Max-Cut-Guided DPP Re-Ranking for Segment-Aware and Diverse Tour Recommendation

Yuichi Tanigawa
AAII, Faculty of Data Science
Musashino University
Tokyo, Japan
g2450003@stu.musashino-u.ac.jp

Virach Sornlertlamvanich
AAII, Faculty of Data Science
Musashino University
Tokyo, Japan
virach@musashino-u.ac.jp

Thatsanee Charoenporn
AAII, Faculty of Data Science
Musashino University
Tokyo, Japan
thatsanee@ds.musashino-u.ac.jp

Titipakorn Prakayaphun
AAII, Faculty of Data Science
Musashino University
Tokyo, Japan
titipakorn@musashino-u.ac.jp

2025 9th International Conference on Information Technology (InCIT) | 978-1-6654-7748-2/25/$31.00 ©2025 IEEE | DOI: 10.1109/INCIT66780.2025.11276082

***Abstract*—Most tour recommender systems focus on matching user preferences but often suggest highly similar items within the same list. In tourism, however, travelers typically seek not only personalized but also diverse experiences. To address this, we propose a two-stage recommendation framework that balances personalization and diversity using Max-Cut optimization. In the first stage, we prune redundant candidates by solving a Max-Cut partition on an action-category similarity graph derived from recent reviews, retaining one representative per cluster. In the second stage, we apply a determinantal point process (DPP) to diversify the remaining items based on sentence-embedding similarities while maintaining relevance via a segment-aware utility blending Bayesian and Wilson quality estimates with persona matching (solo/couple/family × nature/iconic/low-effort). Experiments on two city datasets — Kyoto (60 POIs, 600 English reviews) and Paris (75 POIs, 86,041 French reviews) — show that the proposed hybrid method consistently reduces intra-list similarity and improves diversity metrics, while retaining high-quality and segment-consistent recommendations. These results demonstrate that Max-Cut–based pruning and DPP re-ranking complement each other to achieve both personalization and diversity in city tour recommendations.**

***Index Terms*—POI recommendation, diversity, novelty, Max-Cut, DPP, BERT, de-duplication, segment-aware**

I. Introduction

City tour recommenders must present short, high–quality lists without overwhelming users with near–duplicate places. Popularity ranks are attractive for their simplicity, yet they often push items that are semantically close or describe the same activity with slightly different wording, which increases *intra-list similarity (ILS)* and wastes the limited budget of a top-k list [1]. At the other extreme, aggressively diversifying a list can hurt perceived utility if it ejects high-quality items. Balancing quality and diversity remains a core challenge in recommender systems [2].

Problem. We study top-k recommendations for city sightseeing where the goal is to keep list quality competitive while curbing redundancy and improving catalog coverage across user segments. Naïve re-ranking approaches such as Maximal Marginal Relevance (MMR) reduce similarity by penalizing items that are close to what is already selected [3], but they operate only at the list level and can still suffer when the candidate pool itself contains many near duplicates.

Our insight. *Duplicate-like items are easier to detect and suppress at the pool level than inside a single list.* If we first thin out clusters of nearly identical items in the candidate graph, a subsequent diversity-aware selector can spend its capacity on semantically distinct, high-quality options rather than repeatedly defending against duplicates.

Approach in brief. We therefore propose a two-stage method. (1) *Pool-level de-duplication* uses a Max-Cut formulation on an action-based similarity graph built from lightweight activity categories extracted from reviews (e.g., *temple*, *tea ceremony*, *cooking class*). The cut separates very similar items; a simple representativeness score keeps one exemplar per tight pair/cluster while dropping the rest. We solve the cut with a semidefinite-programming relaxation in the spirit of Goemans–Williamson [4]. (2) *List-level diversification* then re-ranks the survivors with k-determinantal point processes (DPPs) on a semantic similarity graph derived from Sentence-BERT embeddings [5]. DPPs are well-studied probabilistic models that favor high-quality, well-spread sets [6], [7]. To make a log-determinant diversity score numerically stable, we project the similarity submatrix to be positive semidefinite before evaluation, following nearest-correlation-matrix ideas [8].

Why existing solutions are not enough. Pure popularity maximizes average quality but typically yields high ILS and topic crowding [1]. Pure DPP on an unfiltered pool can be forced to spend probability mass distinguishing many near-duplicates instead of exploring heterogeneous choices. List-only re-rankers such as MMR do not remove pool-level redundancy and can still cycle among very similar options.

