Weaving Visual Narratives: Agentic Image Bundle Composition Beyond Atomic Visual Matching

Anonymous ACL submission

Abstract

Image retrieval has traditionally been formulated as a point-wise matching problem, where each candidate image is scored in isolation. However, this atomic paradigm fails to capture the complexity of human search intent within personal photo collections, where users often seek compact visual stories bound by structural relations rather than isolated snapshots. To address this limitation, we introduce **Image Bundle Composition (IBC)**, a novel paradigm that shifts the objective from ranking individual images to dynamically composing cohesive image bundles from a massive, unstructured photo pool. Since target bundles are not predefined, IBC presents a severe combinatorial explosion challenge and demands modeling non-decomposable joint relevance. To establish this paradigm, we construct IBCBench, the first IBC benchmark dataset containing 109,467 images and 667 verified queries, built via a semi-automated verification pipeline. Furthermore, we propose **BundleWeaver** , an agentic framework that reformulates IBC as query-conditioned incremental hyperedge discovery. By employing a Large Language Model to adaptively search for missing relational roles and utilizing a Vision-Language Model for whole-bundle verification, BundleWeaver effectively navigates the combinatorial space. Extensive experiments demonstrate that while state-of-the-art embedding models and static decompose-and-rerank paradigms suffer from relational blindness, BundleWeaver achieves substantial performance gains, highlighting the necessity of shifting from atomic scoring to dynamic relational composition. Our dataset and code are available[1].

1 Introduction

Image retrieval has traditionally been formulated as a point-wise matching problem between a textual query and individual visual assets (Li et al., 2026; Xu et al., 2026; Deng et al., 2026). Despite the progress of multimodal embedding models (Xue et al., 2025; Jiang et al., 2024b), the dominant paradigm remains fundamentally atomic: given a query, each candidate image is scored in isolation, yielding a ranked list of independently relevant items.

As illustrated in Figure 1(a), standard evaluation on benchmarks such as MSCOCO (Lin et al., 2014), Flickr30K (Plummer et al., 2015), and MMEB (Jiang et al., 2024b) operate strictly under this atomic assumption. Recently, some approaches (Deng et al., 2026) have extended this paradigm to multi-hop text-to-image retrieval (Figure 1(b)), navigating through the image pool via cross-image evidence discovery. Yet, despite incorporating intermediate visual reasoning, the ultimate retrieval target is still a ranked list of individually scored, disconnected images.

However, this atomic paradigm fails to capture the complexity of human memory and search intent. In real-world scenarios, users rarely recall experiences as isolated snapshots, rather, they remember compact visual stories (Andrews et al., 2011; Kurby and Zacks, 2008). A user might search for the key highlights of a concert night, the gradual transition of a landscape from day to dusk, or a continuous journey spanning multiple landmarks. In such cases, the ideal target is not a single optimal image, nor a disjointed set of individually relevant results, but a cohesive *image bundle* whose members jointly satisfy the query. While the individual images within a bundle may be visually heterogeneous, they are bound together by an underlying relation, *e.g.*, temporal progression, event summarization, or spatial continuity, which renders them meaningful only as a collective whole.

Motivated by this observation, we introduce **Image Bundle Composition (IBC)**, a novel paradigm that shifts the retrieval objective from ranking in-

[1] https://anonymous.4open.science/r/Image-Bundle-Composition-7313

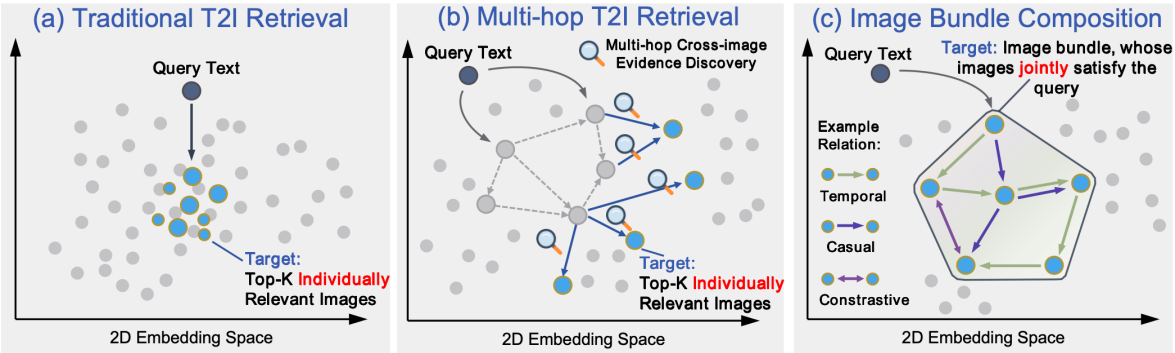


Figure 1: Comparison of different text-to-image retrieval paradigms.

dividual images to dynamically composing cohesive image bundles from a vast, unstructured photo pool. Crucially, these bundles are *not* pre-defined or indexed a priori. As any arbitrary subset of images could theoretically form a bundle, the search space poses a challenge of combinatorial explosion. Therefore, IBC demands more than just matching a description to an image. It requires the system to actively navigate a massive combinatorial space to discover and compose a unique combination of images that collectively instantiates the intended narrative or relational structure.

Establishing this new paradigm necessitates a dedicated benchmark, yet creating a dataset for IBC presents significant challenges. Exhaustive subset annotation is computationally impossible due to the combinatorial nature of the task, while unconstrained discovery using Vision-Language Models (VLMs) often yields bundles that lack uniqueness or structural coherence. To tackle this, we design a semi-automated mining and verification pipeline. By anchoring the search space to spatiotemporal sessions, we extract candidate sliding windows and employ a VLM to rigorously evaluate their cross-image relations. Following this, human annotators refine the VLM-generated queries and filter out ambiguous cases, ensuring that each ground-truth bundle represents a unique, unambiguous answer within the global image pool. Finally, the resulting **IBCBench** consists of 109,467 images and 667 meticulously verified queries.

To tackle the combinatorial search space of IBC, we propose **BundleWeaver**, an agentic framework that reformulates the task as query-conditioned hyperedge discovery. Viewing the unstructured image corpus as a graph of vertices, a target bundle represents a hyperedge connecting a specific subset through a higher-order semantic relation. To avoid exhaustive enumeration, BundleWeaver explores this space incrementally. Starting from diverse seed images, it expands partial bundles by employing a LLM reasoning agent to deduce adaptive search directions, conditioned on both the original query and the current bundle state. Instead of greedily retrieving more locally relevant images, our method deliberately searches for images that fulfill *missing roles* within the ongoing visual composition. Supported by adaptive beam search, contextual candidate pruning, and a whole-bundle VLM reranker, BundleWeaver effectively identifies promising cohesive sets. Empirical results on IBCBench demonstrate that while current state-of-the-art multimodal embeddings and VLMs excel at finding isolated matches, they struggle to compose them logically. BundleWeaver bridges this gap, offering substantial performance gains by explicitly modeling and searching for cross-image compositions.

Our contributions are summarized as follows:

- We formulate **Image Bundle Composition** (IBC), a novel retrieval paradigm that shifts the focus from isolated image matching to the dynamic composition of coherent image bundles.
- We construct and open-source the first IBC benchmark **IBCBench**, comprising 109,467 images and 667 meticulously verified queries, built via a robust semi-automated pipeline.
- We further propose **BundleWeaver**, an innovative agentic framework that models IBC as query-conditioned hyperedge discovery. By incrementally fulfilling missing bundle roles and employing whole-bundle reranking, BundleWeaver demonstrates promising results and sets a strong baseline for the task.

2 Task Formulation

To elucidate the fundamental differences between our proposed task and existing paradigms, we first formalize the traditional image retrieval pipeline,

and then introduce the mathematical formulation of Image Bundle Composition (IBC).

2.1 Preliminaries: Traditional Text-to-Image Retrieval

Let $\mathcal{I} = \{x_1, x_2, \ldots, x_N\}$ denote a large, unstructured image pool of size N, and q denote a natural language query. In traditional T2I retrieval, the objective is formulated as a point-wise matching problem. The system defines a scoring function $f(x, q)$ that independently measures the semantic similarity between the query and each individual candidate image $x \in \mathcal{I}$. The goal is to retrieve an image x^* (or a ranked list of top-k images) that maximizes this independent relevance score:

$$x^* = \arg\max_{x \in \mathcal{I}} f(x, q). \tag{1}$$

In this paradigm, the relevance of any image is atomic and entirely decoupled from the presence or absence of other images in the retrieved list.

2.2 Image Bundle Composition (IBC)

Unlike traditional retrieval, IBC does not seek a single optimal image or a disconnected list of items. Instead, it aims to dynamically compose a cohesive *image bundle*, *i.e.*, a compact subset of images, that collectively fulfills the user's relational or narrative intent.

Formally, let $B \subseteq \mathcal{I}$ denote a candidate image bundle consisting of K distinct images, where the bundle size $K = |B|$ is dynamically determined but bounded by a small integer K_{max} (*e.g.*, $2 \leq K \leq K_{max}$). The objective of IBC is to discover the optimal subset B^* that maximizes a *joint* relevance scoring function $\Phi(B, q)$:

$$B^* = \arg\max_{B \subseteq \mathcal{I}, 2 \leq |B| \leq K_{max}} \Phi(B, q), \tag{2}$$

where $\Phi(\cdot, \cdot)$ evaluates the cross-image composition as a whole. Rather than measuring independent visual-textual alignment, Φ captures the structural, temporal, or spatial relations dictated by the query q across all elements in B.

2.3 Discussions on Challenges of IBC

IBC introduces two profound mathematical and computational challenges that fundamentally distinguish it from standard retrieval:

- **Combinatorial Explosion:** Because bundles are *not predefined or indexed a priori*, the system must actively search over the subsets of $\mathcal{I}$. The size of this search space, denoted as $\mathcal{S}$, scales combinatorially with the pool size N:

$$|\mathcal{S}| = \sum_{k=2}^{K_{max}} \binom{N}{k} \approx \mathcal{O}(N^{K_{max}}). \tag{3}$$

  For a typical personal photo collection where $N \sim 10^4$, evaluating all candidate combinations exhaustively is computationally intractable, making brute-force subset selection impossible.

- **Non-Decomposability of Joint Relevance:** Crucially, the joint relevance function $\Phi(B, q)$ is non-decomposable. This means the bundle-level score cannot be approximated by aggregating individual image-level scores:

$$\Phi(B, q) \neq g\Big(\{f(x_i, q) \mid x_i \in B\}\Big), \tag{4}$$

  where g is any monotonic aggregation operator (*e.g.*, sum or average). For instance, if q asks for *the transition of the Eiffel Tower from day to night*, two visually stunning daytime photos of the tower might each receive high independent scores $f(x, q)$. However, their combination fails to satisfy the *transition* relation, yielding a low joint score $\Phi(B, q)$. Consequently, a greedy strategy that simply selects the top-K independently retrieved images will inherently fail in IBC, necessitating a paradigm shift from independent scoring to relational composition.

We provide more detailed theoretical analysis on limitations of atomic retrieval in IBC in Appendix G.

