

VIRTUAL FASHION PHOTO-SHOOTS: BUILDING A LARGE-SCALE GARMENT-LOOKBOOK DATASET

Yannick Hauri Luca A. Lanzendörfer Till Aczel

ETH Zurich

ABSTRACT

Fashion image generation has so far focused on narrow tasks such as virtual try-on, where garments appear in clean studio environments. In contrast, editorial fashion presents garments through dynamic poses, diverse locations, and carefully crafted visual narratives. We introduce the task of virtual fashion photo-shoot, which seeks to capture this richness by transforming standardized garment images into contextually grounded editorial imagery. To enable this new direction, we construct the first large-scale dataset of garment–lookbook pairs, bridging the gap between e-commerce and fashion media. Because such pairs are not readily available, we design an automated retrieval pipeline that aligns garments across domains, combining visual–language reasoning with object-level localization. We construct a dataset¹ with three garment–lookbook pair accuracy levels: high quality (10,000 pairs), medium quality (50,000 pairs), and low quality (300,000 pairs). This dataset offers a foundation for models that move beyond catalog-style generation and toward fashion imagery that reflects creativity, atmosphere, and storytelling.

Index Terms— Virtual Photo-Shoot, Dataset Curation, Fashion Image Generation, Garment-Lookbook Pairs, Image Retrieval

1. INTRODUCTION AND RELATED WORK

Advances in image generation and the growth of the fashion industry have driven research in virtual try-on, fashion image editing, clothing recognition, and garment classification. Virtual try-on, in particular, allows users to upload an image and generate how different garments would appear when worn.

Building on virtual try-on, we introduce the task of *virtual photo-shoot*, which aims to generate editorial-style images of models wearing a given garment in diverse, complementary settings. This enables designers and fashion houses to automatically produce creative photo-shoot material, moving beyond the studio-like outputs of existing try-on systems.

Training such models requires data linking garment-level product images with lookbook-style photography. Existing datasets [1–5] provide rich annotations but focus on shop environments. Figure 1 illustrates the difference between shop and lookbook style images. These datasets pair isolated garment images (with uniform white backgrounds) with shop-lookbook images, which exhibit minimal variation in poses, backgrounds, and styling. Consequently, current datasets do not capture the creative diversity of real fashion media.

To address this gap, we construct the first dataset of garment–lookbook pairs. In this setting, the garment image provides a standardized product-level reference, while the lookbook image captures the same garment in diverse poses, backgrounds, and artistic styles. By linking these two domains, the dataset enables training



Fig. 1. Difference in garment, shop, and lookbook image. Existing datasets provide clean shop images, not suitable for virtual photo-shoot model training.

models that can generate lookbook-style photographs conditioned on a garment image, analogous to how virtual try-on models generate studio-style outputs from garment–shop pairs. Unlike try-on datasets, which can be collected directly from e-commerce product pages, garment–lookbook pairs are not co-located and must be assembled from separate sources. We therefore gather unpaired garment and lookbook images from diverse collections and create pairs automatically through garment retrieval.

Several existing retrieval approaches have been applied to fashion matching. Proxynca++ [6] leverages proxy-based contrastive learning to capture fine-grained visual similarity, enabling accurate image-to-image matching in structured datasets. Hyp-DINO [7] encodes hierarchical embeddings in hyperbolic space, capturing relationships between visually similar garments more effectively than standard Euclidean embeddings. While these models perform well on clean, curated datasets, they are less robust to the diverse backgrounds, poses, and editing styles found in editorial fashion imagery.

To overcome these limitations, we develop a retrieval pipeline combining vision–language models (VLMs), object detection (OD), and SigLIP-based similarity estimation [8]. VLMs identify garment categories in natural language, OD isolates relevant regions in lookbook images, and SigLIP provides robust similarity scores between garment crops and query images. This combination is particularly effective in noisy, heterogeneous settings, where existing metric-learning models like Proxynca++ and Hyp-DINO alone struggle.