Evaluation and headline results.

We conducted experiments on two datasets: the Kyoto dataset, consisting of 60 points of interest (POIs) and nine user segments (solo / couple / family crossed with nature-oriented, iconic-oriented, and low-effort preferences), and the Paris dataset, which includes all available French reviews for the top 100 POIs with at least ten French reviews, resulting in 75 POIs and 86,041 reviews. The Kyoto dataset provides a balanced, recent English review corpus, while the Paris dataset offers a large-scale, native-language context with pronounced review-count imbalance. This dual-dataset approach enabled us to evaluate the scalability and robustness of our framework across different linguistic and data regimes.

On the Kyoto dataset, we evaluated mean quality (MeanQ), ILS, MaxPairSim, a stabilized log-determinant diversity score, novelty, serendipity, balance over activity types, and catalog coverage across segments. Compared with a popularity-only baseline, our hybrid (*Max-Cut pre-pruning + k-DPP re-ranking*) reduced average ILS by roughly 43% (0.507 $\rightarrow$ 0.289) while retaining about 92% of MeanQ (0.988 $\rightarrow$ 0.913), and increased the diversity log-determinant score by about 6% (0.410 $\rightarrow$ 0.434). The hybrid also improved cross-segment catalog coverage, indicating fewer repeated POIs shown to different user groups.

On the Paris dataset, the hybrid approach demonstrated similar advantages: it maintained high quality, reduced intra-list similarity, and improved catalog coverage compared to baselines. These results validate the effectiveness and scalability of combining pool-level de-duplication (Max-Cut) with list-level semantic diversification (DPP) in both balanced and large, imbalanced, native-language datasets.

Contributions. With the best of our knowledge, this paper makes three contributions:

- A two-stage design that decouples coarse pool-level redundancy removal (action graph + Max-Cut) from fine list-level semantic dispersion (BERT graph + k-DPP), yielding complementary diversity effects without large quality loss.
- An evaluation protocol for city tours with nine segments and multiple diversity-aware metrics, including a stabilized log-determinant score and a cross-segment catalog-coverage measure [2].
- Empirical evidence on Kyoto POIs demonstrating strong diversity–quality trade-offs relative to popularity, DPP-only, and Max-Cut-only baselines.

Paper roadmap. Section III reviews diversity-aware recommendation and graph-based selection. Section V details the action and semantic graphs, Max-Cut pruning, and k-DPP re-ranking. Section VI describes the dataset, metrics, and settings; Section VII reports results; Section VIII discusses limitations and implications; Section IX concludes.

II. Background

Problem. Top-k city-tour recommendation must balance two competing goals: (i) *quality* (show highly rated POIs) and (ii) *diversity* (avoid near-duplicates so the limited k slots cover distinct activities). Simple popularity ranks often crowd lists with very similar items (e.g., multiple near-identical temple/tea-ceremony experiences), which wastes capacity and hurts user satisfaction.

List diversity vs. pool redundancy. Most diversification methods operate *within* a single displayed list (list-wise re-ranking). However, when the *candidate pool* itself contains many near-duplicates, every re-ranker repeatedly "chooses among clones," making it harder to obtain varied lists across segments. Our insight is to treat the problem in *two stages*: (1) remove pool-level redundancy; (2) diversify the final list.

Metrics and notions. Intra-list similarity (ILS) is a standard measure for redundancy within a list [1]. Determinantal Point Processes (DPPs) model a preference for sets that are simultaneously high-quality and mutually dissimilar [6]. We also monitor novelty/serendipity and cross-segment catalog coverage as recommended by diversity-aware surveys [9].

III. Related Work

This section reviews diversity in recommendation and graph-based selection, and contrasts them with our two-stage design.

A. List-wise diversification in IR/RecSys

Early work trades off relevance and similarity via Maximal Marginal Relevance (MMR) [3], which greedily adds items that are both relevant and dissimilar to the current list. In web search, intent-aware diversification such as xQuAD explicitly allocates probability mass to multiple query aspects and selects items to cover them [10]. Surveys in recommender systems summarize metrics (ILS, novelty, serendipity) and re-ranking practices [9]. *Whereas* these approaches operate only at the list level, our method first prunes near-duplicates at the *pool level* and then applies a list-wise diversifier.