3 Dataset Construction

Constructing a dataset for IBC is highly challenging. Exhaustive manual annotation is computationally impossible due to combinatorial explosion, while unconstrained VLM generation often yields ambiguous or decomposable image lists. To fill this blank, we introduce **IBCBench**, the first dataset dedicated to IBC. As illustrated in Figure 4, We design a five-stage semi-automated pipeline. Due to space constraints, we outline the core principles here and defer more details to Appendix C.

Property Requirements. A valid candidate subset $B \subset \mathcal{I}$ must satisfy four strict criteria to be accepted: **(C1) Joint Completeness:** Every image plays an indispensable role; **(C2) Cross-image**

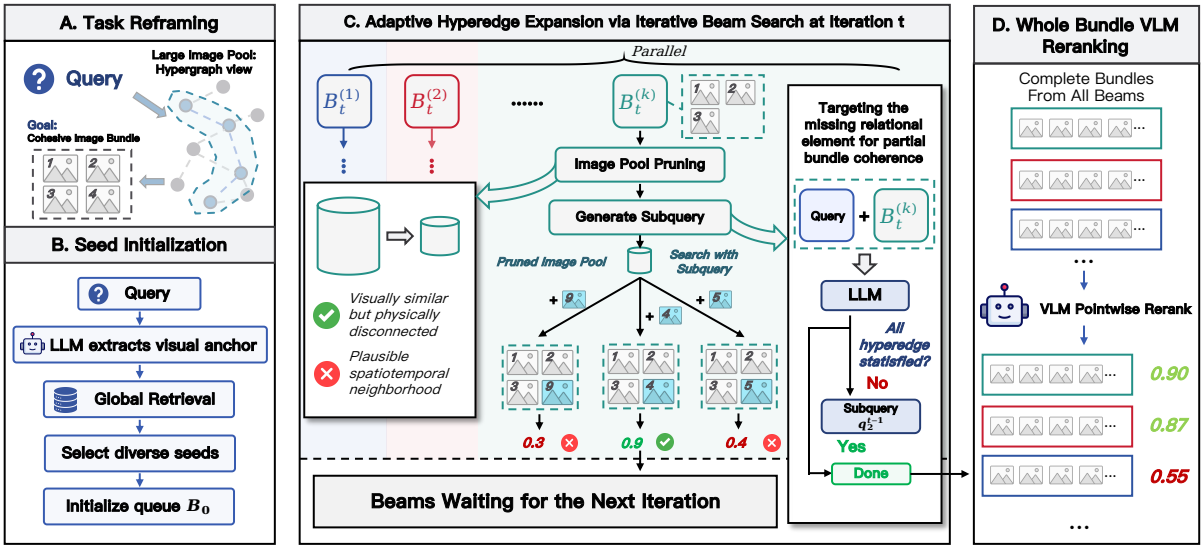


Figure 2: Framework illustration of our proposed BundleWeaver.

Binding: The query enforces a higher-order relation (*e.g.*, temporal progression) rather than merely listing independent visual targets; **(C3) Uniqueness:** The bundle should be unambiguous, and alternatives should be minimized; **(C4) Bounded Redundancy:** Images within B must be visually distinct.

Candidate Mining. We source our raw pool from the YFCC-100M dataset (Thomee et al., 2016), retaining 109,467 photos with valid spatiotemporal metadata. We pre-extract dense captions and tags using GPT-4o and compute multimodal embeddings to penalize redundancy. To bypass the $\mathcal{O}(N^K)$ search space, we group user photos into spatiotemporal sessions and extract sliding windows of size $K \in [3, 5]$. A composite heuristic score aggressively prunes these windows based on spatiotemporal coverage and semantic richness, yielding 7,460 highly diverse candidate windows.

VLM Verification & Human Review. We employ Claude-Opus-4.5 (Anthropic, 2025a) as a rigorous semantic verifier to check candidates against predefined relational templates (*i.e.*, *Same-location Dynamics* and *Cross-location Structures*). The VLM must generate a natural query and articulate a specific cross-image shared anchor. Finally, four expert annotators carefully review the VLM-approved candidates against C1-C4, filtering out ambiguous cases to ensure global uniqueness.

Final Dataset Statistics. Through this exhaustive pipeline (with an end-to-end acceptance rate $< 9\%$), the final IBCBench dataset comprises 667 high-quality, verified queries. Bundle sizes are distributed across 3 (24.3%), 4 (32.2%), and 5 (43.5%) images. The semantic relations are well-balanced, with 52.5% focusing on same-location dynamics and 47.5% capturing cross-location structural ties.

4 Methodology

Addressing IBC by exhaustively evaluating all possible image subsets is computationally intractable. Furthermore, because bundles are dynamically defined by the user's relational intent rather than indexed a priori, traditional atomic retrieval mechanisms cannot be directly applied. To resolve these challenges, we propose **BundleWeaver**, an agentic framework that reformulates IBC as a problem of *query-conditioned hyperedge discovery* and solves it through an adaptive, incremental search process.

As is shown in Figure 2, the BundleWeaver framework consists of three core stages, *i.e.*, Diverse Seed Initialization, Adaptive Hyperedge Expansion via Beam Search, and Whole-Bundle VLM Reranking.

4.1 Reformulation: Incremental Hyperedge Discovery

Let the massive image pool $\mathcal{I}$ be the vertex set $\mathcal{V}$ of an implicit, fully connected hypergraph. A coherent image bundle B corresponds to a hyperedge connecting a small subset of vertices. Since the set of valid hyperedges $\mathcal{E}$ is not predefined, we can only dynamically *construct* the hyperedges from scratch. BundleWeaver models this construction as an incremental pathfinding problem. Starting from a single seed vertex (*i.e.*, an image), the system sequentially discovers new vertices that fulfill miss-

ing narrative or relational roles, eventually closing the hyperedge.

4.2 Diverse Seed Initialization

The discovery process begins by anchoring the search space to a set of highly promising starting vertices. Given the original natural language query q, we prompt a LLM to extract the primary visual anchor and generate an initial search direction d_1. Using a dense multimodal embedding model, we encode d_1 and perform an initial global Nearest Neighbor search across the entire unstructured pool of N images.

To prevent the search from collapsing into a single local optimum or retrieving near-duplicate starting points, we implement a **Diverse Seed Initialization** strategy. From the top-K globally retrieved images, we select a set of N_s seed images that exhibit maximum semantic and visual diversity (*e.g.*, spanning different sub-clusters or contextual sessions). Each selected seed then acts as the independent root for a parallel hyperedge construction tree. This parallelized initialization ensures a broad coverage of candidate starting spaces across the global image pool without loss of generality.

4.3 Adaptive Hyperedge Expansion via Parallel Beam Search

Starting from the initialized diverse seeds, BundleWeaver executes parallel branch expansions, searching for complementary images via a beam search of width B_u. At step k within a specific search branch, let the current partial bundle be $B_{k-1} = \{x_1, x_2, \ldots, x_{k-1}\}$. The expansion involves three key mechanisms:

- **Contextual Candidate Pruning:** Many IBC queries seek cohesive visual narratives grounded in real-world contexts, *e.g.*, *a trip to Florida*. Such queries naturally imply coarse temporal or geographic feasibility constraints. Therefore, rather than exposing the reasoning agent to the entire unconstrained pool, which is filled with visually similar but physically disconnected distractors, we allow it to dynamically bound the candidate expansions x_k to a plausible spatiotemporal neighborhood relative to the seed image. However, this does not by itself solve the task, as images nearby in time or space may still be relationally incompatible. The following stages are therefore essential for composing a coherent bundle.

- **Adaptive Sub-query Generation:** Standard retrieval models lack the reasoning capability to deduce what is missing in a partial bundle. Therefore, we deploy an LLM as an active reasoning agent. Conditioned on the original query q and the visual captions of the already selected images in B_{k-1}, the LLM dynamically generates an adaptive sub-query d_k that specifically targets the missing relational element needed to advance the partial bundle coherence.

- **Path Scoring:** We encode the sub-query d_k and retrieve the top-C local candidates. To determine which branches survive in the beam search, we evaluate each candidate path using a composite scoring function that balances step-wise precision with whole-bundle coherence:

$$\text{Score}(B_k) = \underbrace{\frac{1}{k}\sum_{i=1}^{k}\cos(\mathbf{e}_{d_i}, \mathbf{e}_{x_i})}_{\text{Step-wise Local Match}} + \lambda \cdot \underbrace{\cos\left(\mathbf{e}_q, \frac{\sum_{i=1}^{k}\mathbf{e}_{x_i}}{\left\|\sum_{i=1}^{k}\mathbf{e}_{x_i}\right\|}\right)}_{\text{Holistic Bundle Alignment}} \tag{5}$$

  where $\mathbf{e}_{d_i}$ and $\mathbf{e}_{x_i}$ are the dense embeddings of the i-th sub-query and selected image, respectively, and $\mathbf{e}_q$ is the embedding of the original query. λ controls the trade-off between fulfilling individual sub-queries and maintaining the overall semantic trajectory of the hyperedge.

4.4 Whole-Bundle VLM Reranking

Instead of expanding to a fixed depth, the beam search adaptively terminates when the LLM deems the bundle complete (bounded by a maximum depth K_{max}), yielding a diverse candidate pool. While embedding scores efficiently guide this heuristic search, they often fail to capture fine-grained cross-image relations like strict identity consistency. To this end, we propose **Whole-Bundle VLM Reranking**. We merge all completed paths into a unified pool, where a VLM acts as a pointwise reranker. By simultaneously evaluating all images within a candidate alongside the query q, the VLM scores its structural coherence and cross-image binding (1-10). The highest-scoring hyperedge is returned as B^*. This explicit decoupling of adaptive generation and relational verification efficiently mitigates combinatorial constraints while maximizing final output quality.

5 Experiments

5.1 Experimental Setup

Evaluation Metrics. Unlike traditional ranked lists, IBC outputs dynamic-length cohesive image

Type	Method	Precision	Recall	F1	EM
Multimodal Embedding	CLIP-ViT-B/32	2.21	2.21	2.21	0.00
	SigLIP2-giant	9.97	9.97	9.97	0.15
	Qwen3-VL-Embed-8B	12.14	12.14	12.14	0.00
	RzenEmbed	14.88	14.88	14.88	0.30
Caption + Text Embedding	BM25	3.29	3.29	3.29	0.00
	BGE-M3	7.05	7.05	7.05	0.15
	Qwen3-Embedding-8B	9.32	9.32	9.32	0.30
Heuristic Metadata Augmentation	Session Clustering	4.84	4.84	4.84	0.60
	User+Session Clustering	8.94	8.94	8.94	1.05
	Time Proximity	6.67	6.67	6.67	0.45
	User+Spatiotemporal	8.80	8.80	8.80	0.75
VLM Two-Stage Decompose & Rerank	Qwen2.5-VL-72B	11.39	10.16	10.67	0.15
	Qwen3-VL-235B	15.78	14.35	14.95	0.90
	Gemini-3-Flash	20.89	15.53	17.06	0.60
	GPT-4o	18.50	17.34	17.84	0.60
	Claude-Sonnet-4.5	<u>24.79</u>	<u>23.01</u>	<u>23.74</u>	<u>1.50</u>
BundleWeaver (Ours)		**30.95**	**30.46**	**30.28**	**7.20**
Relative Improvement		24.8%	32.4%	27.5%	380.0%

Table 1: Performance comparison of different methods on IBCBench. Best result is given in bold, and the second best is underlined. *Relative Improvement* is computed against the best baseline result.

sets. Thus, we evaluate performance using set-level *Precision*, *Recall*, and *F1 score*, alongside *Exact Match (EM)*, *i.e.*, the percentage of perfectly predicted ground-truth bundles. The metrics are evaluated at per-query-level and reported with the average.

