Attribute-Conditioned Multimodal Slot Factorization for Controllable Fashion Retrieval

Najmeh Forouzandehmehr
Walmart Global Tech
Sunnyvale, California, USA
najmeh.forouzandehmehr@walmart.com

Topojoy Biswas
Walmart Global Tech
Sunnyvale, California, USA
topojoy.biswas@walmart.com

Evren Korpeoglu
Walmart Global Tech
Sunnyvale, California, USA
evren.korpeoglu@walmart.com

Kannan Achan
Walmart Global Tech
Sunnyvale, California, USA
kannan.achan@walmart.com

Abstract

Fashion retrieval often requires satisfying multiple attributes at once, such as category, color, pattern, and demographic. Monolithic embeddings mix these signals into a single vector, making attribute-specific control difficult at retrieval time. Many existing semantic-ID methods provide discrete item codes, but these codes are typically optimized as item-level or residual addresses and do not expose named, independently controllable attribute slots.

We introduce **MM-slotgate**, a multimodal slot encoder that factorizes Fashion-CLIP text and image embeddings into four named attribute slots. Each slot learns its own text–image gate, so visually grounded attributes such as color and pattern can rely more on image evidence, while taxonomy-oriented attributes such as category and demographic can remain more text-driven.

On H&M, using a combined slot-similarity and slot-logit retrieval score, MM-slotgate achieves 0.7566 macro ConstraintSatisfied@10, outperforming equal-weight multimodal fusion (0.7142) and fCLIP text-only retrieval (0.4755). The largest gain is on color, which improves from 0.321 to 0.889 (+0.568 absolute), as the learned color gate assigns 57.4% weight to image evidence. The learned gates are interpretable without modality supervision: color is image-leaning, category is text-leaning, and pattern and demographic lie near the middle.

The resulting slots also remain controllable: linear probes show no measured excess leakage beyond the label-correlation baseline, and quantized slot codes support targeted intervention, including a 15.3× lift for color. These results suggest that controllable fashion retrieval benefits from typed, attribute-conditioned multimodal slots rather than either a single global embedding or opaque item-level semantic IDs.

RecSys '26, Minneapolis, Minnesota

ACM ISBN 978-1-4503-XXXX-X/26/09
https://doi.org/10.1145/XXXXXXX.XXXXXXX

CCS Concepts

• **Information systems** → **Recommender systems**; *Content-based filtering*; • **Computing methodologies** → *Neural networks*; **Multimodal learning**.

Keywords

controllable recommendation, multimodal fashion retrieval, attribute-conditioned retrieval, semantic IDs, slot factorization, vector quantization, real-time retrieval

ACM Reference Format:
Najmeh Forouzandehmehr, Topojoy Biswas, Evren Korpeoglu, and Kannan Achan. 2026. Attribute-Conditioned Multimodal Slot Factorization for Controllable Fashion Retrieval. In *Proceedings of the 16th ACM Conference on Recommender Systems (RecSys '26), September 2026, Minneapolis, Minnesota.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/XXXXXXX.XXXXXXX

1 Introduction

Fashion retrieval systems must simultaneously satisfy constraints across multiple orthogonal attributes: a query for "women's patterned summer dresses" requires matching category, demographic, and pattern axes at once. Standard retrieval pipelines encode queries as monolithic dense vectors, making it difficult to selectively emphasize one attribute at query time without constructing an entirely new query.

Supervised slot factorization improves constraint-satisfaction retrieval by decomposing item representations into named attribute slots, each optimized for a different semantic dimension. However, slot factorization alone does not determine which input modality each slot should use. This is especially important in fashion: color and pattern are often primarily visual, while category and demographic are often more explicit in product text and catalog taxonomy. A text-only slot encoder therefore underuses visual evidence, while a single fixed text–image blend applies the same modality mixture to every attribute. We instead argue that modality fusion should be attribute-conditioned: each slot should learn how much to rely on text versus image evidence.