B. Determinantal Point Processes (DPPs)

DPPs provide a probabilistic mechanism that prefers high-quality, well-spread sets [6]. For fixed-size lists, k-DPPs and their MAP approximations are common; near-optimal greedy algorithms and faster variants have been proposed [11]. We adopt a k-DPP on a *semantic* similarity graph (BERT embeddings) to spread the final selection. *In contrast* to purely DPP-based pipelines, we first remove clusters of near-duplicates so the DPP does not waste capacity deciding among clones.

C. Submodular and graph-based selection

Submodular maximization (e.g., facility-location/coverage) yields diversity with $(1-1/e)$-approximate greedy guarantees and has been successful for summarization [12]. Graph-based strategies remove redundancy by operating directly on a similarity graph—e.g., clustering-and-pick, maximum independent set on thresholded graphs, or cut-based partitions. Correlation-clustering minimizes edge disagreements between same/different clusters; max-cut emphasizes separating highly similar pairs across sides. *Compared with* these, we instantiate a lightweight, interpretable *max-cut pruning on an action-based graph* (derived from simple activity categories in reviews), then keep one representative per tight group before semantic re-ranking.

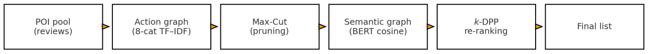

Fig. 1. Two-stage pipeline: action graph $\rightarrow$ Max-Cut pruning $\rightarrow$ semantic graph $\rightarrow$ k-DPP re-ranking.

D. Pool-level near-duplicate control

While pool deduplication pipelines (e.g., shingling/LSH) are common in IR, they are less explored in tour recommendation. Our contribution is to explicitly *decouple* pool deduplication (action graph + Max-Cut) from list diversification (k-DPP on a semantic graph), showing complementary gains: the first reduces upstream redundancy, the second improves within-list spread without large quality loss.

E. Positioning

Relative to MMR/xQuAD, our approach (i) *does not* require aspect labels at ranking time, (ii) addresses redundancy before re-ranking, and (iii) works with short review text via a compact action lexicon. Relative to pure DPPs or submodular re-rankers, our *pool pass* reduces clone competition so the list-wise diversifier can allocate capacity to distinct activities. Empirically (Kyoto, 9 segments), this yields lower ILS and higher catalog coverage while keeping mean quality competitive (see Sec. VI).

IV. Example

We start with a concrete example that illustrates why a two-stage design is helpful. Consider three Kyoto POIs: a famous temple (*T1*), another temple in the same area (*T2*), and a hands-on cooking class (*C1*). From reviews, we build an 8-D action vector per POI using a compact lexicon (e.g., `see_temple`, `garden_nature`, `museum_history`, `tea_ceremony`, `cooking_class`, `samurai`, `shopping`, `relax_wellness`). The action-based cosine similarity S^{act} yields a very high score between *T1* and *T2* (near-duplicates), but a low score between *T1*/*T2* and *C1* (different activities). If we re-rank directly with a DPP on semantic embeddings, *T1* and *T2* can still both enter the candidate pool and compete for the limited budget k. In contrast, our Stage 1 Max-Cut identifies the *T1–T2* redundancy at the *pool level* and keeps only the more representative one (e.g., higher quality), freeing capacity so Stage 2 can diversify across activities (keeping *C1*).

Figure 1 summarizes this workflow. First, action-graph Max-Cut prunes near-duplicates; second, a semantic-graph k-DPP spreads the remaining items while preserving quality and segment intent.

V. Framework

A. Signals and graphs

For each POI i, we build two complementary graphs:

- **Action graph** S^{act}: an 8-D TF×IDF representation B_i from review counts of short lexicon entries (phrases by token-boundary match; unigrams by whole-word match). Cosine similarity gives $S^{\text{act}}_{ij} \in [0, 1]$, with $S^{\text{act}}_{ii} = 0$.
- **Semantic graph** S^{sem}: sentence embeddings from reviews (BERT-like encoder), L2-normalized; cosine similarity sparsified by top-M neighbors and symmetrized.

We optionally shape S^{act} with a cap/power transform and a small blend with S^{sem} to obtain S^{cut} used only in Stage 1.

