

1. Motivation

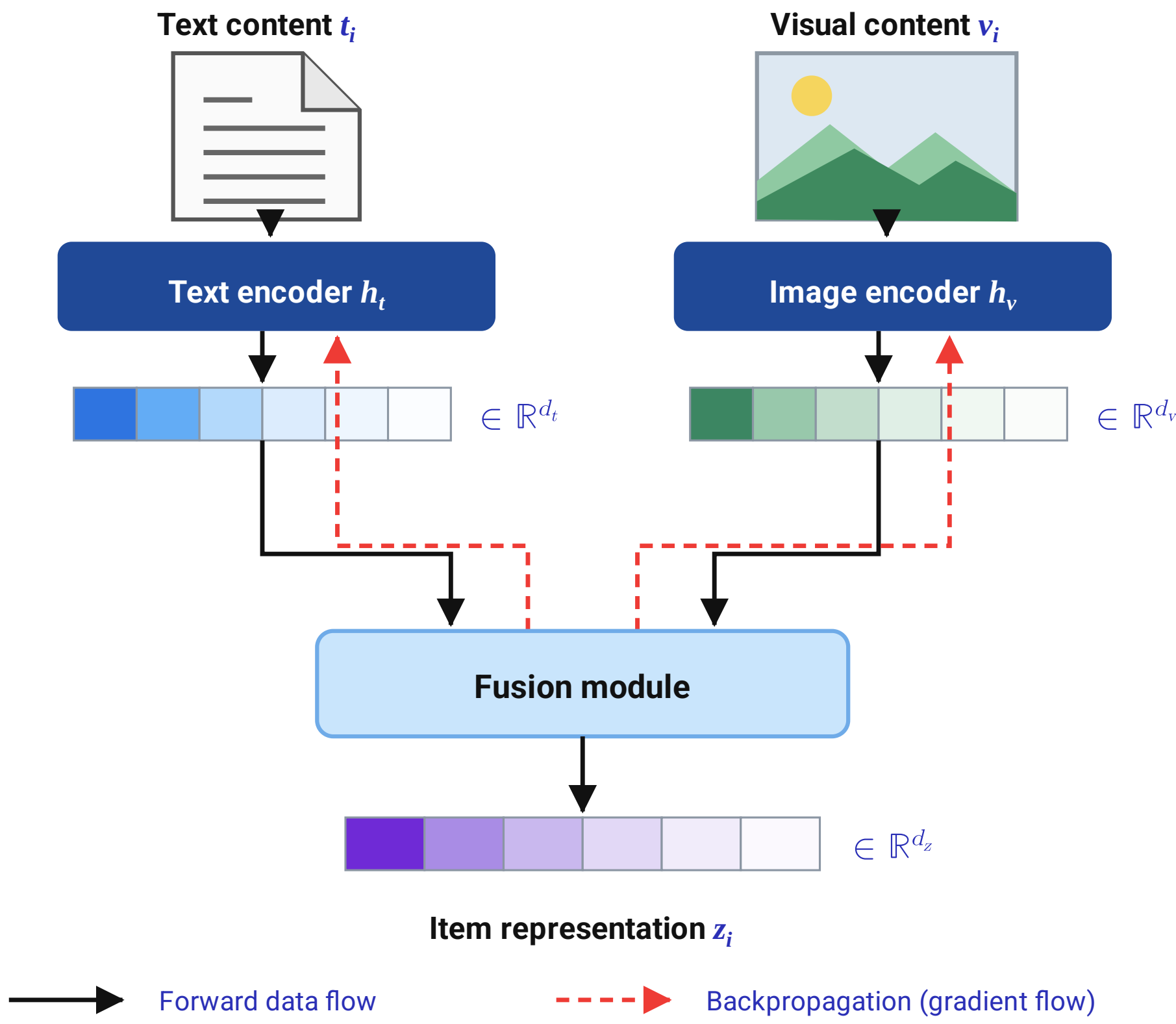
- Modern recommender systems operate in multimodal environments, where items have rich textual and visual content.
- Foundation models provide powerful pretrained encoders, but their role in recommendation pipelines is not yet well understood.
- We study how to train and evaluate multimodal item representations in a unified framework and how they adapt to different recommendation tasks and domains.
- A recommender system connects users with a catalog of candidate items, which can include text, images, and other modalities, based on user interactions.

Research question

How can we learn robust, transferable item representations that improve recommendations across datasets, cold-start settings, and domains?