We address this gap with **MM-slotgate**: a multimodal slot encoder that fuses Fashion-CLIP (fCLIP) [2] text and image embeddings through per-slot learnable gates. Each slot s receives a scalar

gate $g_s = \sigma(a_s)$, trained end-to-end, that controls the text-versus-image blend entering that slot's VQ bottleneck. Rather than prescribing a fixed modality ratio, the model learns that color is best resolved by images ($g_{\text{color}} = 0.426$, image-leaning), while category and demographic are resolved by text ($g_{\text{categ}} = 0.545$, $g_{\text{demo}} = 0.509$, text-leaning)—without any modality supervision.

We make three contributions:

(1) **MM-slotgate architecture**: a per-slot learnable gate over fCLIP text and image embeddings that learns attribute appropriate modality allocation end-to-end. A graceful missing-image fallback ensures full catalog coverage.
(2) **Strong empirical results**: MM-slotgate achieves 0.7566 macro ConstraintSatisfied@10 on H&M, surpassing equal-weight MM-global fusion (0.7142, +5.9% relative) and fCLIP text-only (0.4755, +59.1%). The color constraint improves from 0.321 (fCLIP-text) to 0.889 (+0.568 absolute), the largest single-attribute gain observed.
(3) **Gate analysis and ablation**: Learned gates converge to attribute-appropriate modality preferences without modality supervision (color image-leaning, category text-leaning) and outperform fixed equal-weight fusion on macro CS@10 (0.7566 vs. 0.7516, +0.005 absolute). A shuffled-image negative control confirms that actual visual content, not merely co-training with an extra input branch, drives the gain.

2 Related Work

2.1 Semantic IDs and Vector Quantization

Semantic ID methods [10, 15, 16] represent items as structured discrete codes for retrieval and generation. TIGER [10] learns item IDs via residual quantization and generates the next item's code sequence autoregressively; subsequent work uses these IDs as ranking features [12] or enforces hierarchical structure [11]. VQ-Rec [7], LETTER [14], and TokenRec [8] learn discrete tokens for sequential and LLM-based recommendation. Recent work also studies disentangled or tag-aligned semantic IDs [3].

Our approach differs in the semantics assigned to each code. Existing methods learn item-level or residual addresses whose meaning emerges implicitly; we instead define four named attribute slots—pattern, color, category, demographic each with its own codebook. This makes the representation directly addressable: one slot can be weighted or intervened on independently, and the semantic role of each code is fixed by supervised alignment rather than discovered implicitly.

2.2 Multimodal Fashion Retrieval

CLIP [9] and its fashion-domain adaptation fCLIP [2] provide strong monolithic embeddings by aligning image and text in a shared space. These embeddings improve per-attribute retrieval precision but blend all attributes into a single vector, making selective emphasis difficult. MM-slotgate uses fCLIP as a backbone but replaces the global embedding with four attribute-specific slots, each with its own learned modality gate.

2.3 Controllable and Disentangled Retrieval

Compositional retrieval [13] edits the *query* representation at inference time using relative attribute feedback. Our approach is complementary: we decompose *index-time item* representations, enabling attribute intervention without recomputing any embedding. Unsupervised disentanglement (β-VAE [5]) does not guarantee alignment with task-relevant attributes; we use supervised slot alignment and evaluate disentanglement relative to a label correlation baseline [4].

3 MM-Slotgate: Multimodal Slot Encoder

Figure 1 gives the retrieval and intervention intuition before we define the model formally.

MM-slotgate has three stages. First, each item is embedded with Fashion-CLIP text and image encoders. Second, each semantic slot learns its own text–image mixture and produces a continuous slot vector $\mathbf{z}_{i,s}$. Third, each slot is quantized into a slot-specific codebook, yielding $\mathbf{z}^q_{i,s}$. Retrieval uses the continuous slots $\mathbf{z}$, while intervention uses the quantized slots $\mathbf{z}^q$.

Inputs:

We encode each item i with the Fashion-CLIP ViT-B/32 backbone [2], producing a 512-dim text embedding $\mathbf{t}_i$ (product name + description) and a 512-dim image embedding $\mathbf{v}_i$ (product photograph). A binary indicator $m_i \in \{0, 1\}$ flags image availability (≈95% on H&M). *Note*: "demographic" denotes product target audience (men's, women's, kids'), not a user attribute.