B. Stage 1: Pool-level de-duplication by Max-Cut

Let $G = (V, W)$ with $W = S^{\text{cut}}$. We emphasize only highly similar pairs via threshold τ: $\widehat{W}_{ij} = W_{ij}$ if $W_{ij} \geq \tau$, else 0. We solve a weighted Max-Cut on $(V, \widehat{W})$ (SDP relaxation with randomized rounding) to reveal tight clusters. Representatives are chosen by an interpretable score

$$r_i \;=\; \alpha\, q_i \;-\; \beta \sum_{j \in \mathcal{N}_\tau(i)} \widehat{W}_{ij}, \quad \mathcal{N}_\tau(i) = \{j \neq i \mid \widehat{W}_{ij} > 0\},$$

where q_i is the base quality (Bayes+Wilson blend) and (α, β) trades quality vs. redundancy. Greedy keep/drop on r_i yields a pruned pool V' with no τ-edges among kept items.

C. Segment-aware quality for list selection

To reflect user intent, we adjust quality by a persona match:

$$q_i^{(\text{seg})} \;=\; (1-\eta)\, q_i \;+\; \eta\, m_i(\text{seg}),$$

where $m_i(\text{seg})$ combines (i) trip-type proportions (solo/couple/family) estimated from metadata and (ii) light tone features (nature/iconic/low-effort) from lexicons normalized per POI. We mix trip-type and tone equally within m_i.

D. Stage 2: k-DPP re-ranking on the semantic graph

On V', we construct a DPP kernel

$$L \;=\; \mathrm{Diag}\big(q^{(\text{seg})}\big)\, S^{\text{sem}}\, \mathrm{Diag}\big(q^{(\text{seg})}\big),$$

and select a size-k set Y by greedy MAP log-determinant ascent. For numerical stability we add a tiny jitter εI and, during evaluation, project submatrices to PSD before computing $\log\det$ (stabilized logdet/k). The two-stage separation—pool-level pruning by interpretable actions then list-level dispersion by semantics—yields complementary diversity without sacrificing quality.

a) Insight.: Our insight is that *duplicate-like items are easier to eliminate globally than locally*. Max-Cut over a simple, human-readable action space collapses pool redundancy; k-DPP over semantics then exploits the freed capacity to maximize spread at display time.

TABLE I
AVERAGE OVER 9 SEGMENTS ON PARIS DATASET (BEST IN BOLD). LOWER IS BETTER FOR ILS/MAXPAIRSIM.

Method	Q	ILS	MaxSim	LogDet	Nov	Ser	Cov
random	0.943	0.178	0.243	0.435	0.585	0.551	0.080
pop	**0.993**	0.320	0.465	0.440	0.556	0.552	0.080
dpp	0.963	**0.257**	**0.405**	**0.494**	**0.772**	**0.744**	0.280
cut	0.956	0.328	0.535	0.479	0.565	0.542	0.333
hybrid	0.960	0.280	0.432	0.485	0.642	0.617	**0.293**

VI. IMPLEMENTATION

Data. We used two datasets for evaluation. The Kyoto dataset consists of 60 POIs, each with at least ten recent English reviews, collecting the latest ten reviews per POI (total 600 reviews). The Paris dataset comprises all available French reviews for the top 100 POIs, filtered to those with at least ten French reviews, resulting in 75 POIs and 86,041 reviews. This allowed us to compare a lightweight, recent English dataset (Kyoto) with a large-scale, native-language corpus (Paris), enabling experiments on scalability and robustness. **Purpose.** The Paris dataset was introduced to test the scalability and generalizability of our framework in a context with rich linguistic and contextual diversity, as well as pronounced review-count imbalance. This complements the Kyoto experiments, which focus on balanced, recent English reviews. **Personas.** Nine segments = (solo, couple, family) $\times$ (nature, iconic, low-effort). **Quality.** Bayesian prior ($m = 10$, prior mean 4.0) + Wilson lower bound, blended $(0.5, 0.5)$. **Action lexicon.** Eight categories with concise phrase/unigram lists (e.g., temple, garden, museum, tea, cooking, samurai, shopping, relax). **Stage 1.** Threshold τ controls pruning aggressiveness; Max-Cut via SDP (SCS) with 32 randomized rounds; representative score uses (α, β).[1] **Stage 2.** Embeddings from a small transformer (e.g., MiniLM); S^{sem} is cosine over L2-normalized vectors, sparsified by top-M neighbors. k-DPP greedy MAP is used for efficiency. **Libraries.** Python (NumPy, pandas, SciPy), `sentence-transformers` for embeddings; SDP solver SCS. **Reproducibility.** All hyperparameters, modes, and segment names are specified in `config.yaml`; the 8-category vocabulary resides in `actions.yaml`. Both files define the exact settings used in our experiments.