Baselines. Since IBC is a novel task, we establish comprehensive baselines of four different types:

- **Multimodal Embedding:** Directly retrieves images using state-of-the-art vision-language multimodal embedding models, including CLIP-ViT-B/32 (Radford et al., 2021), SigLIP2-giant (Tschannen et al., 2025), Qwen3-VL-Embedding-8B (Li et al., 2026), and RzenEmbed-7B (Jian et al., 2025).

- **Caption + Text Embedding:** First generates captions for all images using GPT-4o, then performs text-to-text retrieval using BM25 (Robertson, 2025), BGE-M3 (Chen et al., 2024), and Qwen3-Embedding-8B (Zhang et al., 2025).

- **Heuristic Metadata Augmentation:** Applies heuristic re-ranking on multimodal embedding results using metadata, *i.e.*, session clustering, user+session clustering time proximity boosting and user+spatiotemporal.

- **VLM Two-Stage Decompose & Rerank:** A strong agentic baseline where a VLM first decomposes the query into multiple sub-queries, retrieves top-K candidates for each independently, and then reranks the final bundle in the candidate pool. Evaluated VLMs include GPT-4o (Hurst et al., 2024), Claude-Sonnet-4.5-20250929 (Anthropic, 2025b), Gemini-3-Flash (Google DeepMind, 2025), Qwen2.5-VL-72B (Bai et al., 2025b) and Qwen3-VL-235B (Bai et al., 2025a).

For baselines that naturally return a ranked list of individual images (*e.g.*, multimodal/text embeddings and metadata augmentation), we follow previous works (Xu et al., 2026; Deng et al., 2026) and adopt an *oracle-size evaluation*: for each query, we truncate the retrieved list to the top-$|B^*|$ images, where $|B^*|$ denotes the cardinality of the GT bundle. Consequently, since the predicted and GT sets have identical sizes, their Precision, Recall, and F1 scores become mathematically equivalent.

Implementation Details. Due to the page limitation, we move the implementation details to Appendix D.

5.2 Main Results

Table 1 summarizes the performance of BundleWeaver and baselines. We make the

Method	Precision	Recall	F1	EM
GPT-4o (baseline)	18.50	17.34	17.84	0.60
BundleWeaver (Ours)	**30.95**	**30.46**	**30.28**	**7.20**
w/o Diverse Seed	27.52	27.66	27.21	6.45
w/o Candidate Pruning	24.57	24.74	24.69	5.85
w/o Beam Search	25.63	25.93	25.35	6.30
w/o VLM Rerank	26.33	26.70	26.13	6.15

Table 2: Ablation study of different components of our proposed BundleWeaver.

following observations:

- **Traditional point-wise matching is fundamentally inadequate for IBC.** Even with model scaling, their performance remains low. This indicates that IBC cannot be solved by simply measuring independent text-image alignment, and cross-image relations are completely lost in atomic point-wise matching paradigms.

- **Naive metadata augmentation is insufficient.** Although spatiotemporal locality naturally correlates with real-world events, explicitly applying it as a rigid heuristic filter into baseline models actually degrades their F1. This degradation highlights that spatiotemporal proximity alone cannot fulfill the specific relational roles demanded by the query, reinforcing the indispensable role of active inline reasoning in IBC.

- **Static decomposition lacks relational constraints.** The decompose-and-rerank paradigm yields a boost in Recall and Precision, as decomposing the query ensures a broader coverage of the visual narrative. However, its EM rate remains devastatingly low. Since sub-queries are retrieved *independently*, the system cannot enforce vital inter-image constraints, highlighting the cross-image composition nature of IBC.

- **BundleWeaver excels via incremental composition.** BundleWeaver achieves state-of-the-art performance across all metrics. This comprehensive improvement is attributed to its dynamic hyperedge discovery: by adaptively expanding hyperedges via parallel beam search, BundleWeaver balances step-wise precision with whole-bundle coherence, effectively enforces the cross-image relational constraints that other paradigms ignore.

5.3 Component Ablation

In Table 2, we ablate the three core stages of BundleWeaver to isolate their contributions. We can observe that:

- Relying on a single starting point artificially restricts the search scope. The performance degradation without diverse seeding can likely be attributed to the search getting trapped in local optima, whereas seeding the search from visually and semantically diverse anchors helps ensure broader coverage across the global image pool.

- Omitting Contextual Candidate Pruning leads to a clear performance drop. Without this constraint, each expansion step must search over a much larger and noisier candidate space, making the beam search more likely to select visually plausible but relationally incompatible distractors. Importantly, the resulting model still outperforms the vanilla GPT-4o baseline by a large margin, suggesting that the performance of BundleWeaver cannot be attributed to metadata pruning alone. Instead, contextual pruning provides a tractable candidate neighborhood, while adaptive expansion and whole-bundle verification remain essential for composing relationally coherent bundles.

- Downgrading the parallel beam search to a greedy expansion strategy drastically hurts performance. Because greedy choices in a massive combinatorial space are highly susceptible to early search errors, maintaining a diverse beam of hypothesis paths appears essential for uncovering the promising visual story.

- Removing the VLM pointwise reranking and relying purely on embedding-based path scores causes performance drop. This suggests that while dense embeddings are efficient for guiding the search direction, they cannot reliably verify complex structural logic or instance-level identity consistency across the completed bundle.

5.4 Breakdown Analysis

Figure 3 further breaks down performance by bundle size and location type. Across all bundle sizes, BundleWeaver consistently outperforms the baselines. As the bundle size increases, all methods become worse, indicating that IBC becomes harder when more images must be jointly recovered. However, BundleWeaver shows a smaller performance drop, suggesting that incremental bundle expansion is more robust than independently retrieving images for decomposed sub-queries.

RzenEmbed performs notably worse on cross-location bundles, since visually diverse images are

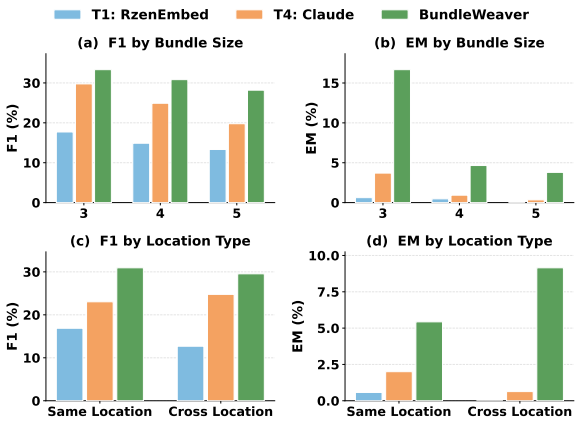


Figure 3: Breakdown analysis on bundle size and location type.

Backbone	Method	Precision	Recall	F1	EM
Qwen3-VL-235B	Dec. & Re.	15.78	14.35	14.95	0.90
	BundleWeaver	25.99	26.23	25.73	5.25
Gemini-3-Flash	Dec. & Re.	20.89	15.53	17.06	0.60
	BundleWeaver	25.52	26.52	25.55	4.35
Claude Sonnet 4.5	Dec. & Re.	24.79	23.01	23.74	1.50
	BundleWeaver	**31.12**	30.39	**30.32**	6.60
GPT-4o	Dec. & Re.	18.50	17.34	17.84	0.60
	BundleWeaver	30.95	**30.46**	30.28	**7.20**

Table 3: Performance w.r.t. different VLM backbones. *Dec. & Re.* represents the baseline *VLM Two-Stage Decompose & Rerank*.

difficult to connect through point-wise similarity alone. The VLM decompose-and-rerank baseline improves F1, but its EM remains low, showing that static decomposition can retrieve partially relevant images but often fails to assemble the exact coherent bundle. In contrast, BundleWeaver achieves the best metrics for both location types. These results indicate that explicitly modeling missing relational roles and verifying the whole bundle are important for robust bundle retrieval.

5.5 Backbone Generalizability

To verify the generalizability of our framework, we evaluate BundleWeaver across various frontier VLM backbones. As shown in Table 3, BundleWeaver consistently outperforms the agentic decompose-and-rerank baseline across diverse architectures, from open-weight models to close-source giants. Notably, the EM rate experiences a massive surge regardless of the chosen backbone. This universal leap confirms that the relational blindness inherent in independent decomposition cannot be rescued merely by deploying a smarter VLM. Instead, it is the structural design of BundleWeaver, *i.e.*, incremental hyperedge discovery coupled with whole-bundle verification, that effectively enforces bundle-level constraints.

5.6 More Experiments

We provide more experiments including reranking strategy comparison, hyperparameter study and case study in Appendix E. We further address potential questions about our paper in Appendix F.

6 Related Works

Traditional T2I retrieval formulates cross-modal matching as an independent, point-wise scoring task, established by benchmarks like MSCOCO and Flickr30K (Lin et al., 2014; Plummer et al., 2015). Subsequent benchmarks have introduced richer contexts, including composed image retrieval (Wu et al., 2021; Baldrati et al., 2023), lifelog retrieval (Gurrin et al., 2023), and multi-hop visual-history search (Xu et al., 2026; Deng et al., 2026). Alongside benchmark evolution, retrieval architectures have advanced from foundational dual-encoders (*e.g.*, CLIP, SigLIP) (Radford et al., 2021; Tschannen et al., 2025) to powerful MLLM-based embeddings (*e.g.*, Qwen3-VL, RzenEmbed) (Li et al., 2026; Jian et al., 2025). Recently, agentic retrieval methods have further replaced one-shot embedding search with reasoning-driven iterative query decomposition (Jiang et al., 2024a; Wang et al., 2026; Liang et al., 2026). Despite these advancements, existing retrieval paradigm remains atomic. IBC redefines the retrieval objective, shifting from independently scoring images to composing cohesive image bundles.

7 Conclusion

In this paper, we introduce **Image Bundle Composition (IBC)**, which shifts the retrieval objective from ranking isolated snapshots to dynamically composing cohesive bundles bound by explicit cross-image relations. To establish this new paradigm and mitigate the computational bottlenecks of combinatorial annotation, we build IBCBench, the first meticulously verified benchmark dataset for this task. Furthermore, we propose BundleWeaver, an agentic framework that reformulates IBC as query-conditioned incremental hyperedge discovery. By leveraging adaptive sub-query generation, parallel beam search, and whole-bundle VLM reranking, BundleWeaver effectively enforces cross-image structural constraints and significantly outperforms existing baseline paradigms.