Per-slot projection and gating:

For each slot $s \in$ {pattern, color, category, demographic}:

$$\mathbf{h}^t_{i,s} = \text{LN}(W^t_s\, \mathbf{t}_i), \quad \mathbf{h}^v_{i,s} = \text{LN}(W^v_s\, \mathbf{v}_i), \tag{1}$$

where $W^t_s, W^v_s \in \mathbb{R}^{128\times512}$ are slot-specific projections with Layer Normalization. A per-slot scalar a_s (initialized to 0, learned end-to-end) defines the text weight $g_s = \sigma(a_s) \in (0, 1)$; thus g_s is the text weight and $1-g_s$ is the image weight for slot s. Values near 0.5 give equal fusion; g_s>0.5 is text-leaning; g_s<0.5 is image-leaning.

Fusion with missing-image fallback:

For items with an available product image, each slot mixes its text and image projections according to the learned gate; for items without an image, the slot uses the text projection only:

$$\tilde{\mathbf{z}}_{i,s} = m_i\big(g_s\, \mathbf{h}^t_{i,s} + (1-g_s)\, \mathbf{h}^v_{i,s}\big) + (1-m_i)\, \mathbf{h}^t_{i,s}, \tag{2}$$

followed by L2-normalization: $\mathbf{z}_{i,s} = \text{Norm}(\tilde{\mathbf{z}}_{i,s}) \in \mathbb{R}^{128}$. When $m_i = 0$, the slot falls back to $\text{Norm}(\mathbf{h}^t_{i,s})$, preserving full catalog coverage.

Vector quantization:

Each slot has an EMA codebook of K_s unit-norm codewords ($K_s \in \{15, 20, 10, 6\}$ for pattern, color, category, demographic). A straight-through estimator [1] passes gradients through the discrete assignment: $\mathbf{z}^q_{i,s} = \mathbf{z}_{i,s} + \text{sg}(\mathbf{e}_{i,s} - \mathbf{z}_{i,s})$. Dead entries (EMA count < 0.05) are restarted from random training items.

Training loss:

$$\mathcal{L} = \sum_{s=1}^{4}\Big[\underbrace{\beta\|\mathbf{z}_{i,s} - \text{sg}(\mathbf{e}_{i,s})\|^2}_{\text{commitment}} + \underbrace{\lambda_a\, \mathcal{L}_{\text{CE}}(\text{logits}_s,\, y_s)}_{\text{alignment}}\Big] + \lambda_\perp \mathcal{L}_{\text{orth}}, \tag{3}$$