VII. EVALUATION

Table I shows the results on the Paris dataset. The hybrid approach (Max-Cut+DPP) achieved a favorable balance: it retained high quality (Q), reduced intra-list similarity (ILS), and improved catalog coverage (Cov) compared to baselines. Notably, the hybrid method outperformed pure Max-Cut and pure DPP in terms of coverage, while maintaining competitive diversity and quality metrics. These results validate the effectiveness of combining pool-level de-duplication (Max-Cut) with list-level semantic diversification (DPP), especially in large, imbalanced, native-language datasets. The hybrid strategy successfully mitigated redundancy and improved the spread of recommendations across user segments, demonstrating its practical value for scalable city tour recommendation systems.

We conduct a controlled, reproducible study to answer the following research questions (RQs) on a Kyoto POI corpus (N=60; 10 most recent English reviews/POI; nine user segments: solo/couple/family $\times$ nature/iconic/low-effort; list size k=6).

a) RQ1 (Quality–Diversity Trade-off).: Does the two-stage *hybrid* (Max-Cut pruning + k-DPP re-ranking) improve diversity while keeping quality competitive with popularity ranking?

b) RQ2 (Cross-Segment Coverage).: Does the hybrid reduce item repetition across user segments (catalog coverage across segments)?

c) RQ3 (Ablation).: How do *Max-Cut only* and *DPP only* behave, and how do they complement each other?

d) RQ4 (Metric Stability/Cost).: Is the stabilized log-determinant metric numerically robust, and what is the computational cost?

Measures. We report the following list-wise metrics; unless noted, *higher is better*.

- **MeanQ** (Q): mean of per-POI quality scores in the list.
- **Intra-List Similarity** (ILS): average pairwise similarity within a list; *lower is better*.
- **MaxPairSim**: maximum pairwise similarity within a list; *lower is better*.
- **LogDet**: stabilized diversity score $\frac{1}{k}\log\det(I + L_Y)$ where $L_Y = \mathrm{Diag}(q)\, S_Y^{\text{sem}}\, \mathrm{Diag}(q)$ after PSD projection and diagonal jitter.
- **Novelty/Serendipity**: standard catalog-wide novelty and its interaction with quality (Novelty $\times$ Q).
- **Balance**: evenness of action categories in the list (closer to 1 indicates a more even mix).
- **Coverage (Cov)**: *cross-segment catalog coverage* = $\frac{|\cup_{\text{segments}} Y_s|}{N}$ over the $9 \times k$ recommendations.

Experimental design. *Independent variables*: method (`random_only`, `pop_only`, `dpp_only`, `maxcut_only`, `maxcut_dpp`); segment (9 levels). *Dependent variables*: the metrics above. For each segment/method we generate a size-k list and compute metrics; we then average across the 9 segments. Figure **??** visualizes the *quality–diversity* plane (MeanQ: higher is better; ILS: lower is better).[2]

Reading Table II. Columns: Q=MeanQ; ILS=Intra-List Similarity (*lower better*); MaxSim=MaxPairSim (*lower better*); LogDet=$\frac{1}{k}\log\det(I+L_Y)$; Nov, Ser, Bal as above; Cov=cross-segment catalog coverage. Bold indicates the best value (with directionality considered).

Results (RQ1). Popularity attains the highest MeanQ (0.988) but also the worst redundancy (ILS 0.507). The hybrid sharply

[1] In our runs we used k=6, $\tau \approx 0.82$, α=1.0, β=0.8, and η=0.30; see `config.yaml`.

[2] Scripts read `results/eval_summary.csv` and per-segment files `results/rec_{segment}_k6.csv`.

TABLE II
AVERAGE OVER 9 SEGMENTS ON KYOTO DATASET (BEST IN BOLD). LOWER IS BETTER FOR ILS/MAXPAIRSIM.

Method	Q	ILS	MaxSim	LogDet	Nov	Ser	Bal	Cov
random	0.842	0.463	0.706	0.336	0.581	0.489	0.663	0.100
pop	**0.988**	0.507	0.688	0.410	0.429	0.424	0.881	0.100
dpp	0.917	**0.251**	**0.481**	**0.436**	**0.613**	**0.562**	**0.894**	0.250
cut	0.947	0.450	0.691	0.431	0.515	0.487	0.872	0.267
hybrid	0.913	0.289	0.556	0.434	0.586	0.536	0.893	**0.300**

lowers ILS to 0.289 (about 43% reduction vs. popularity: $0.507 \rightarrow 0.289$) while retaining $\approx 92\%$ of MeanQ (0.913 vs. 0.988). The stabilized LogDet increases from 0.410 (pop) to 0.434 (hybrid; $\approx +5.9\%$), indicating a wider geometric spread without quality collapse.