By integrating these complementary approaches into an ensemble, we improve retrieval accuracy and robustness across diverse datasets containing complex poses, backgrounds, and editorial styles. This high-quality garment–lookbook matching enables the construction of a dataset suitable for training generative models to

¹<https://huggingface.co/datasets/disco-eth/lookbook>

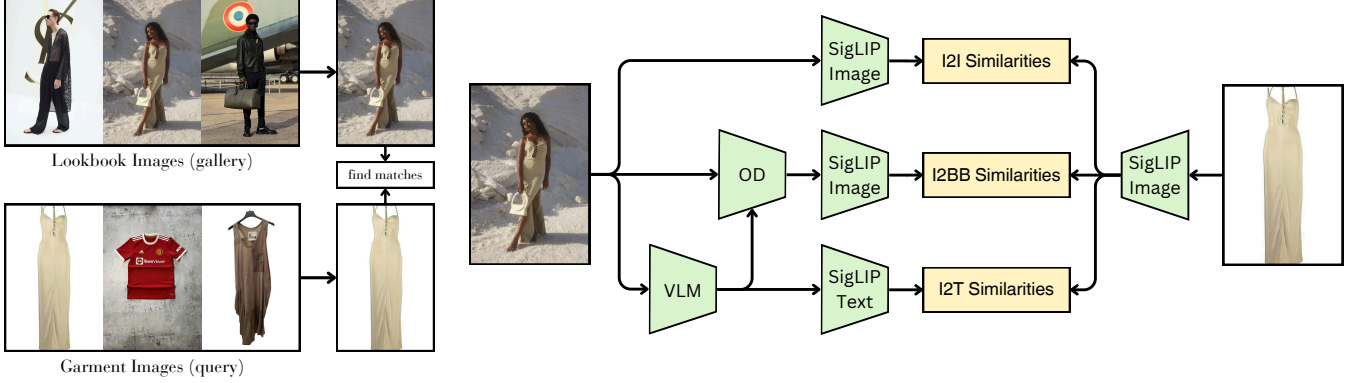


Fig. 2. Left: Examples of lookbook images (gallery) and garment images (query), showing a sample where one garment appears in both a gallery and a query image. Matching gallery-query pairs form the basis of our dataset, and the task is to find all such matches. **Right:** Overview of our retrieval pipeline. Query images are embedded with SigLIP2, while garment descriptions for gallery images are generated with a vision-language model (VLM). Object detection (OD) pconditioned on the garment description reduces bounding boxes for individual garments. Embeddings of gallery images, descriptions, and bounding boxes are compared with SigLIP2 to compute image-to-image, image-to-bbox, and image-to-text similarities.

produce rich, contextually grounded virtual photo-shoots.

In summary, this work makes three contributions: (1) We define the new task of virtual photo-shoot and present the first large-scale dataset of garment–lookbook pairs. (2) We propose a zero-shot retrieval pipeline integrating VLMs, OD, and SigLIP for automatic garment–lookbook matching. (3) We provide an ensemble-based retrieval strategy that further improves the quality of paired data.

2. METHODOLOGY

To support the task of virtual photo-shoot, we construct a large-scale dataset of garment–lookbook pairs linking product-level garment images with lookbook images. The dataset is built in two stages: first, we collect unpaired garment and lookbook images; then, we create pairs via retrieval. This approach reflects the reality that product pages and editorial media are rarely co-located for most brands. A visualization of the retrieval stage is shown in Figure 2.

2.1. SigLIP2 Retrieval

Garment–lookbook retrieval is challenging because lookbook images contain multiple garments, diverse poses, and complex backgrounds, while product-level garment images are clean and standardized. Directly embedding the garment and full lookbook image produces suboptimal matches. We refer to this baseline as **SigLIP2-FI2I** (Full Image-to-Image).

To improve retrieval, we introduce **SigLIP2-T2I** (Text-to-Image). A vision–language model (gpt-4.1-mini [9]) parses each lookbook image into individual garment components and generates concise natural language descriptions. We then compute similarity between the product garment’s SigLIP2 image embedding and the SigLIP2 text embedding of each description, improving robustness by filtering out background clutter and focusing on garment content.