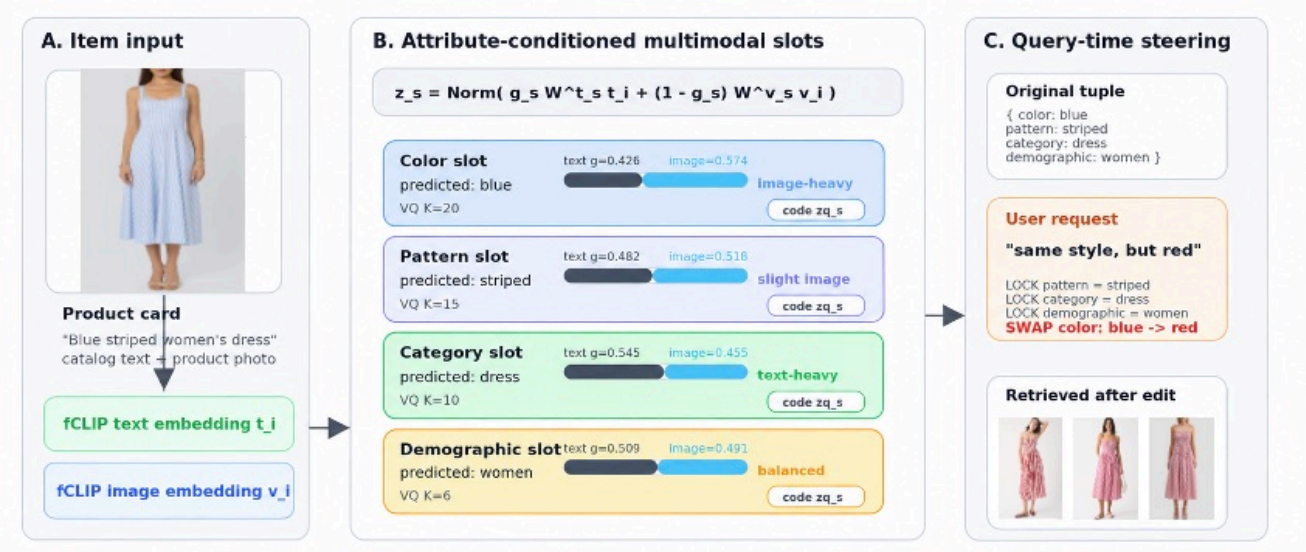

Figure 1: MM-slotgate decomposes a fashion item into four named multimodal slots. Each slot learns its own text–image gate before vector quantization. Continuous slots support weighted retrieval; quantized slot codes support targeted intervention, illustrated by changing only the color slot from blue to red.

with β=0.25, λ_a=5.0, and $\lambda_\perp$=2.0. $\mathcal{L}_{\text{orth}}$ penalizes off-diagonal elements of the slot Gram matrix, discouraging slot collapse. Optimization details are given in Section 4.

Combined retrieval:

At query time, the score for item i given constrained query q is:

$$\text{score}(q, i) = \cos\big(\mathbf{w}(q),\, \mathbf{w}(i)\big) + \alpha \sum_{s \in C} \log \hat{P}(y_s = c_s^* \mid i), \qquad (4)$$

where $\mathbf{w}(\cdot)$ is slot-weighted concatenation (constrained slots 3×, others 1×, L2-normalized; *Wsel-z*), $\hat{P}$ from alignment logit heads, and α^* tuned by 5-fold CV. Equation (4) is called the *combined* score.

4 Experimental Setup and Retrieval Evaluation

We first ask whether slot-specific multimodal fusion improves constraint satisfaction.

Dataset. We evaluate on the H&M Personalized Fashion Recommendations catalog [6], using a 50K-item subset with product text, product images, and four attribute labels: pattern, color, category, and demographic. We use a 90/10 split: 45K items for training and 5K held out for evaluation. Image availability is approximately 95%; items without images use the text-only fallback defined in Eq. (2).

Training protocol. All MM-slotgate variants are trained for 80 epochs on the 45K-item training split with AdamW (learning rate 10^{-3}, cosine decay), EMA codebook decay 0.99, and dead-code restart.

Retrieval protocol. We evaluate CS@n on the 5K held-out split at n=10. For each constraint set, we sample 500 random query items and retrieve from the held-out gallery excluding the query itself.

Metric. ConstraintSatisfied@n (CS@n): fraction of top-n retrieved items where *all* constrained slots match the query's ground-truth labels. *Chance* is $\prod_{s \in C} \sum_c p(y_s = c)^2$.

Methods.

- **CLIP-text**: combined retrieval score on CLIP ViT-B/32 text embeddings with slot encoder.
- **fCLIP-text-only**: combined retrieval on fCLIP text embeddings (no image input; text-only MM-slotgate variant with $g_s \approx 1$).
- **MM-global**: fuses text and image *before* the slot encoder, using the precomputed $\text{Norm}(0.5\,\mathbf{t}_i + 0.5\,\mathbf{v}_i)$ as input—a single global mixture with no per-slot gating. (Distinct from **fixed_half** in Table 4, which keeps separate per-slot text/image projections but freezes each gate at g_s=0.5.)
- **MM-slotgate (ours)**: combined retrieval with per-slot learned gates (Eq. (4)).
- **MetaFilter+fCLIP**† (oracle): the item catalog is filtered to items whose attribute labels exactly match the query's constrained slots, then ranked by fCLIP cosine similarity. Requires knowing the answer labels at retrieval time and is therefore unreachable in practice; represents the best result a perfect attribute filter could achieve.