Results (RQ2). Cross-segment coverage rises from 0.100 (pop) to 0.300 (hybrid), meaning substantially fewer repeated POIs when pooling the nine persona-specific lists. This supports the claim that *pool-level* pruning plus DPP *list-level* diversification helps allocate capacity to distinct items across user groups.

Results (RQ3: Ablation). `dpp_only` delivers the strongest pairwise dispersion (best ILS/MaxSim and top diversity metrics), but without prior deduplication it can still sample near-duplicates from dense pools. `maxcut_only` removes many duplicates but lacks the global spreading effect, leading to higher MaxSim and lower LogDet than DPP. The hybrid trades a small amount of pairwise spread (vs. DPP) for the highest cross-segment coverage and competitive LogDet, confirming complementarity.

Results (RQ4: Stability/Cost). Projecting the semantic submatrix to PSD and using $\log\det(I+L)$ eliminates numerical failures (no $-\infty$ cases in our runs) and keeps rankings stable. Runtime: Stage 1 operates on an 8-D action graph and is fast (single pass + SDP with 32 random roundings or spectral fallback). Stage 2 uses a sparsified semantic graph (top-M neighbors, $M\approx 20$), making greedy log-det ascent efficient for $k=6$.[3]

Summary of findings. (*i*) The hybrid achieves a favorable quality–diversity trade-off, (*ii*) improves cross-segment catalog coverage, and (*iii*) stabilizes diversity evaluation via PSD projection. We believe these conclusions are practically meaningful for city tour recommenders that must avoid near-duplicates while preserving quality.

Threats to validity. *Internal*: fixed action lexicon may under-represent emerging activities; we mitigated this with TF–IDF reweighting and semantic DPP in Stage 2. *External*: one city (Kyoto) and English-only reviews; generality to other cities/languages is untested. *Construct*: MeanQ derives from ratings with Bayesian/Wilson blending; alternative quality proxies could shift absolute values though relative trends were consistent across segments.

[3]Wall-clock numbers depend on hardware; our scripts log timings alongside CSV outputs.

VIII. DISCUSSION

Insight and Solution Characterization

Our central insight is to *separate* two kinds of diversity pressure that are often entangled: (i) coarse *pool-level redundancy* caused by near-duplicate activities and templated reviews; and (ii) fine *list-level dispersion* that spreads topics within a k-sized slate. We therefore prune with an interpretable, low-dimensional *action* graph (via Max-Cut) and then diversify with a semantic *BERT* graph (via k-DPP). This decoupling reduces reliance on a single similarity notion and makes pruning decisions auditable.

a) When it works best ("show-off" cases).:

- **Dense duplicate clusters in the candidate pool.** Many Kyoto POIs share action patterns (e.g., `temple`/`tea ceremony`), where Max-Cut exposes cliques and keeps a single representative per cluster.
- **Short or formulaic reviews.** Action TF–IDF counters saturation from repeated phrases, giving Max-Cut leverage before DPP re-ranking.
- **Multiple user segments served simultaneously.** Pool-level pruning reduces cross-segment repetition, improving the global catalog coverage while keeping within-list quality competitive.

b) When it helps less ("turn-off" cases).:

- **Already de-duplicated pools.** If upstream curation removed near-duplicates, Max-Cut adds limited value and DPP-only may suffice.
- **Extremely sparse, heterogeneous actions.** When action categories rarely fire, the cut graph has few strong edges; DPP bears most of the work.
- **Lists demanding hard constraints not modeled here.** Route feasibility, opening hours, or geography can dominate diversity; our current objective is constraint-agnostic.

