated semantic descriptions, consistent with the late fusion setup.

6.3.1 ***Hyperparameter Search.*** QCFR introduces additional hyperparameters beyond those used in score-level fusion:

- M_{img}: Number of image-based candidates retrieved during the first retrieval pass from the image index.
- M_{txt}: Number of text-based candidates retrieved during the first retrieval pass from the text index.
- α: Controls the relative contribution of visual similarity and semantic similarity in the hybrid compatibility score used to rank first-pass candidates and select pseudo-positive and pseudo-negative samples ***(Equation 5)***.
- l_{positive}: Number of top-ranked pages selected as pseudo-positive examples for constructing the positive feedback centroid.
- l_{negative}: Number of lowest-ranked pages selected as pseudo-negative examples for constructing the negative feedback centroid.
- (a, b, c, d): Coefficients in the Rocchio-style query update that weight the original query, positive feedback, negative feedback, and semantic intent, respectively ***(Equation 9)***.

For simplicity and to reduce the number of tunable hyperparameters, we adopt the following fixed settings throughout all QCFR experiments. We retrieve a larger pool of semantic candidates than visual candidates by setting $M_{\text{txt}} = 3M_{\text{img}}$, reflecting the fact that a single page may contain multiple characters. We enforce symmetric feedback by setting $l_{\text{positive}} = l_{\text{negative}}$. In the Rocchio-style query update, we fix the coefficients of the original query and semantic term to $a = 1$ and $d = 0.21$, respectively.

6.3.2 ***QCFR Performance.*** We compare QCFR against the strongest late fusion baselines identified in the previous section, namely image-first reranking with $\alpha = 0.8$ and $M = 100$. All metrics are averaged across queries and reported at the page level.

Table 7: Recall@k performance comparison between QCFR and late fusion baselines ($\alpha = 0.8$, $M = 100$).

Method	@10	@20	@30	@40	@50
Late Fusion (Image First)	0.365	0.590	0.710	0.791	0.838
QCFR (Proposed)	**0.383**	**0.593**	**0.718**	**0.827**	**0.855**

Table 8: mAP@k performance comparison between QCFR and late fusion baselines ($\alpha = 0.8$, $M = 100$).

Method	@10	@20	@30	@40	@50
Late Fusion (Image First)	0.338	0.501	0.567	0.602	0.618
QCFR (Proposed)	**0.352**	**0.506**	**0.574**	**0.615**	**0.631**

QCFR Configuration Details. Unless otherwise stated, all QCFR results are reported using the following parameter configuration. We retrieve $M_{\text{img}} = 100$ image-based candidates and $M_{\text{txt}} = 300$ text-based candidates in the first retrieval pass. The number of pseudo-positive and pseudo-negative samples is set to $l_{\text{positive}} = l_{\text{negative}} = 20$. In the Rocchio-style query refinement step, we set the feedback coefficients to $b = 0.35$ for the positive centroid and $c = 0.30$ for the negative centroid. All QCFR results are obtained using a single refinement iteration, followed by a second nearest-neighbor search over the full image index.

Discussion. QCFR consistently outperforms both late fusion baselines across all Recall@k and mAP@k metrics. The improvements are most pronounced at higher recall cutoffs, indicating that query refinement enables QCFR to recover relevant pages that are absent from the initial retrieval stage. Unlike late fusion, which is restricted to reranking a fixed candidate set, QCFR allows semantic intent to influence candidate generation itself. These results demonstrate that incorporating semantic feedback at the query level yields a more expressive retrieval mechanism for character-centric manga search.

A limitation of QCFR is its increased computational cost relative to late fusion. QCFR requires additional similarity computations, centroid construction, and a second nearest-neighbor search over the full image index, resulting in higher query-time latency. However, this overhead is incurred only at inference time and remains practical for interactive retrieval settings, as all embeddings are precomputed and the refinement process involves a single additional retrieval pass.

6.4 Failure Mode Analysis

We identify and categorize failure modes observed across different methods:

- **CLIP Image Encoder Limitations**: Padding images to square aspect ratios or masking backgrounds can introduce artifacts (e.g., large white regions) that cause the model to focus on these irrelevant features rather than character-specific details. This likely explains the modest performance degradation observed in the character-isolated (masked) configuration compared to full page retrieval. However, the substantial improvements achieved through late fusion reranking demonstrate that text embeddings can provide complementary semantic information that mitigates these visual limitations.
- **CLIP Text Encoder Constraints**: The 77-token limit for CLIP's text encoder severely restricts the richness and detail of character descriptions. Complex characters with multiple distinctive attributes cannot be fully captured within this constraint, limiting the discriminative power of text-only retrieval and potentially affecting the semantic component of late fusion methods.
- **Ambiguous Character Appearances**: Characters with similar visual features (e.g., similar hair color, clothing style) but different identities are difficult to distinguish using visual features alone, leading to false positives in image-based retrieval methods.
- **Artistic Style Dominance**: The strong clustering by manga title (Figure 1) indicates that CLIP embeddings heavily encode artistic style. This can cause the model to retrieve visually similar pages from the same manga even when they do not contain the target character, particularly affecting text-first reranking strategies.

7 Conclusion

This study investigates how automatically generated semantic representations can be used to simplify and enhance multimodal image retrieval when user intent targets specific visual attributes rather than holistic image similarity. By converting user instructions and visual observations into structured, attribute-centric descriptions, semantic representations provide an interpretable abstraction that bridges low-level visual features and high-level retrieval intent. This abstraction allows users to specify fine-grained appearance constraints without requiring precisely cropped or visually optimized queries, while remaining compatible with contrastively trained vision–language embedding spaces.