Relying only on text, however, discards fine visual cues such as patterns, texture, and subtle design elements. To recover these, we propose **SigLIP2-BB2I** (Bounding Box-to-Image), where an open-vocabulary object detector (YOLO-World [10]), guided by the text descriptions, predicts bounding boxes for each garment in the lookbook. We crop these regions, embed them with SigLIP2, and com-

pare them to the product garment embedding, yielding localized image-to-image similarities.

Since SigLIP2-BB2I produces multiple scores per lookbook image, we aggregate them with the full-image similarity from SigLIP2-FI2I by taking the maximum. This final strategy, **SigLIP2-I2I** (Image-to-Image), leverages both global and localized cues while avoiding dilution by weaker matches.

2.2. Ensemble Retrieval

While the SigLIP2 pipeline provides strong zero-shot performance, further gains can be achieved by incorporating complementary retrieval models. We therefore extend our approach into an ensemble retrieval system that combines SigLIP2 with specialized metric learning methods. Specifically, we include two state-of-the-art distance metric learning models: Proxynca++ [6] and Hyp-DINO [7]. These models capture garment-specific details and structural cues beyond what SigLIP2 alone provides, producing similarity scores that complement the SigLIP2-based similarities.

Because the similarity distributions of different models are not directly comparable, we normalize them before combining. For each model m , we estimate the mean μ_m and standard deviation σ_m of its similarity scores. We then transform each score s_{mij} for query–gallery pair (i, j) into a standardized score

$$s'_{mij} = \frac{s_{mij} - \mu_m}{\sigma_m}, \quad (1)$$

so that all models operate on a common standardized scale.

With this normalization, we combine similarities from multiple models. Merging SigLIP2-I2I and SigLIP2-T2I yields the **SigLIP2-Ensemble**, while combining all four models (SigLIP2-I2I, SigLIP2-T2I, Proxynca++, Hyp-DINO) produces the **Total-Ensemble**.

2.3. Dataset

We collect approximately 550,000 lookbook and runway images with associated metadata from SHOWstudio² and Tagwalk³. We

²<https://www.showstudio.com/>

³<https://www.tag-walk.com/>

augment these collections with roughly 9.5 million garment images from e-commerce platforms such as Farfetch⁴, VestiaireCollective⁵, Grailed⁶, and Depop⁷, using brand names from the editorial metadata as queries. Metadata such as brand name and short descriptions are retained to assist search and filtering. The resulting corpus contains about ten million images and, to our knowledge, represents the first large-scale resource tailored to the virtual photo-shoot task. We include runway images under the lookbook category as a practical compromise, since runway photography contributes editorial diversity that is rarely available at scale from single-brand lookbooks.

Pairing images follows the methodology described in Sections 2.2. For each query garment, we compare its brand name with those of gallery candidates using fuzzy string matching (Rapid-Fuzz [11]). Only gallery images with sufficiently similar brand names are retained. Among these candidates, we select the lookbook image with the highest ensemble similarity score to form a garment–lookbook pair. Sorting all pairs by similarity produces a curated dataset aligned with both visual and semantic consistency.

To create quality splits, we rank each garment image by its highest similarity score. The top 10,000 pairs form the *high-quality* set, the top 50,000 pairs form the *medium-quality* set, and the top 300,000 pairs form the *low-quality* set.

3. EXPERIMENTAL SETUP

To evaluate retrieval strategies for pairing images in our dataset, we require a benchmark dataset that provides ground truth garment–lookbook pairs. We consider four established datasets: DeepFashion In-Shop [1], DeepFashion Consumer-to-Shop [1], DeepFashion2 [3], and DressCode [5]. Qualitative inspection of our collected data indicates that DressCode is the closest in style and content, with DeepFashion2 as the next most similar.