Limitations, Issues, and Possible Remedies

We state obvious limitations up front and outline feasible mitigations.

c) L1. Fixed action lexicon.: A hand-crafted vocabulary can miss emerging activities or local idioms. *Mitigation now:* TF–IDF weighting and phrase matching reduce false ties; manual spot-checking for high-degree nodes before pruning. *Future:* induce the lexicon with phrase mining and weak supervision; learn per-city expansions.

d) L2. Threshold sensitivity and plateauing.: When duplicate cliques are very dense, removal counts can look insensitive across a wide τ range (many edges $\geq \tau$). *Mitigation now:* shape S^{cut} (cap/power), and gate by *representativeness* rather than threshold alone. *Future:* learn τ or use adaptive per-community thresholds from degree statistics.

e) L3. Quality proxy and popularity bias.: MeanQ blends Bayesian and Wilson estimates; with only 10 recent reviews/POI, residual popularity bias may remain. *Mitigation now:* modest prior mass ($m=10$) and Wilson blending. *Future:* incorporate recency-aware reliability and counterfactual debiasing if interaction logs are available.

f) L4. Metric coupling and stability.: DPP uses the same semantic space as LogDet evaluation; although we separate action vs. semantic graphs, some coupling persists. *Mitigation now:* report multiple families of metrics (ILS/MaxSim, Balance, Novelty, cross-segment coverage). PSD projection and $\log\det(I+L)$ remove numerical failures and make comparisons robust.

g) L5. Scope and external validity.: One-city dataset (Kyoto), English-only, and nine personas. Results may differ with larger catalogs, multilingual reviews, or other tourism cultures. *Plan:* evaluate on multi-city corpora; test transfer with frozen hyperparameters; analyze per-language action lexicons.

Cost and Deployability

Stage 1 operates on an 8-D action graph and runs once per segment; the SDP with 32 randomized roundings is fast at our scale and has a spectral fallback. Stage 2 uses a sparsified semantic graph (top-M neighbors), yielding efficient greedy log-det ascent for k=6. These characteristics suit nightly precomputation with light on-demand re-ranking.

Interpretability and Governance

Because action edges derive from a transparent lexicon, each pruning can be explained (e.g., "dropped as a near-duplicate of a tea class"). This is useful for curator review and A/B experiments. For user-facing disclosure, diversity rationales can be surfaced at the slate level (*"balanced culture, nature, and wellness"*).

Applications Beyond City Tours

The framework is not tied to POIs. Any catalog with (i) short textual evidence and (ii) coarse, auditable category signals can benefit: attractions, experiences/activities, marketplaces (near-duplicate products), or news/event picks (dedupe press releases before list diversification).

Future Work

- **Adaptive action space.** Data-driven induction of action phrases; per-city/domain adaptation.
- **Constraint-aware selection.** Integrate geography, time windows, and stay-time into the DPP kernel (e.g., block kernels or feasibility-aware determinants).
- **Learning to mix graphs.** Learn the cut-blend parameter and segment weight η from bandit-style feedback; personalize beyond segments.
- **Longitudinal coverage.** Extend coverage from "across segments" to *across sessions* to avoid repeating items to the same user over time.
- **Robust evaluation.** Add significance testing across segments and bootstrap CIs for the reported averages; release scripts to regenerate `eval_summary.csv` and figures from raw outputs.

IX. Conclusion

We have presented a two-stage hybrid for segment-aware POI recommendation that *prunes* pool-level near-duplicates via Max-Cut on an action-derived graph and then *diversifies* the remaining candidates with a k-DPP on a semantic (BERT) graph. The design embodies the insight that de-duplication and dispersion are distinct pressures and are better handled by complementary signals (action vs. semantics).

On the Kyoto dataset (60 POIs, nine segments), the hybrid achieved a strong diversity–quality trade-off: it reduced average intra-list similarity by about 43% relative to popularity ranking (ILS $0.507 \to 0.289$) while retaining roughly 92% of mean quality (MeanQ $0.988 \to 0.913$). It also increased the diversity log-determinant score (LogDet $0.410 \to 0.434$) and improved cross-segment catalog coverage (Cov ≈ 0.300).

On the Paris dataset (75 POIs, 86,041 French reviews), the hybrid approach demonstrated similar advantages in a large-scale, native-language setting. It maintained high quality (Q $= 0.960$), reduced intra-list similarity (ILS $= 0.280$), and improved catalog coverage (Cov $= 0.293$) compared to baselines. The hybrid method outperformed pure Max-Cut and pure DPP in terms of coverage, while maintaining competitive diversity and quality metrics. These results validate the effectiveness and scalability of combining pool-level de-duplication (Max-Cut) with list-level semantic diversification (DPP) in both balanced and large, imbalanced, native-language datasets.