Our experiments show that semantic information is most effective when integrated at the query level rather than applied solely as a post-retrieval reranking signal. While score-level late fusion improves over single-modality baselines by combining visual and semantic similarities during reranking, it remains constrained by the initial candidate set. In contrast, Query-Conditioned Feedback Retrieval (QCFR) refines the query representation prior to a second retrieval pass, enabling semantic intent to influence candidate generation itself. Across all evaluated metrics, QCFR consistently outperforms both image-first and text-first late fusion strategies, achieving higher Recall@k and mAP@k, with particularly strong gains at larger recall cutoffs.

We evaluate our approach on a custom-curated manga dataset of approximately 600 pages. Although limited in scale and not intended as a benchmark, this dataset provides a controlled environment for character-centric retrieval, where recurring characters exhibit relatively stable visual attributes across pages. This setting allows us to isolate the effect of semantic guidance and query-level feedback without confounding factors related to ambiguous identity or large intra-class variation. Despite the modest dataset size, the consistent improvements observed across multiple retrieval metrics suggest that query-conditioned feedback using LLM is a promising mechanism for fine-grained multimodal search.

8 Future Work

Several directions remain open for extending this work. First, the quality and expressiveness of semantic representations can be further enhanced by incorporating alternative language encoders with higher token capacity. While CLIP's text encoder provides strong alignment with visual embeddings, its fixed 77-token limit constrains the level of detail that can be expressed in semantic descriptions. Integrating language models such as BERT, which support substantially longer token sequences, could enable richer and more precise semantic representations of visual attributes.

One promising direction is to learn a mapping between the semantic embedding space and the CLIP image embedding space, allowing high-capacity language representations to be projected into a shared retrieval space. Such remapping could preserve the fine-grained semantic structure captured by longer textual descriptions while maintaining compatibility with CLIP-based visual retrieval. This approach would enable more expressive semantic queries without sacrificing the efficiency and scalability of contrastive vision–language models.

Beyond language modeling, future work includes scaling QCFR to larger and more diverse datasets, as well as extending the framework to iterative or interactive retrieval scenarios. In particular, the query-conditioned feedback mechanism naturally lends itself to retrieval-augmented pipelines, where retrieved images and their associated semantic representations can be used to refine subsequent queries or support downstream reasoning tasks. Exploring these directions could further improve alignment between user intent and retrieval behavior in complex, multimodal search settings.

References

[1] Yanbei Chen, Shaogang Gong, and Loris Bazzani. 2020. Image Search With Text Feedback by Visiolinguistic Attention Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2998–3008. doi:10.1109/CVPR42600.2020.00307

[2] Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. 2024. Vision-by-Language for Training-Free Compositional Image Retrieval. arXiv:2310.09291 https://arxiv.org/abs/2310.09291

[3] Yuanhua Lv and ChengXiang Zhai. 2009. A comparative study of methods for estimating query language models with pseudo feedback. (2009), 1895–1898. doi:10.1145/1645953.1646259

[4] Toru Ogawa, Atsushi Otsubo, Rei Narita, Yusuke Matsui, Toshihiko Yamasaki, and Kiyoharu Aizawa. 2018. Object Detection for Comics using Manga109 Annotations. arXiv:1803.08670 https://arxiv.org/abs/1803.08670

[5] X. Pang, Ying Cao, Rynson W. H. Lau, and Antoni B. Chan. 2014. A Robust Panel Extraction Method for Manga. In *Proceedings of the 22nd ACM international conference on Multimedia*. https://api.semanticscholar.org/CorpusID:18439714

[6] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. 2021. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. arXiv:2103.17249 https://arxiv.org/abs/2103.17249

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020 https://arxiv.org/abs/2103.00020

[8] J. J. Rocchio. 1971. Relevance Feedback in Information Retrieval. In *The SMART Retrieval System: Experiments in Automatic Document Processing*, Gerard Salton (Ed.). Prentice-Hall.

[9] Yong Rui, T.S. Huang, M. Ortega, and S. Mehrotra. 1998. Relevance feedback: a power tool for interactive content-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology* 8, 5 (1998), 644–655. doi:10.1109/76.718510

[10] Ragav Sachdeva and Andrew Zisserman. 2024. The Manga Whisperer: Automatically Generating Transcriptions for Comics. arXiv:2401.10224 https://arxiv.org/abs/2401.10224

[11] Aneeshan Sain, Ayan Kumar Bhunia, Vaishnav Potlapalli, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. 2022. Sketch3T: Test-Time Training for Zero-Shot SBIR. arXiv:2203.14691 https://arxiv.org/abs/2203.14691

[12] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. 2023. Pic2Word: Mapping Pictures to Words for Zero-shot Composed Image Retrieval. arXiv:2302.03084 https://arxiv.org/abs/2302.03084

[13] Conghao Tom Shen, Violet Yao, and Yixin Liu. 2023. MaRU: A Manga Retrieval and Understanding System Connecting Vision and Language. arXiv:2311.02083 https://arxiv.org/abs/2311.02083

[14] Xuemeng Song, Haoqiang Lin, Haokun Wen, Bohan Hou, Mingzhu Xu, and Liqiang Nie. 2025. A Comprehensive Survey on Composed Image Retrieval. arXiv:2502.18495 https://arxiv.org/abs/2502.18495

[15] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. 2018. Composing Text and Image for Image Retrieval - An Empirical Odyssey. arXiv:1812.07119 https://arxiv.org/abs/1812.07119

[16] Tianshi Wang, Fengling Li, Lei Zhu, Jingjing Li, Zheng Zhang, and Heng Tao Shen. 2024. Cross-Modal Retrieval: A Systematic Review of Methods and Future Directions. arXiv:2308.14263 https://arxiv.org/abs/2308.14263