Results. Table 1 shows macro CS@10. MM-slotgate (0.7566) outperforms MM-global (0.7142) by +5.9% relative. Gate ablation analysis (§5) shows that learned per-slot allocation slightly outperforms fixed equal-weight fusion while providing interpretable attribute-specific modality allocation (0.7566 vs. 0.7516). Both multimodal methods far exceed fCLIP text-only (0.4755), with gains of +58.8% (MM-slotgate) and +50.2% (MM-global) relative to the text baseline—demonstrating that image information is essential for constraint-satisfying retrieval on H&M.

Table 1: Macro ConstraintSatisfied@10 (H&M, combined retrieval). MetaFilter† is an oracle (GT labels used to filter).

Method	Macro CS@10
MetaFilter+fCLIP†	0.990
MM-slotgate (ours)	**0.7566**
MM-global fusion	0.7142
fCLIP-text-only	0.4765
CLIP-text	0.4640

† oracle upper bound; combined = Wsel-z + slot logits

Table 2: Per-constraint CS@10 (H&M, combined retrieval). Bold = best per row. MM-slotgate leads on all nine constraint sets.

Constraint	fCLIP-txt	MM-global	MM-slotgate
color	0.321	0.852	**0.889**
category	0.861	0.861	**0.889**
demographic	0.749	0.804	**0.849**
pattern	0.647	0.812	**0.824**
cat+demo	0.642	0.705	**0.757**
color+categ	0.225	0.670	**0.725**
color+demo	0.210	0.643	**0.718**
patt+categ	0.492	0.659	**0.685**
cat+demo+color	0.141	0.422	**0.474**
Macro	0.4765	0.7142	**0.7566**

Table 2 reports per-constraint CS@10 for all three methods across the nine constraint sets. MM-slotgate leads on every row, with the largest gains on color-containing constraints. The strongest single-attribute gain is on color: fCLIP-text combined retrieval yields 0.321, while MM-slotgate reaches 0.889 (+0.568 absolute). This aligns with the learned color gate ($g_{\text{color}} = 0.426$, image weight 57.4%): product images provide a more direct signal for color class than product descriptions, where lexical variants such as "navy," "midnight," and "ink" can map to the same label class.

5 Gate Analysis and Ablation

We next examine whether the learned gates match intuitive modality needs for each attribute.

Learned gate values. Table 3 shows the per-slot gate values $g_s = \sigma(a_s)$ after 80 training epochs. Recall that g_s weights the text projection: $g_s > 0.5$ is text-leaning, $g_s < 0.5$ is image-leaning (image weight $= 1 - g_s$).

The color gate diverges furthest from 0.5 toward the image modality, consistent with its dominant per-constraint gain (0.321 → 0.889). Category and demographic lean text, consistent with these slots being determined by taxonomic labeling conventions rather than visual appearance. Pattern falls between the two: pattern names ("striped", "floral") have reasonable text coverage but benefit from visual confirmation. Critically, these preferences emerge without any modality supervision—the gate is trained solely to minimize alignment CE loss and commitment loss on multi-attribute labels.

Table 3: Learned gate values g_s at convergence. Text weight $= g_s$; image weight $= 1 - g_s$. All gates initialized to $g_s = 0.5$ ($a_s = 0$).

Slot	g_s	Image wt.	Modality preference
color	0.426	0.574	image-leaning
pattern	0.482	0.518	balanced
demographic	0.509	0.491	text-leaning
category	0.545	0.455	text-leaning

Table 4: Gate ablation: macro CS@10 and color slot accuracy. Learned gates give the best CS@10, while fixed-half is close; the main benefit of learned gates is interpretable slot-specific modality allocation.

Mode	Macro CS@10	Color acc	Val loss
learned	**0.7566**	0.762	28.91
fixed_half	0.7516	0.763	28.77
image_only	0.6618	**0.786**	29.82
text_only	0.4765	0.304	36.33

Gate ablations. Table 4 compares four gate-mode variants on validation loss and per-slot classification accuracy (80 epochs).