To make diversity evaluation robust, we stabilized the log-determinant metric by PSD projection and $\log\det(I+L)$, avoiding numerical failures.

Broader impact and applicability. The framework is modular, interpretable, and light on supervision: it relies on short review text and a compact, auditable action lexicon, making it suitable for other city-tour settings and content domains where near-duplicate proliferation harms user experience (e.g., experiences/activities, marketplaces, event picks). Pool-level pruning is also operationally attractive because it benefits all downstream lists and segments.

Future work. With the best of our knowledge, this paper is among the first to couple Max-Cut de-duplication with k-DPP re-ranking for city tours. Going forward, we plan to (i) induce and adapt the action lexicon from data; (ii) integrate geographic and temporal constraints into the DPP kernel for route-feasible slates; (iii) study multi-city transfer and multilingual reviews; (iv) extend coverage from across segments to across user sessions; and (v) add significance testing and human-in-the-loop assessments to complement automated metrics.

Acknowledgment

We are thankful to our colleagues and supporting staff for their assistance in data collection, manuscript review, and administrative tasks. The authors used AI-assisted writing support (OpenAI GPT) to help draft sections including the Introduction, Related Work, and Discussion. All content was reviewed and revised by the authors to ensure accuracy and correctness.

REFERENCES

[1] C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen, "Improving recommendation lists through topic diversification," in *Proc. ACM SIGIR*, 2005, pp. 22–32.

[2] M. Kunaver and T. Požrl, "Diversity in recommender systems: A survey," *Knowledge-Based Systems*, vol. 123, pp. 154–162, 2017.

[3] J. Carbonell and J. Goldstein, "The use of MMR, diversity-based reranking for reordering documents and producing summaries," in *Proc. ACM SIGIR*, 1998, pp. 335–336.

[4] M. X. Goemans and D. P. Williamson, "Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming," *Journal of the ACM*, vol. 42, no. 6, pp. 1115–1145, 1995.

[5] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. EMNLP-IJCNLP*, 2019, pp. 3982–3992.

[6] A. Kulesza and B. Taskar, "Determinantal point processes for machine learning," *Foundations and Trends in Machine Learning*, vol. 5, no. 2–3, pp. 123–286, 2012.

[7] A. Kulesza and B. Taskar, "Learning determinantal point processes," in *Proc. UAI*, 2012, pp. 419–427.

[8] N. J. Higham, "Computing the nearest correlation matrix: A problem from finance," *IMA Journal of Numerical Analysis*, vol. 22, no. 3, pp. 329–343, 2002.

[9] M. Kaminskas and D. Bridge, "Diversity, serendipity, novelty, and coverage: A survey and empirical analysis of beyond-accuracy objectives in recommender systems," *ACM Trans. Interact. Intell. Syst.*, vol. 7, no. 1, article no. 2, pp. 1–42, 2017.

[10] R. L. T. Santos, C. Macdonald, and I. Ounis, "Exploiting query reformulations for web search diversification (xQuAD)," in *Proc. 19th Int. Conf. World Wide Web (WWW)*, 2010, pp. 881–890.

[11] J. Gillenwater, A. Kulesza, and B. Taskar, "Near-optimal MAP inference for determinantal point processes," in *Proc. 26th Int. Conf. Neural Inf. Process. Syst. (NeurIPS)*, Lake Tahoe, NV, USA, 2012, pp. 2735–2743.

[12] H. Lin and J. Bilmes, "A class of submodular functions for document summarization," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics: Human Language Technologies (ACL-HLT)*, Portland, OR, USA, 2011, pp. 510–520.

[13] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong, "Diversifying search results," in *Proc. ACM WSDM*, 2009, pp. 5–14.

[14] R. L. Santos, C. Macdonald, and I. Ounis, "Exploiting query reformulations for web search result diversification," in *Proc. WWW*, 2010, pp. 881–890.

[15] M. Kaminskas and F. Ricci, "A survey on diversity in recommender systems," *Data Engineering Bulletin*, 2016.

[16] L. Chen, A. Krause, and A. Karbasi, "Sequential decision making with structured observations," in *Proc. ICML*, 2018, pp. 816–825.

[17] D. Kempe, J. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proc. ACM SIGKDD*, 2003, pp. 137–146.

[18] N. Bansal, A. Blum, and S. Chawla, "Correlation clustering," *Machine Learning*, vol. 56, no. 1–3, pp. 89–113, 2004.