Nevertheless, our dataset is over two orders of magnitude larger and far more variable and noisy. Lookbook images span diverse backgrounds, poses, and editing effects, while garment presentation varies in quality. Although DressCode is simpler than our dataset, it remains the best available proxy for evaluating retrieval performance. We report results on DressCode, noting these scores likely overestimate performance on our noisier, more diverse data. To train supervised models (Proxynca++ and Hyp-DINO), we use the remaining datasets (DeepFashion In-Shop, DeepFashion Consumer-to-Shop, DeepFashion2).

The SigLIP2-based retrieval models (SigLIP2-I2I, SigLIP2-T2I, and SigLIP2-BB2I) operate in a zero-shot fashion and require no training. In contrast, the two metric learning based approaches: Proxynca++ [6] and Hyp-DINO [7], require training. Proxynca++ is trained for 80 epochs with 5 warm-up epochs, while Hyp-DINO is trained for 400 epochs. The metric standardization uses means and standard deviations estimated on a random subsample of our raw dataset. For evaluation, we compute query–gallery similarities using the FAISS [12] library.

4. RESULTS

4.1. Retrieval Model

The pairwise correlations between our retrieval models (Figure 3) show that the models are only weakly correlated. Values range from

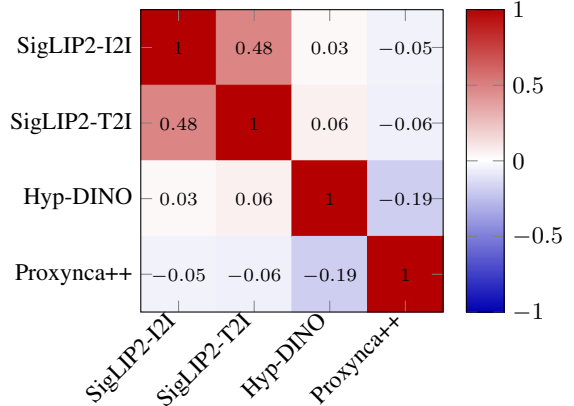


Fig. 3. Rank correlation heatmap between retrieval models. Values are rounded to two decimals and centered in each cell.

slightly negative to moderately positive, indicating that each model captures complementary aspects of garment similarity. This observation motivates our ensemble approach, as combining multiple diverse models allows us to leverage their individual strengths and improve overall retrieval performance.

Table 1 compares Proxynca++, Hyp-DINO, and the SigLIP2-Ensemble. Proxynca++ and Hyp-DINO are trained on the combined training split, while all models are evaluated on the validation sets. Despite never seeing these datasets during training, the SigLIP2-Ensemble performs competitively on clean benchmarks and even surpasses metric learning approaches on DeepFashion2. This advantage likely arises from DeepFashion2 containing more contextual noise around garments, where the VLM+OD pipeline of SigLIP2 effectively isolates relevant features. To the best of our knowledge, we achieved state-of-the-art on the DeepFashion2 dataset.

Table 2 reports Recall@K on DressCode for Proxynca++, Hyp-DINO, SigLIP2-I2I, SigLIP2-T2I, SigLIP2-Ensemble, and our Total-Ensemble. Among individual models, Hyp-DINO performs best, outperforming both Proxynca++ and SigLIP2-I2I. Our Total-Ensemble, which combines SigLIP2-I2I, SigLIP2-T2I, Proxynca++, and Hyp-DINO, achieves 89.3% R@1, 96.6% R@5, and 98.0% R@10. This is nearly 12 points higher at R@1 than the strongest single model (Hyp-DINO), highlighting the complementary strengths of SigLIP2-based and metric learning approaches.

These results demonstrate that while SigLIP2-based retrieval is highly effective in a zero-shot setting. Further gains are possible by combining it with specialized metric-learning models. The ensemble strategy ensures robustness, mitigates outliers, and integrates diverse similarity signals, which is especially important for noisy, heterogeneous datasets such as our garment–lookbook collection.