Learned gates produce the best macro CS@10, but the margin over fixed_half is small (0.7566 vs. 0.7516). We therefore interpret the gate primarily as an interpretable and controllable mechanism for attribute-specific modality allocation, rather than as the sole source of the retrieval gain. The large drops for text_only (0.4765) and image_only (0.6618) show that both modalities are necessary: text_only collapses on color (acc 0.304, near chance), while image_only degrades on category and demographic where text provides the primary signal.

Negative control. We test whether the image branch helps because it contains the correct product image, rather than simply because the model has an extra input. To do this, we train the same model after randomly shuffling image embeddings across items, so each product is paired with another product's image. Under this shuffle, the model learns to rely less on images: all gates become text-leaning ($g \approx 0.57$–0.63), and validation loss becomes close to the text-only model (36.95 vs. 36.33). This suggests that MM-slotgate benefits from correctly aligned visual content, not merely from adding an extra image branch.

Representation diagnostics. We also run a simple diagnostic to check where color information is stored before slot factorization. We train logistic classifiers on frozen fCLIP text, image, and global-fusion embeddings, independently of MM-slotgate. These classifiers are not retrieval baselines; they only measure how much attribute information is present in each input representation. The image and global multimodal embeddings achieve high color accuracy (0.802 and 0.789), showing that color is strongly available in the visual branch. This matches the learned MM-slotgate behavior: the color slot assigns more weight to the image stream.

Table 5: Linear probe AUC matrix for MM-slotgate. Rows = probe target, cols = feature slot. Diagonal mean = 0.897; off-diagonal mean = 0.671; excess leakage = −0.019, indicating no measured excess leakage beyond the label-correlation baseline.

	patt	color	categ	demo
pattern	**0.866**	0.652	0.746	0.645
color	0.607	**0.912**	0.601	0.603
category	0.720	0.623	**0.912**	0.747
demographic	0.678	0.660	0.772	**0.899**

Table 6: Linear probe AUC matrix for text-only CLIP slot encoder (*prior model, shown for reference*). Rows = probe target, cols = feature slot. Excess leakage = 0.704−0.690 = 0.014.

	patt	color	categ	demo
pattern	**0.790**	0.709	0.734	0.652
color	0.606	**0.641**	0.587	0.613
category	0.768	0.810	**0.904**	0.753
demographic	0.721	0.739	0.753	**0.858**

6 Measuring Slot Disentanglement

High retrieval accuracy alone is not sufficient for controllability; the slots must also avoid excess cross-slot leakage.

Multimodal fusion can, in principle, increase cross-slot information leakage if visual features are correlated across attribute axes. We measure this with the linear probe AUC matrix protocol from prior work.

Linear probe AUC matrix (MM-slotgate). We fit logistic regression probes from each slot's $\mathbf{z}_s$ to each slot's ground-truth label on a held-out 10% split. The 4×4 AUC matrix (rows = probe target, cols = feature slot) distinguishes own-slot encoding (diagonal) from cross-slot leakage (off-diagonal).

Table 5 shows the MM-slotgate AUC matrix. Diagonal mean AUC is 0.897, substantially higher than the prior text-only slot encoder (0.798). The largest gains are on color (0.641 → 0.912) and pattern (0.790 → 0.866)—the two visually grounded slots.

Although the raw off-diagonal AUC is 0.671, this does not by itself imply model-induced leakage because H&M labels are correlated. The label-prior probe, which uses only dataset co-occurrence, has off-diagonal AUC 0.690. MM-slotgate is slightly *below* this baseline, giving excess leakage of −0.019. Thus, under linear probes, the model does not add measurable cross-slot leakage beyond what is already present in the label structure.

Reference: text-only CLIP baseline. The prior text-only slot encoder (CLIP backbone) achieved diagonal mean AUC = 0.798, off-diagonal mean = 0.704, and excess leakage = 0.014— also near zero after controlling for dataset co-occurrence. Table 6 shows this matrix for comparison.

7 Targeted Attribute Intervention

Finally, we test whether the quantized slot codes can be directly edited at query time.