Limitations

As a pioneering effort in formulating the Image Bundle Composition (IBC) paradigm, this work naturally possesses certain limitations that offer exciting avenues for future research. First, our current investigation is primarily constrained to static images within personal photo collections (*i.e.*, YFCC). We have not yet extended this paradigm to more specialized domains (*e.g.*, medical imaging sequences or legal evidentiary archives) or dynamic modalities such as Video Bundle Retrieval, where temporal dynamics are continuous rather than discrete. Second, BundleWeaver currently operates in a training-free, zero-shot manner by leveraging off-the-shelf foundation models. While this elegantly demonstrates the generalized relational reasoning power of our incremental search mechanism, we have not yet explored end-to-end fine-tuning strategies on specific datasets. Developing parameter-efficient tuning methods tailored specifically for IBC remains a highly promising direction for future work.

References

Pierre Andrews, Fausto Giunchiglia, Javier Paniagua, and 1 others. 2011. Clues of personal events in online photo sharing.

Anthropic. 2025a. Claude opus 4.5 system card. Technical report, Anthropic. System card.

Anthropic. 2025b. System card: Claude sonnet 4.5. Technical report, Anthropic. System card.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025b. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. 2023. Zero-shot composed image retrieval with textual inversion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 15338–15347.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 4(5).

Chenlong Deng, Mengjie Deng, Junjie Wu, Dun Zeng, Teng Wang, Qingsong Xie, Jiadeng Huang, Shengjie Ma, Changwang Zhang, Zhaoxiang Wang, and 1 others. 2026. Deepimagesearch: Benchmarking multimodal agents for context-aware image retrieval in visual histories. *arXiv preprint arXiv:2602.10809*.

Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2025. The faiss library. *IEEE Transactions on Big Data*.

Google DeepMind. 2025. Gemini 3 flash model card. Technical report, Google DeepMind. Model card.

Cathal Gurrin, Björn Þór Jónsson, Duc Tien Dang Nguyen, Graham Healy, Jakub Lokoc, Liting Zhou, Luca Rossetto, Minh-Triet Tran, Wolfgang Hürst, Werner Bailer, and 1 others. 2023. Introduction to the sixth annual lifelog search challenge, lsc'23. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, pages 678–679.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Weijian Jian, Yajun Zhang, Dawei Liang, Chunyu Xie, Yixiao He, Dawei Leng, and Yuhui Yin. 2025. Rzenembed: Towards comprehensive multimodal retrieval. *arXiv preprint arXiv:2510.27350*.

Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanmin Wu, Jiayi Lei, Pengshuo Qiu, Pan Lu, Zehui Chen, Chaoyou Fu, Guanglu Song, and 1 others. 2024a. Mmsearch: Benchmarking the potential of large models as multi-modal search engines. *arXiv preprint arXiv:2409.12959*.

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. 2024b. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *arXiv preprint arXiv:2410.05160*.

Christopher A Kurby and Jeffrey M Zacks. 2008. Segmentation in the perception and memory of events. *Trends in cognitive sciences*, 12(2):72–79.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, and 1 others. 2026. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*.

Qian Liang, Yujia Wu, Kuncheng Li, Jiwei Wei, Shiyuan He, Jinyu Guo, and Ning Xie. 2026. Mm-r1: Unleashing the power of unified multimodal large language models for personalized image generation. In

Proceedings of the AAAI Conference on Artificial Intelligence, volume 40, pages 6835–6843.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Wenbo Lyu, Yingjun Du, Jinglin Zhao, Xianton Zhen, and Ling Shao. 2025. Vischainbench: A benchmark for multi-turn, multi-image visual reasoning beyond language priors. *arXiv preprint arXiv:2512.06759*.

Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Tianshuo Yang, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, and 1 others. 2025. Mmiu: Multimodal multi-image understanding for evaluating large vision-language models. In *International Conference on Learning Representations*, volume 2025, pages 38405–38453.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Stephen Robertson. 2025. Bm25 and all that–a look back. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 5–8.

Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, and 1 others. 2025. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*.

Teng Wang, Rong Shan, Jianghao Lin, Junjie Wu, Tianyi Xu, Jianping Zhang, Wenteng Chen, Changwang Zhang, Zhaoxiang Wang, Weinan Zhang, and 1 others. 2026. Oscar: Optimization-steered agentic planning for composed image retrieval. *arXiv preprint arXiv:2602.08603*.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. 2021. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11307–11317.

Tianyi Xu, Rong Shan, Junjie Wu, Jiadeng Huang, Teng Wang, Jiachen Zhu, Wenteng Chen, Minxin Tu, Quantao Dou, Zhaoxiang Wang, and 1 others. 2026. Photobench: Beyond visual matching towards personalized intent-driven photo retrieval. *arXiv preprint arXiv:2603.01493*.

Youze Xue, Dian Li, and Gang Liu. 2025. Improve multi-modal embedding learning via explicit hard negative gradient amplifying. *arXiv preprint arXiv:2506.02020*.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, and 1 others. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

A Ethical Considerations

The foundation of our benchmark, IBCBench, relies on the public YFCC-100M corpus (Thomee et al., 2016). To ensure strict adherence to data privacy and copyright standards, we exclusively utilize images distributed under Creative Commons licenses that explicitly authorize research applications. We have also audited the associated spatiotemporal metadata to ensure compliance. The release of our dataset will rigorously follow all original redistribution guidelines to respect and protect content creators' rights.

As Image Bundle Composition (IBC) introduces the capability to model complex spatiotemporal trajectories and bundle-level relational constraints, it is crucial to address potential privacy implications. Within the scope of this dataset, privacy risks are minimized because the visual assets are already publicly available under Creative Commons licenses, and our cross-image modeling focuses purely on advancing multimodal relational reasoning rather than profiling real-world identities. More importantly, the core motivation of the IBC paradigm is user-centric. The technology is envisioned as an intelligent agent for personal photo management, empowering individuals to dynamically compose visual stories within their *own* private galleries or encrypted local storage. It is neither intended for, nor optimized for, the unauthorized surveillance or analysis of external, third-party subjects.

B Impact Discussion

While BundleWeaver achieves remarkable relative improvement in Exact Match (EM) over the strongest baseline, we acknowledge that the absolute performance remains modest. This substantial headroom underscores the intrinsic difficulty and profound significance of the IBC task. Perfect recovery of a non-decomposable visual narrative from a massive $\mathcal{O}(N^K)$ combinatorial space poses a formidable challenge for current multimodal agents, exposing a critical blind spot in contemporary retrieval methodologies. Consequently, BundleWeaver is positioned not as a definitive endpoint, but as a foundational baseline and a pioneering starting point. We hope our framework and the benchmark will attract broader attention from the research community to tackle this challenging yet highly meaningful problem. Ultimately, IBC highlights a profound paradigm shift for the community, demonstrating that future retrieval systems must evolve from passive, point-wise rankers into active, reasoning-driven constructors capable of inline relational composition.

C More Details of Dataset Construction

C.1 Property Requirements of a Valid Image Bundle

Before mining candidates, we formally define the criteria for a high-quality bundle. A candidate subset $B \subseteq \mathcal{I}$ must satisfy four strict criteria to be accepted:

- **(C1) Joint Completeness:** Every image in B must play an indispensable role in fulfilling the query. Removing any single image would break the narrative or structural integrity of the answer.
- **(C2) Cross-image Binding:** The query q must enforce a higher-order relation across the images (*e.g.*, temporal progression, recurring rituals, or before-and-after contrast), rather than merely listing independent visual targets.
- **(C3) Uniqueness within Pool:** To ensure robust evaluation, B must be unambiguous. There must be minimized subset in the entire image pool $\mathcal{I}$ that satisfies q better than or equally well as B.
- **(C4) Bounded Redundancy:** The images within B must be visually distinct and not near-duplicates (*e.g.*, burst shots).

C.2 Source Data and Visual Profiles

We source our raw image pool from the YFCC-100M dataset (Thomee et al., 2016), selecting 57 users with rich metadata. The global pool comprises 109,467 photos, all possessing valid EXIF timestamps and over 98.3% having GPS coordinates. Crucially, we discard all user-defined album ids from YFCC, ensuring that our bundles emerge from visual and spatiotemporal relations rather than pre-existing manual groupings.

To enable semantic filtering, we cache rich visual signals for all 109,467 images. We utilize GPT-4o (Hurst et al., 2024) to extract dense text captions, word-level tags, and event-level signals (*e.g.*, *wedding, summit*), whose prompt is shown in Appendix C.7.1. Additionally, we compute multimodal embeddings using RzenEmbed, which serves to penalize redundancy during the candidate mining phase.

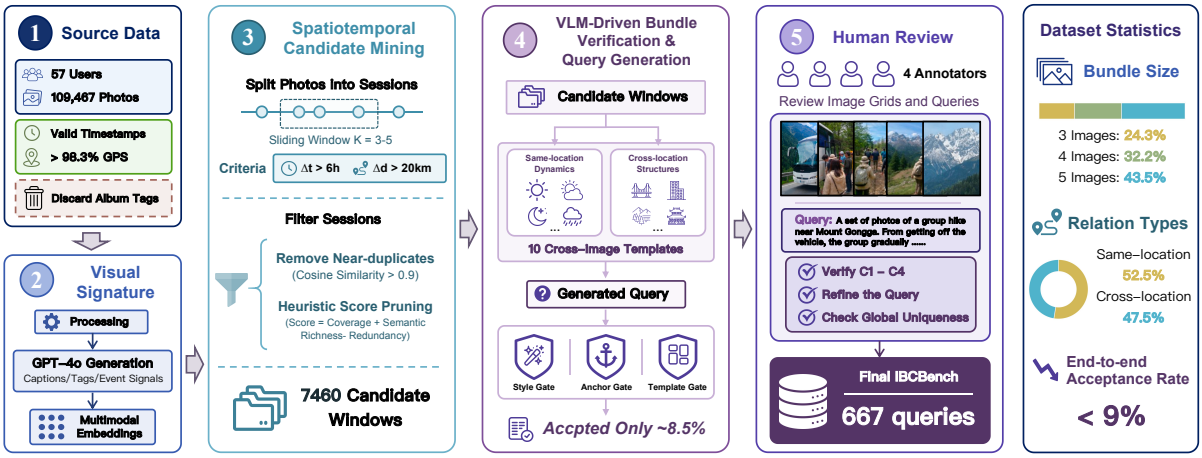


Figure 4: The demonstration of dataset construction pipeline for IBCBench.

C.3 Spatiotemporal Candidate Mining

To bypass the $\mathcal{O}(N^K)$ combinatorial explosion, we anchor our search space using metadata constraints. We first group each user's photos into spatiotemporal sessions, breaking sequences wherever there is a time gap $\Delta t > 6$ hours or a geographic jump $\Delta d > 20$ km.