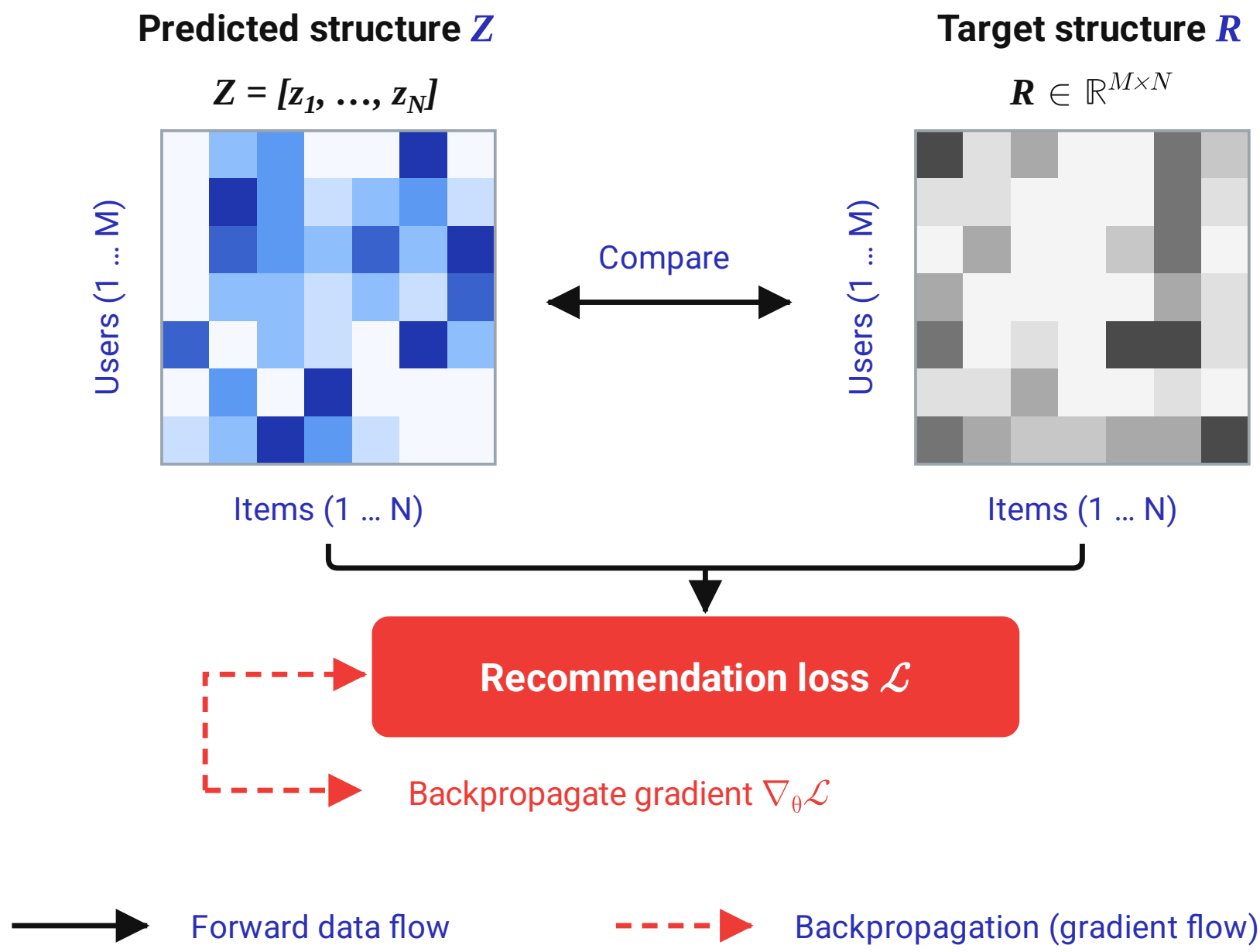
2. Model Architecture

We learn a unified item representation from text and visual content using pre-trained encoders and a fusion module.



3. Training Objective

We optimize the item representations using interaction-derived learning objectives, e.g., contrastive and ranking losses.



4. REEF Evaluation Framework

REEF isolates the contribution of the item backbone and provides a reproducible pipeline for evaluation.