Table 7: Codebook intervention results at n=10 (MM-slotgate). PreserveDelta = mean change in other-slot match rate.

Slot	Hit@10	Null@10	Lift	ΔPres
color	0.297	0.019	**15.3**×	−0.026
category	0.222	0.021	10.7×	−0.042
pattern	0.099	0.011	9.4×	−0.021
demographic	0.370	0.044	8.5×	−0.080

CLIP-text reference: color 1.5×, category 13.6×, pattern 3.0×, demo 4.5×.

The VQ codebooks in MM-slotgate support direct attribute steering without re-embedding. For each query, we choose one slot and replace its quantized code with the codebook entry most associated with a different target class, using a mapping computed from training items only. We then retrieve with the modified representation and measure whether the top-10 results shift toward the target class while preserving other attributes.

We report *Lift* = HitRate@10 / NullHitRate@10 (prevalence-adjusted target-class retrieval rate) and *PreserveDelta@10* (mean change in other-slot match rate; negative values indicate collateral disruption).

MM-slotgate intervention results. Table 7 shows results at n=10. Color achieves 15.3× lift (Hit = 0.297 vs. Null = 0.019), a 10.2× improvement over the CLIP-text color lift (1.5×), consistent with the much stronger color slot encoding (AUC 0.641 → 0.912). Pattern (9.4×), category (10.7×), and demographic (8.5×) all show strong lifts, with PreserveDelta near zero on non-intervened slots.

The largest relative change is on color: lift increases from 1.5× in the prior text-only slot encoder to 15.3× under MM-slotgate. This matches the stronger color slot AUC (0.641 → 0.912) and shows that visual grounding improves both retrieval and codebook-level steering.

8 Limitations

The current taxonomy contains only four slots; extending the method to richer fashion concepts such as occasion, style aesthetic, and brand remains future work.

9 Conclusion

We introduced MM-slotgate, a multimodal slot encoder for controllable fashion retrieval. Instead of representing each item with a single monolithic embedding, MM-slotgate factorizes FashionCLIP text and image embeddings into four named attribute slots and lets each slot learn its own text–image gate. This allows visually grounded attributes such as color and pattern to draw more from image evidence, while taxonomy-oriented attributes such as category and demographic can remain more text-driven.

On H&M, MM-slotgate achieves 0.7566 macro ConstraintSatisfied@10 using the combined retrieval score, outperforming fCLIP text-only retrieval (0.4755) and equal-weight MM-global fusion (0.7142). The largest gain is on color, where CS@10 improves from 0.321 to 0.889. This improvement aligns with the learned color gate, which assigns 57.4% weight to image evidence (g_{color} = 0.426). More

broadly, the learned gates are interpretable without modality supervision: color leans image, category leans text, and pattern and demographic lie near the middle.

The same slot structure also preserves controllability. The MM-slotgate slots show stronger own-slot predictive accuracy than the prior text-only slot encoder, with diagonal AUC improving from 0.798 to 0.897, while adding no measured excess leakage beyond the label-correlation baseline. Quantized slot codes further support direct codebook-level intervention: color achieves 15.3× hit-rate lift over unmodified retrieval, compared with 1.5× for the prior text-only slot encoder.

Overall, the results suggest that controllable fashion retrieval benefits from typed, attribute-conditioned multimodal representations. A single global text–image blend is not sufficient: different fashion attributes require different modality mixtures. MM-slotgate provides a simple mechanism for learning those mixtures while keeping the representation explicitly addressable for retrieval weighting and targeted intervention.

References

[1] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. 2013. Estimating or Propagating Gradients Through Stochastic Neurons for Conditional Computation. In *arXiv preprint arXiv:1308.3432*.