Within each session, we enumerate sliding windows of size $K \in [3, 5]$. To prioritize structurally promising candidates, we evaluate each window using a composite heuristic score:

$$\text{Score} = S_{\text{st}} + S_{\text{sem}} - P_{\text{red}}, \tag{6}$$

where S_{st}, S_{sem}, and P_{red} denote spatiotemporal coverage, semantic richness, and redundancy penalty, respectively. The three components are detailed below:

- **Spatiotemporal Coverage (S_{st}).** A valid visual story typically requires a meaningful progression in time or space, rather than a static burst of photos taken simultaneously. The spatiotemporal coverage component rewards candidate windows that span a reasonable duration and physical distance:

$$S_{\text{st}}(W) = \min\left(2, \frac{T}{2}\right) + \min\left(2, \frac{D}{3}\right) + \mathbf{1}_{\text{anchor}}(W). \tag{7}$$

  where T denotes the total time span of the window in hours, and D represents the cumulative geodesic distance in kilometers between consecutive images. The indicator function $\mathbf{1}_{\text{anchor}}(W)$ returns 1 if the photos share at least one identical reverse-geocoded address prefix, which rewards spatial anchoring. The $\min(\cdot)$ operations act as clipping functions to prevent extreme outliers (*e.g.*, intercontinental flights) from disproportionately dominating the score.

- **Semantic Richness (S_{sem}).** Beyond physical progression, a high-quality bundle must encapsulate a rich set of visual and narrative elements. We leverage the pre-cached tags and event signals generated by GPT-4o to measure semantic diversity:

$$S_{\text{sem}}(W) = \min\left(2, \frac{|\mathcal{T}_W|}{6}\right) + \min\left(1.5, \frac{|\mathcal{E}_W|}{4}\right) \tag{8}$$

  where $|\mathcal{T}_W|$ and $|\mathcal{E}_W|$ denote the cardinality of the union of all word-level tags and event-level signals within the window W, respectively. This component explicitly favors bundles that exhibit diverse semantic concepts, ensuring that the retrieved images are information-dense.

Candidate windows with a total score below a strict threshold (default to 2.5 in our implementation) are discarded. This heuristic pruning aggressively reduces the search space, yielding a manageable pool of 7,460 highly diverse candidate windows.

C.4 VLM-Driven Bundle Verification and Query Generation

The heuristically mined windows satisfy spatiotemporal constraints, but they do not necessarily exhibit narrative coherence (C1-C2). We employ Claude-Opus-4.5 (Anthropic, 2025a) as a rigorous semantic verifier. The VLM is instructed to evaluate each candidate against several predefined cross-image templates divided into two families:

- **Same-location Dynamics:** *e.g.*, visible state changes, event phase progressions (preparation

→ climax → end), or physical transitions (day-to-night).

- **Cross-location Structures:** *e.g.*, repeating travel rituals, route-level continuity, or shared thematic activities across different venues.

For a candidate to be accepted, the VLM must successfully map it to one of these templates, generate a natural language query, and articulate a specific, instance-level *shared anchor* (required by C2, *i.e.*, *the same kayaker*). To prevent the VLM from generating lazy enumerations, we implement a three-layer programmatic gate (*i.e.*, Style Gate, Anchor Gate, Template Gate, see Appendix C.7.3 for the prompt) that automatically rejects outputs containing forbidden grammatical structures (*e.g.*, strings of isolated gerunds) or vague anchors. This strict verification accepts only about 8.5% of the candidates.

C.5 Human Review

In the final stage, all VLM-approved candidates undergo rigorous human review. Four expert annotators examine the image grid, the generated query, and the global pool context. Annotators are tasked with verifying constraints C1-C4, refining the query text for naturalness, and explicitly rejecting any queries that lack pool-wide uniqueness (C3).

Specifically, for each candidate, the interface (shown in Figure 5) displays: (i) the generated query and VLM rationale (including `template_id`, `shared_anchor`, and all audit fields); (ii) the bundle images rendered as a grid with per-image metadata (capture time, GPS-derived address, and device); and (iii) a *uniqueness panel* that supports the annotator in verifying constraint (C3).

Embedding-based uniqueness assistance. A key challenge for human reviewers is judging whether a given bundle is the *unique best answer* within a pool of 109K images, since visually similar alternatives may exist. To assist this decision, we precompute pairwise cosine similarities using the Rzenembed embedding model and, for each bundle image, retrieve its top-5 most similar photographs from the *same user* within a ±3-day time window. These nearest neighbors are displayed alongside the bundle image with their similarity scores and metadata, enabling reviewers to quickly assess whether a near-duplicate or semantically

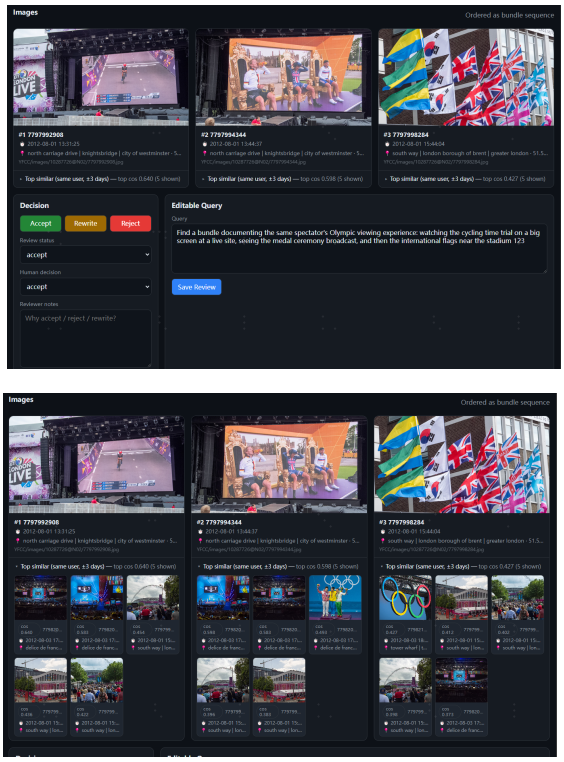


Figure 5: The web demo for human review and deduplication.

interchangeable image exists that would undermine the bundle's uniqueness. A global *uniqueness risk* indicator is also computed as the maximum neighbor similarity across all bundle images, color-coded as HIGH ($\cos \geq 0.85$), MEDIUM ($0.75 \leq \cos < 0.85$), or LOW ($\cos < 0.75$).

Annotator actions. For each candidate, the reviewer selects one of three decisions: ACCEPT (the bundle satisfies all four constraints and the query is accurate), REJECT (any constraint is violated), or REWRITE (the bundle is valid but the query needs editing). When accepting or rewriting, the annotator may edit the query text to improve naturalness or specificity. Free-text notes are also supported for flagging borderline cases.

Agreement and yield. A total of 1,674 candidates were reviewed, of which 667 (39.8%) were accepted. Per-annotator acceptance rates range from 30.2% to 49.9%, reflecting differences in individual strictness during the initial review. To reduce annotator-specific bias, accepted candidates were further cross-validated by other annotators, and borderline or disputed cases were discussed collectively before inclusion in the final dataset. Moreover, we leverage the upstream programmatic

gates and VLM audit fields to enforce a consistent quality floor throughout the review process.

Furthermore, we manually inspect the predicted bundles from BundleWeaver and the baseline that were automatically evaluated as incorrect. As detailed in Section E.3, the rate of genuine false negatives (*i.e.*, unannotated but reasonable alternative bundles) is negligible. This further suggests the uniqueness of the ground-truth bundles established during the dataset construction process.

C.6 Final Dataset Statistics

Following human curation, the final IBCBench dataset contains **667 high-quality queries** evaluated against the massive **109,467-image** pool. The bundle sizes are distributed across 3 images (24.3%), 4 images (32.2%), and 5 images (43.5%). The semantic relations are well-balanced, with roughly 52.5% focusing on same-location dynamics and 47.5% capturing cross-location structural ties. Through this exhaustive pipeline with an end-to-end acceptance rate under 9%, we aim to ensure that every ground-truth bundle in the dataset is unique and cohesive that challenges the limits of modern retrieval systems.

C.7 Prompt Demonstration

C.7.1 Prompt for Building Visual Profiles

Image Profiling

System: You are helping construct Image Bundle Composition data. Return strict JSON with keys: caption, tags, event_signals, bundle_notes. Keep everything in English.
User: Describe this image for downstream bundle mining. Match this JSON schema exactly:
{ "caption": "Detailed but concise English description of the image.", "tags": ["short tags capturing objects, scene and activity"], "event_signals": ["possible event or state-change clues"], "bundle_notes": "Short note about whether the image may help a time/state/route/event progression bundle."}

C.7.2 Prompt for Bundle Verification and Query Generation

Bundle Verification and Query Generation

System: You are a strict IBC dataset judge. Follow every rule literally.
User: You are a STRICT validator for a Image Bundle Composition (IBC) dataset. The bundle is ACCEPTED only if ALL of the following hold:

1. The images fit exactly one template in the allowed list below.
2. There is a CONCRETE instance-level shared_anchor (specific person / object / ritual / named route / repeated action visible across images). Generic phrases like 'temporal progression', 'same day', 'the trip', 'the route', 'cross-image reasoning' are NOT acceptable anchors.
3. Removing any single image would clearly break the bundle's meaning.
4. The query describes one UNIQUE bundle within the available pool. It is not an album dump.
5. The query does NOT read like a list of independent single-image targets.

FORBIDDEN QUERY PATTERNS (auto-reject):

- 'showing A, B, C, and D' style enumeration where each item is a separate scene
- 'from A, through B, to C, and finally D' when A/B/C/D are heterogeneous scenes without a shared anchor
- 4+ gerund verbs in a comma/and list ('doing X, doing Y, doing Z, doing W')

SAME-LOCATION FAMILY RULES:

- The defining logic must come from what CHANGES, PROGRESSES, or becomes STRUCTURALLY DIFFERENTIATED within one location, not from the fact images share a place.
- Reject queries that are a list of different objects/views/scenes at the same place.
- Reject if the query can be paraphrased as 'one photo of A, one of B, one of C at the same place'.
- Reject 'different aspects / different views / different angles / different corners' phrasing.
- A good query answers 'what changes or progresses within this place across the images?'. It must not merely answer 'what different things can be seen at this place?'.

CROSS-LOCATION FAMILY RULES:

- The defining logic must come from what STAYS STRUCTURALLY THE SAME across locations (repeated ritual / same semantic role / recurring activity / route-level progression).
- Reject checklists of destinations, attractions, landmarks, or 'different aspects of a trip'.
- Reject if the query can be paraphrased as 'one photo from place A, one from place B, one from place C'.
- cross.5 (generic shared pattern) is a last-resort template. Prefer cross.1/2/3/4 whenever a concrete ritual / role / activity / route applies. Do NOT use cross.5 for 'different nice streets/beaches/landmarks'.
- A good query answers 'what is being repeated or structurally preserved across places?'. It must not merely answer 'which places should appear?'.