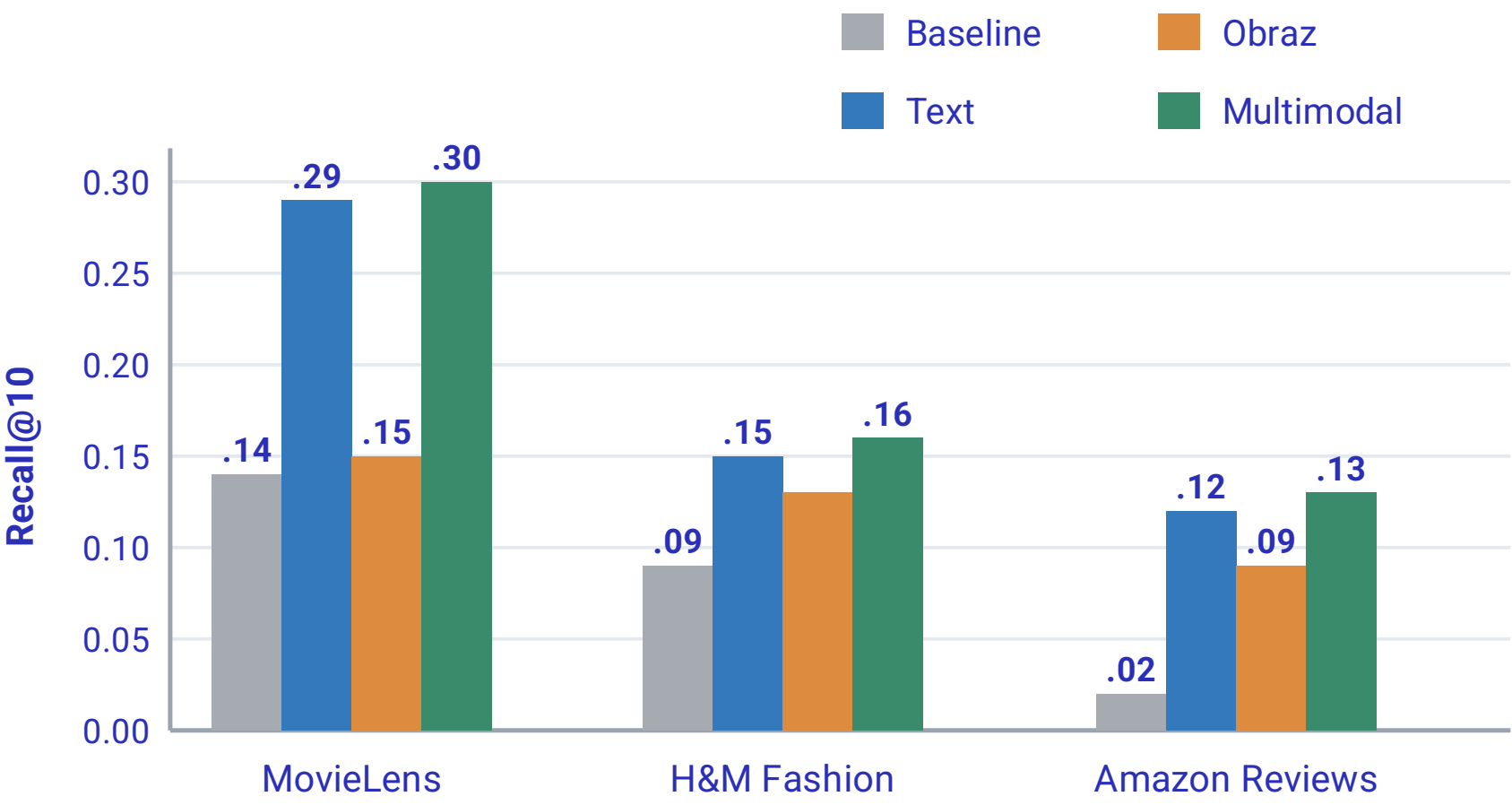


5. Evaluation Scenarios

- Standard recommendation**
General performance on full data
- Cold-start**
New items or users
- Sequential recommendation**
Next-item prediction
- Domain transfer**
Training on one dataset, evaluating on another
- Dataset transfer**
Generalizability across datasets