4.2. Dataset

Building on the strong performance of our Total-Ensemble retrieval model, we use it to construct garment–lookbook pairs from our image corpus. With over 550,000 gallery images, retrieving matches for every query is computationally infeasible. To address this, we limit retrieval to the top 2,000 most similar gallery images per query. Because the four retrieval models often rank different lookbook images in their top matches, computing a simple mean similarity is not possible. Instead the mean, we adopt the second-highest similarity score across the ensemble as a robust indicator.

⁴<https://www.farfetch.com>

⁵<https://us.vestiairecollective.com/>

⁶<https://www.grailed.com/>

⁷<https://www.depop.com/>

	DeepFashion Shop			DeepFashion Consumer			DeepFashion2		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Proxynca++	90.0%	96.9%	98.0%	33.2%	51.1%	56.7%	45.2%	63.8%	71.1%
Hyp-DINO	90.0%	97.0%	98.1%	42.5%	61.1%	65.9%	49.9%	68.5%	74.7%
SigLIP2-Ensemble (Ours)	83.6%	95.5%	97.4%	23.1%	40.5%	48.0%	53.6%	71.8%	78.8%

Table 1. Retrieval performance (%) across three fashion benchmarks. Proxynca++ and Hyp-DINO were trained on gallery-query pairs from the DeepFashion in-shop, DeepFashion consumer-to-shop, DeepFashion2, and DressCode datasets. Note that the SigLIP2-Ensemble was not trained on any of these datasets. The highest recall is marked in bold. Even with training on the datasets, the SigLIP2-Ensemble performs on par, or outperforms the baselines.

	R@1	R@5	R@10
SigLIP2-FI2I (Ours)	67.7%	80.8%	84.9%
SigLIP2-T2I (Ours)	63.6%	81.4%	86.2%
SigLIP2-I2I (Ours)	80.6%	91.6%	94.1%
Proxynca++	72.3%	87.4%	91.5%
Hyp-DINO	77.6%	90.2%	93.1%
Total-Ensemble (Ours)	89.3%	96.6%	98.0%

Table 2. Retrieval performance (%) for DressCode benchmark. The highest recall is marked in bold and the second best is marked with underline. Our model outperforms all previous models by over 10 percentage points.

To assemble the dataset, we rank all garment images by their highest similarity match and select the top-K pairs for each quality tier. We determine cutoffs based on Figure 4, where we manually annotate 200 sampled pairs at indices 100, 2,000, 8,000, 32,000, 128,000, 512,000, and 2,048,000. A pair is considered a true match only if the garments are visually indistinguishable, ensuring high fidelity for generative modeling. Based on these observations, we divide the dataset into three quality tiers to support different experimental needs: **High Quality:** 10,000 pairs suitable for precise evaluation or fine-tuning. **Medium Quality:** 50,000 pairs, providing a larger set for training with moderate noise. **Low Quality:** 300,000 pairs, enabling large-scale pretraining.

The dataset is designed for training diffusion models that generate lookbook images from garment inputs. High-quality pairs provide clean correspondences for fine-tuning, while medium and low-quality pairs, though noisier, increase diversity and scale to capture variations in poses, backgrounds, and editorial styles. By balancing fidelity and scale, the dataset supports robust training, enabling models to produce realistic, contextually grounded virtual photo-shoots from standardized garment images.

5. DISCUSSION AND CONCLUSION

We introduced the task of *virtual photo-shoot*, aiming to generate editorial-style fashion imagery that goes beyond standard virtual try-on systems. To support this, we constructed the first large-scale dataset of garment-lookbook pairs, bridging standardized product images with diverse, creative fashion visuals. Unlike existing virtual try-on datasets, which primarily contain clean e-commerce product and shop-lookbook images, our dataset includes a variety of backgrounds, poses, editing styles, and creative compositions, enabling research into more context-rich fashion image generation.