GENERIC FAMILY RULES: Only accept when the bundle is held together by a true cross-image relation. ALLOWED TEMPLATES for this candidate (choose exactly one template_id):

- same.1: Same location + visible state or quantity change. The bundle must show a gradual, observable change within a fixed setting (street sparse->crowded, table empty->full->finished, campsite quiet->active).
- same.2: Same event in one place + stage progression. The bundle must move through distinct phases of one episode (setup -> peak -> aftermath; decorations -> celebration -> cleanup).
- same.3: Same place + before/after transition. Location identifiable, visible state change between earlier and later condition (before guests vs. during vs. after; before decorations lit vs. after).
- same.4: Same place + day-to-night / lighting progression. Location preserved, lighting/atmosphere visibly shifts over the day.
- same.5: Same place + complementary event summary. Tightly linked slices of ONE coherent episode in one place; each image plays a distinct but jointly-necessary role. Use this ONLY when progression is not monotonic but the slices are clearly one episode.
- cross.1: Repeated travel ritual across locations. Same personal ritual repeated in several different places (first meal after arrival, first hotel-window view, entrance-sign arrival photo).
- cross.2: Same moment right after arrival across cities. A fixed temporal role in a repeated travel routine (first street scene after arrival, first check-in view, first stop after getting there).
- cross.3: Route-level progression across locations. Ordered movement through space during ONE continuous trip (hotel -> old town -> viewpoint; trailhead -> mid-route -> summit; city center -> waterfront -> harbor at sunset).
- cross.4: Repeated concrete activity across locations. Same specific visually-grounded act documented the same way in different places (breakfast-before-eating shots across cities, market-visit shots across trips, campsite-setup across camps).
- cross.5: Shared pattern stronger than location identity (generic fallback). Higher-level visual/semantic pattern stable across different places. Use ONLY when the shared pattern is specific, concrete, and strong enough to define a unique bundle. Do NOT use for 'different nice streets / beaches / landmarks'.

CALIBRATION EXAMPLES THAT MUST BE REJECTED:

- REJECT: "Find photos documenting a day trip in Mendoza, Argentina that includes relaxing at a hotel pool, viewing the hotel's glass architecture, walking to a stone-facade restaurant, dining with a group, and the hotel's Christmas-decorated entrance at night." -> decomposable checklist of 5 independent single-image targets; uses 'including ...'
- REJECT: "Find images showing different aspects of a day trip to Santorini, Greece: the black sand beach, small boats on the Aegean Sea, traditional Cycladic architecture, and the cliffside town of Fira at dusk." -> tourism checklist; uses 'different aspects of'; each item is a separate single-image task
- REJECT: "Find all photos from a single-day hiking expedition in the Andes mountains near Aconcagua, showing hikers on trail, mountain valley landscape, pack mules carrying supplies, and local wildlife." -> uses 'Find all photos'; the content is a 'showing A, B, C, and D' enumeration with no shared anchor
- REJECT: "Find all photos from the same day showing this group's holiday gatherings at both the restaurant with ornate lantern lighting and the home with red walls and Christmas decorations in Mexico City." -> uses 'Find all photos'; blurs bundle boundary; two scene slices not tied by explicit progression

Return STRICT JSON matching this schema (no markdown, no commentary):

```
  {
"accept": "true or false",
"query": "Natural English retrieval
↪  request, or null when rejected.",
"image_ids": [photo_id_1, photo_id_2, ...],
"rationale": {
  "template_id": "one of same.1..same.5 or
  ↪  cross.1..cross.5",
  "shared_anchor": "REQUIRED. A concrete,
  ↪  instance-level cross-image anchor
  ↪  that holds the bundle together. Must
  ↪  name a specific person / piece of
  ↪  equipment / landmark / ritual /
  ↪  repeated action / named route. MUST
  ↪  NOT be a generic phrase such as
  ↪  'temporal progression', 'same day',
  ↪  'single trip', 'the event', 'the
  ↪  route', 'cross-image reasoning'.
  ↪  Example: 'the same paddler wearing a
  ↪  blue hat, together with the same
  ↪  green canoe, appears in every
  ↪  image'.",
  "why_complete": "Why the set is only
  ↪  complete jointly. One sentence.",
  "why_unique": "Why this bundle is the
  ↪  most specific answer within the
  ↪  available pool.",
  "why_linked": "What cross-image
  ↪  relationship binds the images
  ↪  together. One sentence.",
  "why_not_decomposable": "Why this is not
  ↪  just multiple single-image tasks.
  ↪  One sentence.",
  "skeleton": "A/B/C"
},
"audit": {
  "single_image_insufficiency": "pass or
  ↪  fail with short justification",
  "non_substitutability": "pass or fail
  ↪  with short justification",
  "low_redundancy": "pass or fail with
  ↪  short justification",
  "natural_language_query": "pass or fail
  ↪  with short justification",
  "template_fit": "pass or fail, explain in
  ↪  <= 20 words why template_id applies",
  "review_risk": "low / medium / high"
},
"reject_reason": null
}
```

If rejecting, set accept=false, query=null, still fill rationale.template_id with the closest template you consid-

ered, fill shared_anchor with whatever concrete anchor (or 'none') you could identify, and put the specific failure in reject_reason.

C.7.3 Details of Programmatic Gates

The judging prompt in Appendix C.7.2 explicitly states the acceptance criteria (forbidden query patterns, concrete anchor requirements, and template-family constraints. Since the VLM does not always comply with these instructions, we deterministically verify each accepted output against the same rule set using regular expressions and keyword matching. Specifically, we re-check: (1) the generated query text for forbidden surface patterns, (2) the `shared_anchor` field for sufficient specificity (rejecting anchors shorter than 4 tokens or dominated by generic phrases from a 28-term blocklist), and (3) the `template_id` for membership in the candidate's allowed template family. Any violation overrides the model's acceptance, converting it to a programmatic rejection.

D Implementation Details

All methods, including baselines and BundleWeaver, are evaluated on the full IBCBench-dataset (*i.e.*, 667 queries and bundles). For our proposed BundleWeaver , we use RzenEmbed with a FAISS IndexFlatIP backend (Douze et al., 2025) for all base retrieval operations. We set the number of seed $N_s = 5$, beam width $B_u = 4$, and the candidates retrieved per step $C = 5$. For candidate pruning, we set the max bound to images captured within $\tau_t = 24$ hours and $\tau_g = 50$ km of the seed. The balancing coefficient λ in Equation 5 is set to 0.3. Unless otherwise specified, GPT-4o is used for both adaptive sub-query generation and whole-bundle pointwise reranking. All experiments are conducted on two A100 GPUs.

D.1 Prompt Demonstration

D.1.1 Prompt for Adaptive Sub-Query Generation

Pointwise Reranking

System: You are assisting in constructing a photo bundle from a personal photo collection.
User: Original query: {query}
Previous search directions:
1. {sub_query_1}
2. {sub_query_2}
...
Images found so far:
Image 1: {caption_1}
Image 2: {caption_2}
...
Based on the original query and images already found, determine what is STILL MISSING from this photo bundle.

- If something is missing, output exactly one line: SEARCH: <concise description of the photo to find next>
- If the bundle appears complete (all aspects of the query are covered), output exactly: DONE

Output only SEARCH: ... or DONE, nothing else.

D.1.2 Prompts for Whole-Bundle VLM Reranking

Pointwise Reranking

System: You are evaluating whether a set of photos forms a coherent bundle matching a user query.
User: Query: {query}
This bundle has {k} photos:
{Photo 1}
{Photo 2}
...
Rate how well this photo bundle satisfies the query on a scale of 1-10:

- 10: Perfect match — every photo is relevant, together they fully satisfy the query.
- 7-9: Strong match — most photos are relevant, the bundle largely satisfies the query.
- 4-6: Partial match — some photos fit but the bundle is incomplete or has irrelevant photos.
- 1-3: Poor match — photos mostly don't relate to the query.

Output EXACTLY: SCORE: <number>
Then one sentence explaining why.

Listwise Reranking

System: You are a photo-album curator. Given a user query and {n} candidate photo bundles, pick the ONE bundle that best satisfies the query as a coherent set.
User: Query: {query}
{candidates}
For each bundle, ask:

1. Do ALL photos belong to the SAME event/scene/trip described in the query?
2. Does the set cover the key moments the query asks for?
3. Are there irrelevant or duplicate photos?

Reply with EXACTLY one line:
BEST: <number>
Then one sentence explaining why.

D.2 Baseline Implementation

Multimodal Embedding. We encode the query text and all pool images independently using four vision-language embedding models: CLIP-ViT-B/32, SigLIP2-giant, Qwen3-VL-Embedding-8B, and RzenEmbed-7B. All image embeddings are L2-normalized and indexed with FAISS `IndexFlatIP` for maximum inner product search, which is equivalent to cosine similarity for normalized vectors. At query time, the query is encoded by the same multimodal embedding model. This type of methods tests whether a single shared embedding space can capture the *joint* semantics of a multi-image bundle from a text query alone.

Caption + Text Embedding. We first generate a detailed English caption for every image in the pool using GPT-4o, then perform text-to-text retrieval between the query and image captions. Three retrieval models are evaluated:

- **BM25**: a sparse lexical model using whitespace-tokenized captions.
- **BGE-M3**: using only the dense retrieval branch with FAISS indexing.
- **Qwen3-Embedding-8B**: a recent instruction-aware text embedding model where queries and passages are encoded with separate prompts.

Both dense models use L2-normalized embeddings with inner product search. This type of methods isolates whether richer textual descriptions of individual images can compensate for the lack of visual features in bridging the query–bundle gap.

Heuristic Metadata Augmentation. Building on the best multimodal model (RzenEmbed), we provide another line of baselines incorporating spatiotemporal metadata available in the photo manifest. Three strategies are tested:

- **Session Clustering:** The retrieved images are sorted chronologically and partitioned into discrete sessions. A new cluster boundary is formed whenever consecutive images exhibit a time gap > 6 hours or a geographic jump > 20 km. The cluster with the highest aggregate embedding score is then selected, yielding its top-$|B^*|$ images.
- **User + Session Clusteing:** The retrieved images are first grouped by photographer. The dominant user is identified as the one whose top-$3|B^*|$ images have the highest aggregate embedding score, where $|B^*|$ denotes the size of the ground-truth bundle. Session clustering is then applied within that user's images only. This leverages the observation that ground-truth bundles always belong to a single user.
- **Time proximity boosting:** For each retrieved image i, a bonus is added based on the embedding scores of temporally nearby images: $\text{boost}_i = \sum_{j:|\Delta t_{ij}|\leq 1\text{h}} 0.1 \cdot s_j \cdot (1 - |\Delta t_{ij}|)$, where s_j is the original embedding score of image j and $|\Delta t_{ij}|$ is the time difference in hours. This linearly-decaying bonus encourages selecting images from the same short-duration event.
- **User + Spatiotemporal:** The retrieved candidates are first grouped by user. For each user, the image with the highest embedding score is designated as the anchor. The user's candidate pool is then explicitly filtered to retain only images falling within a ± 24-hour window and a $50\,\text{km}$ geographic radius relative to the anchor. The dominant user is identified by the highest aggregate score of their top-$3|B^*|$ filtered candidates, and their top-$|B^*|$ images are returned as the final bundle.

VLM Decompose & Rerank. This is the strongest baseline, employing a two-stage agentic pipeline. In the *decomposition* stage, a VLM is prompted to break the bundle query into m sub-queries ($m \leq 5$), each describing one individual image expected in the bundle. In the *retrieval* stage, each sub-query is independently encoded using the RzenEmbed model and the top-5 most similar images are retrieved from the FAISS index, yielding up to $5m$ candidates. In the *reranking* stage, the VLM receives all candidate images alongside the original query and is prompted to select exactly one image per sub-query group. If the VLM fails to parse or produces invalid selections, a fallback selects the top-1 retrieval result from each sub-query group. We evaluate five VLMs: GPT-4o, Claude-Sonnet-4.5, Gemini-3-Flash-Preview, Qwen2.5-VL-72B, and Qwen3-VL-235B. Note that the number of selected images m is determined by the VLM's decomposition and may differ from the ground-truth bundle size $|B^*|$.