[2] Po-Yao Chia, Giuseppe Attanasio, Federico Bianchi, Silvia Terragni, Ana Rita Pires, Gianmarco Liu, and Ciro Legrand. 2022. Contrastive Language-Image Pre-Training for Fashion. In *Proceedings of the ACM Web Conference*. doi:10.1145/3485447.3512241

[3] Dengzhao Fang, Jingtong Gao, Chengcheng Zhu, Yu Li, Xiangyu Zhao, and Yi Chang. 2025. HiD-VAE: Interpretable Generative Recommendation via Hierarchical and Disentangled Semantic IDs. arXiv:2508.04618 [cs.IR] https://arxiv.org/abs/2508.04618

[4] Xintong Han, Zuxuan Wu, Phoenix X Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S Davis. 2017. Automatic Spatially-aware Fashion Concept Discovery. In *Proceedings of the IEEE International Conference on Computer Vision*.

[5] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. *International Conference on Learning Representations* (2017).

[6] H&M Group. 2022. H&M Personalized Fashion Recommendations. Kaggle competition dataset. https://www.kaggle.com/competitions/h-and-m-personalized-fashion-recommendations

[7] Yupeng Hou, Zhankui He, Julian McAuley, and Wayne Xin Zhao. 2023. Learning Vector-Quantized Item Representation for Transferable Sequential Recommenders. In *Proceedings of the ACM Web Conference 2023*. 1162–1171. arXiv:2210.12316 [cs.IR] doi:10.1145/3543507.3583434

[8] Haohao Qu, Wenqi Fan, Zihuai Zhao, and Qing Li. 2025. TokenRec: Learning to Tokenize ID for LLM-Based Generative Recommendations. *IEEE Transactions on Knowledge and Data Engineering* 37, 10 (2025), 6216–6231. arXiv:2406.10450 [cs.IR] doi:10.1109/TKDE.2025.3599265

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*.

[10] Shashank Rajput, Nikhil Mehta, Anima Singh, Raghunandan Keshavan, Trung Vu, Lukasz Heldt, Lichan Hong, Yi Tay, Vinh Q Tran, Jonah Samost, et al. 2023. Recommender Systems with Generative Retrieval. In *Advances in Neural Information Processing Systems*.

[11] Zihua Si, Zhongxiang Sun, Jiale Chen, Guozhang Chen, Xiaoxue Zang, Kai Zheng, Yang Song, Xiao Zhang, Jun Xu, and Kun Gai. 2024. Generative Retrieval with Semantic Tree-Structured Identifiers and Contrastive Learning. In *Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*. 154–163. arXiv:2309.13375 [cs.IR] https://arxiv.org/abs/2309.13375

[12] Anima Singh, Trung Vu, Nikhil Mehta, Raghunandan Keshavan, Maheswaran Sathiamoorthy, Yilin Zheng, Lichan Hong, Lukasz Heldt, Li Wei, Devansh Tandon, Ed H. Chi, and Xinyang Yi. 2024. Better Generalization with Semantic IDs: A Case Study in Ranking for Recommendations. In *Proceedings of the 18th ACM Conference on Recommender Systems*. 1039–1044. arXiv:2306.08121 [cs.IR] doi:10.1145/3640457.3688190

[13] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. 2019. Composing Text and Image for Image Retrieval – an Empirical Odyssey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[14] Wenjie Wang, Honghui Bao, Xinyu Lin, Jizhi Zhang, Yongqi Li, Fuli Feng, See-Kiong Ng, and Tat-Seng Chua. 2024. Learnable Item Tokenization for Generative Recommendation. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. 2400–2409. arXiv:2405.07314 [cs.IR] doi:10.1145/3627673.3679569

[15] Han Zhu, Daqing Chang, Ziru Xu, Pengye Zhang, Xiang Li, Jie He, Han Li, Jian Xu, and Kun Gai. 2019. Joint Optimization of Tree-Based Index and Deep Model for Recommender Systems. In *Advances in Neural Information Processing Systems*, Vol. 32. arXiv:1902.07565 [cs.IR] https://arxiv.org/abs/1902.07565

[16] Han Zhu, Xiang Li, Pengye Zhang, Guozheng Li, Jie He, Han Li, and Kun Gai. 2018. Learning Tree-Based Deep Model for Recommender Systems. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 1079–1088. arXiv:1801.02294 [cs.IR] doi:10.1145/3219819.3219826