Because such pairs are not naturally available, we developed

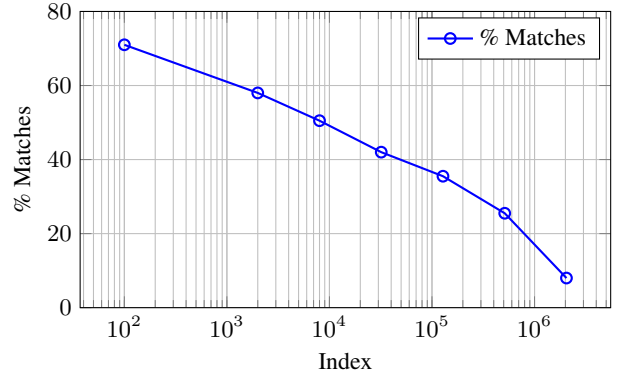


Fig. 4. Shows the garment retrieval accuracy of our dataset at indices 100, 2000, 8000, 32000, 128000, 512000, and 2048000, obtained with qualitative evaluation of 200 garment-lookbook image pair samples at each index, where the dataset is sorted by the similarity scores between the garment and lookbook image pairs.

an automated zero-shot retrieval pipeline combining SigLIP2-based similarity estimation (image-to-image and image-to-text), object-level reasoning, and vision-language alignment. An ensemble of SigLIP2, Proxynca++, and Hyp-DINO further improved robustness and coverage across noisy, heterogeneous data, substantially outperforming individual models in recall@K across DeepFashion in-shop and consumer-to-shop, DeepFashion2, and DressCode benchmarks.

The dataset is organized hierarchically: high-quality pairs provide precise correspondences for fine-tuning generative models, while medium and low-quality pairs increase scale and diversity, crucial for training diffusion models to capture variations in poses, backgrounds, and styles. This structure balances fidelity with coverage, supporting controlled, large-scale training and realistic, contextually grounded virtual photo-shoot generation.

Overall, our contributions highlight the potential of combining vision-language models, object detection, and metric learning for robust garment retrieval and dataset construction. Looking ahead, future work could extend the dataset beyond luxury fashion to include smaller brands, refine retrieval through fine-grained attribute supervision, and leverage the dataset for generative modeling tasks such as controllable virtual photo-shoot synthesis. By bridging the gap between e-commerce product imagery and creative fashion photography, we hope this work inspires new research at the intersection of computer vision, fashion, and generative modeling.

6. REFERENCES

- [1] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang, “DeepFashion: Powering robust clothes recognition and retrieval with rich annotations,” in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 1096–1104, IEEE.
- [2] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S. Davis, “VITON: An image-based virtual try-on network,” in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7543–7552, ISSN: 2575-7075.
- [3] Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, and Ping Luo, “DeepFashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images,” in *Computer Vision – ECCV 2020*, 2019, pp. 5337–5345.
- [4] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo, “VITON-HD: High-resolution virtual try-on via misalignment-aware normalization,” in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2021, pp. 14126–14135, IEEE.
- [5] Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara, “Dress code: High-resolution multi-category virtual try-on,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022, pp. 2230–2234, ISSN: 2160-7516.
- [6] Eu Wern Teh, Terrance DeVries, and Graham W. Taylor, “ProxyNCA++: Revisiting and revitalizing proxy neighborhood component analysis,” in *Computer Vision – ECCV 2020*, Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, Eds., vol. 12369, pp. 448–464. Springer International Publishing, 2020, Series Title: Lecture Notes in Computer Science.
- [7] Aleksandr Ermolov, Leyla Mirvakhabova, Valentin Khrulkov, Nicu Sebe, and Ivan Oseledets, “Hyperbolic vision transformers: Combining improvements in metric learning,” .
- [8] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai, “SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features,” .
- [9] Aram Bahrini, Mohammadsadra Khamoshifar, Hossein Abbasimehr, Robert J. Riggs, Maryam Esmaeili, Rastin Mastali Majdabadkohne, and Morteza Pasehvar, “ChatGPT: Applications, opportunities, and threats,” .
- [10] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev, “Reproducible scaling laws for contrastive language-image learning,” in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 2818–2829, IEEE.
- [11] “Rapidfuzz,” 2025, Accessed: 2025-09-16.
- [12] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou, “The faiss library,” .