Strategy	Precision	Recall	F1	EM
No-rerank	26.33	26.70	26.13	6.15
Listwise	29.41	**31.28**	29.94	6.90
Pointwise	**30.95**	30.46	**30.28**	**7.20**

Table 4: Performance comparison of different reranking strategies.

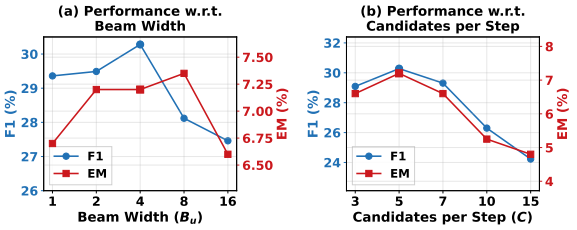


Figure 6: Hyperparameter study of the beam width and candidates per step.

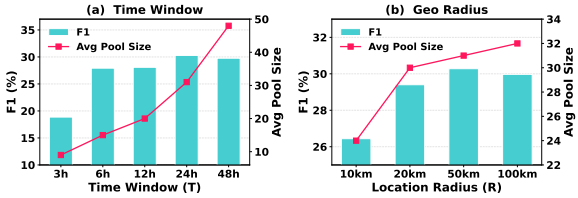


Figure 7: Hyperparameter study of the candidate pruning range.

Method	Precision	Recall	F1	False Negative Rate
Claude-Sonnet-4.5	22.20	21.48	21.66	0%
BundleWeaver	26.27	24.47	25.08	2%

Table 5: Manual audit of strict false negatives on 100 randomly sampled queries, where both BundleWeaver and Claude-Sonnet-4.5 fail under automatic set-level evaluation.

E Additional Experiments

E.1 Reranking Strategy Comparison

For reranking strategy in Section 4.4, we compare *Pointwise* scoring against *Listwise* selection (*e.g.*, asking the VLM to pick the best from all constructed bundles). From Table 4, we can see that rerank generally helps, and pointwise evaluation is more rigorous and scalable. While Listwise evaluation achieves competitive Recall, Pointwise scoring yields the best F1 and EM. We attribute this to attention dilution in Listwise prompts: when faced with dozens of images simultaneously, VLMs struggle to verify intricate constraints, which is consistent with previous works (Meng et al., 2025; Lyu et al., 2025). Pointwise scoring forces the VLM to independently and rigorously audit each bundle, also offering better scalability than Listwise reranking.

E.2 Hyperparameter Study

Parameters of Beam Search. In Figure 6, we investigate the sensitivity of BundleWeaver to its core exploration hyperparameters: the parallel beam width (B_u) and the number of local candidates retrieved per step (C). Initially scaling up these parameters improves performance by providing the reasoning agent with a richer, more diverse set of visual building blocks. However, over-expanding the beam width or retrieving an excessive number of candidates per step actively degrades the final retrieval quality. This occurs because unbounded exploration floods the candidate pool with noisy, locally similar but globally disjointed images, which ultimately derails the LLM's adaptive reasoning trajectory. The existence of a clear optimal (*i.e.*, $B_u = 4, C = 5$) empirically indicates that solving IBC requires a delicate balance: while diverse exploration is necessary to avoid local optima, aggressive and strict pruning is equally vital to maintain a coherent narrative direction within a massive combinatorial space.

Parameters of Candidate Pruning. We study the effect of the temporal window T and geographic radius R used by the agent in contextual candidate pruning. Figure 7 reports F1 together with the average local pool size induced by each constraint. The pool size is computed with the first ground-truth image as the anchor only for diagnostic analysis, and this oracle anchor is not used by the retrieval method.

As shown in the figure, an overly tight temporal window ($T = 3$h) substantially hurts performance, since it excludes valid bundles spanning longer events or day-long trips. Increasing T to 12–24h yields the best performance, while further enlarging it to 48h increases the candidate pool without improving F1. A similar trend can be observed for location: $R = 10$km is too restrictive for cross-location bundles, whereas performance remains stable from 20km to 100km. These results show that our default setting ($T = 24$h, $R = 50$km) maintains performance while keeping the candidate pool compact.

E.3 Human Validation of Model Results

To assess whether strict set matching unfairly penalizes methods due to missing alternative valid bundles, we manually audit 100 randomly sampled queries where both BundleWeaver and Claude-

Sonnet-4.5 are marked incorrect by automatic exact set-level evaluation. As shown in Table 5, unannotated valid alternatives are rare (required by C3 in Section C.1): only 2% of BundleWeaver predictions are judged as reasonable alternative bundles, while none of the Claude-Sonnet-4.5 predictions are human-valid. This suggests that incomplete ground-truth annotation does not substantially distort our strict evaluation protocol. Meanwhile, even when both methods fail to recover the complete target bundle, BundleWeaver achieves higher Precision, Recall, and F1, indicating better partial recovery of the ground-truth bundle images in difficult failure cases.

E.4 Case Study

To intuitively demonstrate the limitations of existing retrieval paradigms and the superiority of BundleWeaver, we provide case studies on two representative queries from IBCBench, as shown in Figure 8.

Case 1: Spatial Route Progression. The query requests a continuous spatial narrative: a single-day walking route through the Dutch countryside, explicitly requiring a progression through distinct visual landmarks (a church, green pastures, a canal path, farmsteads, a wooded estate, and a river weir).

- **Multimodal Embedding (RzenEmbed):** The traditional embedding model completely fails to capture the narrative progression. Instead, it suffers from *visual collapse*, aggressively retrieving images that are visually homogeneous (*e.g.*, predominantly green canal paths). This perfectly illustrates the flaw of atomic point-wise matching: it optimizes for isolated semantic similarity but ignores the sequential completeness of the query.

- **VLM Decompose & Rerank (Claude-Sonnet-4.5):** The agentic decomposition baseline successfully identifies the distinct landmarks by breaking the query into independent sub-queries. However, because these sub-queries are retrieved in isolation, the resulting bundle is a *Frankenstein* composition. The images belong to different trips, different users, or entirely different geographical regions, severely violating the real-world route continuity constraint required by IBC.

- **BundleWeaver (Ours):** By anchoring the search within valid spatiotemporal metadata windows and using an LLM to adaptively search for the *next missing landmark* conditioned on the current path, our method successfully constructs a coherent, authentic walking route that perfectly matches the user's compositional intent.

Case 2: Temporal Event Progression. The query demands a temporal progression of a specific event: a wedding day, spanning from the empty setup and the waiting groom to the ceremony, reception table, and guest book. This query requires *instance-level identity consistency* (*i.e.*, all photos must belong to the exact same wedding).

- **Multimodal Embedding (RzenEmbed):** The embedding model retrieves a redundant set of highly similar ceremony shots. The strong overarching semantic signal of *wedding* overshadows the fine-grained requirements (like the empty setup or guest book), showing that dense embeddings struggle to compose diverse narrative slices.

- **VLM Decompose & Rerank (Claude-Sonnet-4.5):** This baseline successfully retrieves the diverse semantic components (the groom, the ceremony, the table setting, the guest book). However, a closer inspection reveals a fatal flaw: **the images depict completely different couples and different weddings**. This is the ultimate proof of the "Relational Blindness" discussed in Section 5. The static divide-and-conquer strategy cannot enforce the implicit constraint that the core subject must remain identical across the independently retrieved subsets.

- **BundleWeaver (Ours):** Our framework excels here. By incrementally expanding the bundle and utilizing pointwise VLM reranking to verify the whole-bundle logic, BundleWeaver seamlessly tracks the temporal progression while maintaining strict identity consistency. It retrieves the exact timeline of a single, unique wedding day, showcasing its robust capability in modeling complex, non-decomposable joint relevance.

F Discussions on Potential Concerns

In this section, we address potential questions regarding our paper.

F.1 Is contextual candidate pruning a benchmark bias?

A potential concern is that IBCBench is mined from spatiotemporal sessions, while BundleWeaver

Method	Precision	Recall	F1	EM
RzenEmbed	14.88	14.88	14.88	0.30
w/ CCP	8.80	8.80	8.80	0.75
GPT-4o Dec. & Re.	18.50	17.34	17.84	0.60
w/ CCP	16.75	16.74	16.74	3.15
BundleWeaver w/o CCP	24.57	24.74	24.69	5.85
BundleWeaver	**30.95**	**30.46**	**30.28**	**7.20**

Table 6: Controlled experiment on contextual candidate pruning (CCP).

also uses contextual candidate pruning (CCP). This raises the question of whether BundleWeaver benefits from a dataset-specific shortcut rather than solving the intended relational composition problem.

We argue that CCP should be viewed as a *query-faithful physical plausibility constraint* to ensure bundle uniqueness, rather than an artificial benchmark bias. Many IBC queries explicitly or implicitly contain temporal and geographic constraints, such as a single trip, a same-day event progression, a route across landmarks, or a repeated activity at nearby venues. In such cases, using timestamps and GPS metadata is part of faithfully executing the user intent, analogous to filtering by date or location in real-world personal photo search. Ignoring such metadata would create an unrealistic retrieval setting in which the model is forced to search globally even when the query itself specifies a physical context.

Importantly, however, CCP only defines a broad feasible candidate region; it does not determine the answer. Physical proximity is neither sufficient nor monotonically beneficial for IBC. As shown in Table 6, adding the same CCP mechanism to RzenEmbed decreases F1 from 14.88 to 8.80, and adding it to GPT-4o Decompose-and-Rerank decreases F1 from 17.84 to 16.74, although EM improves slightly. This indicates that spatiotemporal locality may help recover some complete local bundles, but it also introduces many physically close yet relationally irrelevant distractors and can hurt partial set recovery.

This observation is also *consistent* with our dataset construction process. Although candidate windows are first mined under spatiotemporal constraints, fewer than 9% survive the subsequent semantic verification and human review. Thus, most spatiotemporally plausible windows are not valid IBC answers. The benchmark therefore does not reward recovering metadata priors alone; it requires identifying which images jointly instantiate the non-decomposable relation specified by the query.

Finally, BundleWeaver remains strong even when CCP is removed. The w/o-CCP variant achieves 24.69 F1, substantially outperforming the strongest GPT-4o static baseline at 17.84 F1. This shows that the core gain of BundleWeaver comes from adaptive missing-role reasoning, beam-based bundle construction, and whole-bundle verification. CCP mainly serves as a practical physical boundary that reduces global combinatorial noise, rather than as a shortcut to the ground-truth bundle.

F.2 Is the "uniqueness" assumption fully validated?

Ground-truth uniqueness leads to robust evaluation. IBC evaluation relies on exact set-matching, which assumes a single ground-truth bundle per query. A critical question is whether bundles annotated in IBCBench have multiple *reasonable alternative* answers, which would cause our evaluation to unfairly penalize valid predictions.

To address this, we highlight the criteria for uniqueness and provide quantified evidence that false negatives (penalizing a valid alternative) are minimized.