6. Results

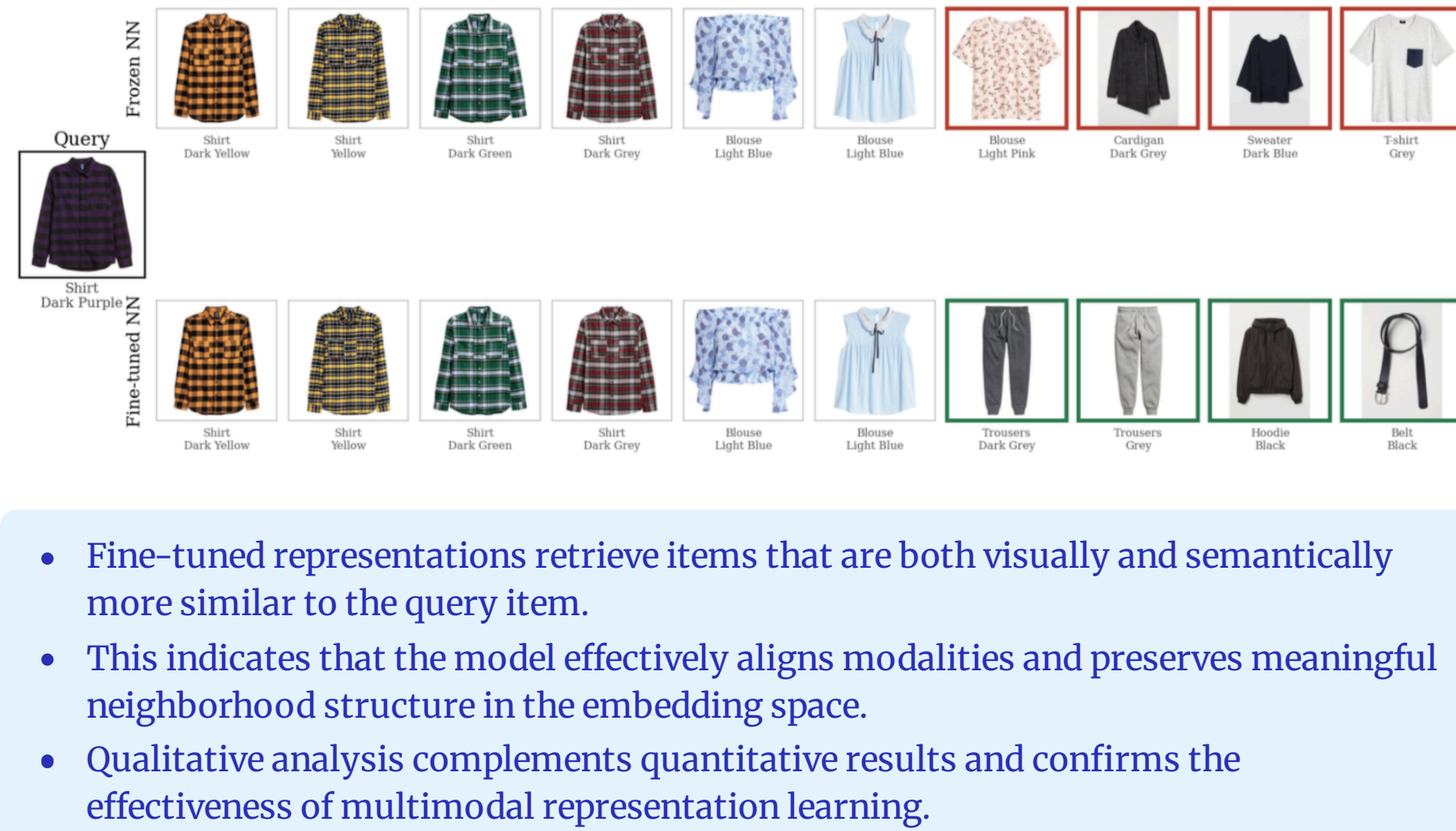
Multimodal item representations consistently outperform single-modality baselines across all datasets. Incorporating both textual and visual content leads to substantial gains, especially on visually rich datasets.



- Multimodal representations bring the largest relative improvement on datasets with high visual content.
- On H&M Fashion, visual content is particularly important, with multimodal representations achieving the best performance.
- Even on the text-heavy Amazon Reviews dataset, multimodal representations outperform all single-modality baselines, showing the robustness of the approach.

7. Representation Analysis

The learned representations preserve semantic structure, which leads to more relevant item neighborhoods after fine-tuning. The example from H&M Fashion shows improved alignment between visually and semantically similar items.



8. Key Findings & Conclusion

- Multimodal item representations improve recommendation performance across tasks and datasets.
- The learned embeddings transfer better to new items, users, and domains.
- A unified evaluation framework enables fair comparison of backbone models.
- Interaction-conditioned training is an effective and scalable strategy.

REEF provides a reproducible framework for understanding and improving foundation models in recommender systems. Multimodal representations and controlled adaptation lead to more robust and transferable recommendations.