Quantified Error Audit. As reported in Table 5, we manually inspected 100 randomly sampled errors where both BundleWeaver and Claude-Sonnet-4.5 failed the automatic evaluation. A predicted bundle is judged as a *reasonable alternative* (*i.e.*, a genuine false negative) **ONLY IF** it fully satisfies constraints C1-C4 described in Section 3 and perfectly aligns with the text query, but consists of different images than the ground truth. Upon rigorous inspection, the false negative rate was merely 2% for BundleWeaver and 0% for Claude. In $> 98\%$ of the failure cases, the predicted bundles were genuinely flawed (*e.g.*, containing irrelevant distractors, breaking the narrative timeline, or violating identity consistency).

How Uniqueness is Enforced During Construction. The low false negative rate is a direct result of our strict human verification design (C3). During annotation, reviewers were equipped with a Nearest-Neighbor visual panel (Figure 5) to explicitly hunt for alternative bundles.

- **Example of a rejected query:** If a VLM generates *"Find 3 photos of the couple kissing at the wedding"*, and the pool contains a burst of 10 kissing photos, any subset of 3 would be a valid

alternative. Because it violates uniqueness, **the annotator rejects this query entirely**.

- **Example of an accepted query:** The annotator refines the query to "Find a progression of the ceremony: from the empty setup, to the groom waiting, and finally the kiss." This structural constraint forces a unique, unambiguous trajectory, eliminating reasonable alternatives.

By explicitly rejecting any queries with interchangeable subset combinations during construction, we try our best to ensure that the exact set-matching evaluation metric remains highly reliable and fair.

G Theoretical Limitations of Atomic Retrieval in IBC

In this section, we mathematically demonstrate why the static Decompose-and-Rerank paradigm inherently suffers from *Relational Blindness* when applied to Image Bundle Composition (IBC). We first prove the non-submodularity of the IBC objective, which subsequently establishes the theoretical bound where local top-k retrieval fails to converge to the global optimum.

G.1 Preliminaries and Problem Definition

As defined in Equation 2, the objective of IBC is to find the optimal image subset B^* that maximizes the joint relevance score:

$$B^* = \arg \max_{B \subset \mathcal{I}, 2 \leq |B| \leq K_{max}} \Phi(B, q)$$

In the Decompose-and-Rerank paradigm, a complex query q is factorized into K independent sub-queries $\mathcal{Q} = \{q_1, q_2, \ldots, q_K\}$. The joint relevance scoring function $\Phi(B, q)$ can be conceptually formulated as the sum of local semantic matching scores and a global cross-image relational constraint:

$$\Phi(B, q) = \sum_{i=1}^{K} f(x_i, q_i) + \gamma \cdot \mathbb{I}_{rel}(B), \quad (9)$$

where:

- $f(x_i, q_i)$ denotes the independent visual-textual alignment score between a single candidate image x_i and the sub-query q_i.
- $\mathbb{I}_{rel}(B) \in \{0, 1\}$ is a non-decomposable Boolean indicator function representing whether the subset B satisfies the global relational constraint dictated by q (*e.g.*, spatiotemporal continuity, identity consistency).
- $\gamma \to \infty$ represents a strict hard constraint penalty (*i.e.*, a bundle that breaks the relational logic is fundamentally invalid).

G.2 Lemma 1: Strict Non-Submodularity of Relational Bundles

In traditional retrieval and subset selection problems, the scoring function is often assumed to be submodular, allowing greedy algorithms to yield bounded approximation guarantees. We demonstrate that IBC violates this foundational assumption.

Definition (Submodularity). Let V be a finite set. A set function $\Phi : 2^V \to \mathbb{R}$ is submodular if and only if for every $A \subseteq B \subset V$ and $x \notin B$, it satisfies the diminishing returns property:

$$\begin{aligned} &\Phi(A \cup \{x\}) - \Phi(A) \\ &\geq \Phi(B \cup \{x\}) - \Phi(B) \end{aligned} \quad (10)$$

Lemma 1. *The joint relevance scoring function $\Phi(B, q)$ for IBC is strictly non-submodular when subject to relational structural constraints.*

Proof. We prove this by construction. Consider a temporal event progression query q (*e.g.*, the wedding progression in Figure 8). Let the target optimal bundle consist of three chronological stages: $B^* = \{x^*_{setup}, x^*_{wait}, x^*_{ceremony}\}$.

Let subset $A = \{x^*_{setup}\}$ and subset $B = \{x^*_{setup}, x^*_{wait}\}$, where $A \subset B$. We now introduce a new candidate image $x^*_{ceremony}$ (a ceremony shot from the exact same wedding).

1. When $x^*_{ceremony}$ is added to A, the subset becomes $\{x^*_{setup}, x^*_{ceremony}\}$. Because the intermediate temporal bridge (x^*_{wait}) is missing, the narrative progression is broken, yielding $\mathbb{I}_{rel}(A \cup \{x^*_{ceremony}\}) = 0$. The marginal gain is strictly the local semantic score:

$$\begin{aligned} &\Phi(A \cup \{x^*_{ceremony}\}) - \Phi(A) \\ &= f(x^*_{ceremony}, q_{ceremony}) \end{aligned} \quad (11)$$

2. When $x^*_{ceremony}$ is added to B, the subset becomes the complete chronological story $\{x^*_{setup}, x^*_{wait}, x^*_{ceremony}\}$. The narrative is

successfully closed, satisfying the joint constraint $\mathbb{I}_{rel}(B \cup \{x^*_{ceremony}\}) = 1$. The marginal gain is:

$$\begin{aligned} &\Phi(B \cup \{x^*_{ceremony}\}) - \Phi(B) \\ &= f(x^*_{ceremony}, q_{ceremony}) + \gamma \end{aligned} \tag{12}$$

Since $\gamma > 0$ denotes the massive reward for satisfying the hard relational constraint, we obtain:

$$\begin{aligned} &\Phi(A \cup \{x^*_{ceremony}\}) - \Phi(A) \\ &< \Phi(B \cup \{x^*_{ceremony}\}) - \Phi(B) \end{aligned} \tag{13}$$

This strict inequality violates the diminishing returns property of submodular functions. ■

Corollary. Since Φ is strictly non-submodular, any greedy selection algorithm relying on independent marginal gains cannot guarantee a constant-factor approximation bound. This necessitates an adaptive, context-aware expansion strategy like Bundle-Search.

G.3 Theorem 1: The Relational Blindness Bound of Independent Top-k

Building upon Lemma 1, we formally establish the failure mechanism of the Decompose-and-Rerank paradigm. In this paradigm, the system independently retrieves the top-k candidates for each sub-query q_i, forming a local candidate space $\mathcal{S}_i = \arg\max_{S \subset \mathcal{I}, |S|=k} \sum_{x \in S} f(x, q_i)$.

Theorem 1. *Assume that for any target image $x^*_i \in B^*$, there exist M "Dominant Distractors" $x_{i,d}$ in the massive unindexed image pool $\mathcal{I}$. These distractors satisfy:*

1. *Higher local semantic alignment: $f(x_{i,d}, q_i) > f(x^*_i, q_i)$.*
2. *Violation of global relational constraints: For any bundle B' containing $x_{i,d}$, $\mathbb{I}_{rel}(B') = 0$.*

If the retrieval system's cutoff parameter satisfies $k \leq M$, the probability $P(B^*)$ of successfully discovering the optimal bundle B^* via the Decompose-and-Rerank paradigm is strictly 0.

Proof. During the independent retrieval stage, the global constraint $\mathbb{I}_{rel}$ is invisible to the atomic scorer. Because there are M dominant distractors satisfying $f(x_{i,d}, q_i) > f(x^*_i, q_i)$, the greedy top-$k$ selection will exclusively populate $\mathcal{S}_i$ with these distractors (since $k \leq M$). Consequently, the true target image is deterministically excluded from the candidate pool: $x^*_i \notin \mathcal{S}_i, \forall i \in \{1, \ldots, K\}$.

During the selection stage, the VLM is restricted to search within the Cartesian product $\mathcal{S}_{global} = \mathcal{S}_1 \times \mathcal{S}_2 \times \cdots \times \mathcal{S}_K$. Since every candidate bundle $B_{cand} \in \mathcal{S}_{global}$ contains at least one dominant distractor $x_{i,d}$, it follows from our assumption that:

$$\forall B_{cand} \in \mathcal{S}_{global}, \quad \mathbb{I}_{rel}(B_{cand}) = 0 \tag{14}$$

Therefore, the system is mathematically forced to output a fragmented composite with a joint relevance score significantly lower than $\Phi(B^*, q)$. Unless the cutoff k is expanded such that $k > M$ (which causes the VLM's combinatorial search space $\mathcal{O}(k^K)$ to explode exponentially, rendering it computationally intractable), the independent splitting strategy will inevitably succumb to Relational Blindness. ■

Remark on Empirical Observations. Theorem 1 directly explains the catastrophic drop in Exact Match (EM) rates for advanced VLMs reported in Table 1. As vividly illustrated in Case Study 2 (Figure 8), visually stunning wedding photos from *different* events act precisely as *Dominant Distractors* ($x_{i,d}$). They easily overwhelm the local top-k ranking due to high semantic similarity to the sub-query, entirely displacing the authentic, identity-consistent images (x^*_i) required to fulfill the global relational constraint.

Case 1. Query: Find a sequence of photos documenting a single-day walking route through the Dutch Overijssel countryside, progressing from a village church across green pastures, along a canal path with a hiker, past traditional thatched-roof farmsteads, through a wooded estate, and ending at a river weir.

(a) Multimodal Embedding: RzenEmbed.
Too similar images, not covering all query intent, e.g. river weir.

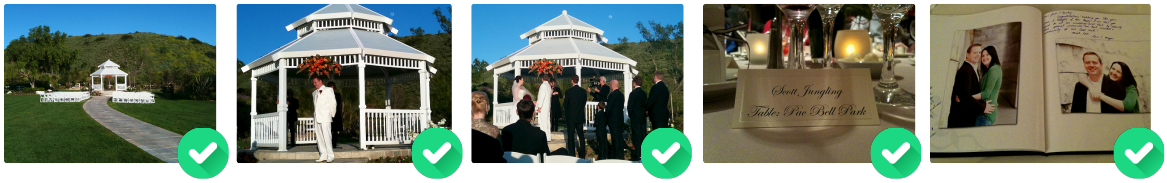
(b) VLM Two-Stage Decompose & Rerank: Claude-Sonnet-4.5.
Some images not from the same user, or not taken in Dutch Overijssel !

(c) BundleWeaver. Successfully construct the right bundle!

Case 2. Query: Find a set of key photos documenting the progression of a wedding day, from the empty ceremony setup through the groom waiting, the ceremony itself, the reception table setting, and the guest book with the couple's engagement photos.

(d) Multimodal Embedding: RzenEmbed.
Too similar images, not covering all query intent, e.g. guest book.

(e) VLM Two-Stage Decompose & Rerank: Claude-Sonnet-4.5.
Some images not from the same wedding !



(f) BundleWeaver. Successfully construct the right bundle!

Figure 8: Examples showing the results of different methods on IBC.