
LRVS-Fashion: Extending Visual Search with Referring Instructions


Simon Lepage^{1,2} Jérémie Mary¹ David Picard²

¹ CRITEO AI Lab, Paris, France

² LIGM, École des Ponts, Marne-la-Vallée, France

{s.lepage, j.mary}@criteo.com david.picard@enpc.fr

Abstract

1 This paper introduces a new challenge for image similarity search in the context of
2 fashion, addressing the inherent ambiguity in this domain stemming from complex
3 images. We present Referred Visual Search (RVS), a task allowing users to define
4 more precisely the desired similarity, following recent interest in the industry. We
5 release a new large public dataset, LRVS-Fashion, consisting of 272k fashion
6 products with 842k images extracted from fashion catalogs, designed explicitly
7 for this task. However, unlike traditional visual search methods in the industry,
8 we demonstrate that superior performance can be achieved by bypassing explicit
9 object detection and adopting weakly-supervised conditional contrastive learning
10 on image tuples. Our method is lightweight and demonstrates robustness, reaching
11 Recall at one superior to strong detection-based baselines against 2M distractors. 

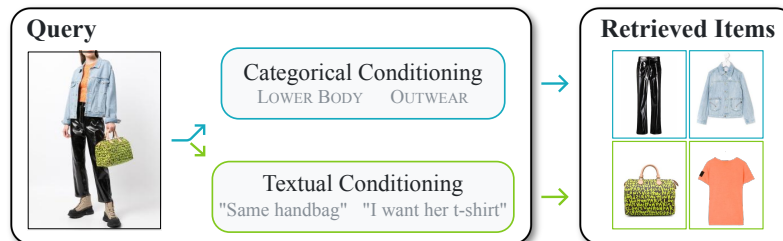


Figure 1: Overview of the Referred Visual Search task. Given a query image and conditioning information, the goal is to retrieve a target instance from a large gallery. *Note that a query is made of an image and an additional text or category, precisizing what aspect of the image is relevant.*

12 1 Introduction

13 Image embeddings generated by deep neural networks play a crucial role in a wide range of computer
14 vision tasks. Image retrieval has gained substantial prominence, leading to the development of
15 dedicated vector database systems [22]. These systems facilitate efficient retrieval by comparing
16 embedding values and identifying the most similar images within the database.

17 Image similarity search in the context of fashion presents a unique challenge due to the inherently
18 ill-founded nature of the problem. The primary issue arises from the fact that two images can be
19 considered similar in various ways, leading to ambiguity in defining a single similarity metric. For

¹The dataset is available at <https://huggingface.co/datasets/Slep/LAION-RVS-Fashion>

20 instance, two images of clothing items may be deemed similar based on their color, pattern, style, or
21 even the model pictured. This multifaceted nature of similarity in fashion images complicates the
22 task of developing a universally applicable similarity search algorithm, as it must account for the
23 various ways in which images can be related.

24 An intuitive approach is to request users furnish supplementary information delineating their interests,
25 such as providing an image of an individual and denoting interest in the hat (see Fig. 1). Numerous
26 industry leaders including Google, Amazon, and Pinterest have adopted this tactic, however academic
27 discourse on potential alternative methodologies for this task remains scarce as the domain lacks
28 dedicated datasets. For convenience, we propose terming this task Referred Visual Search (RVS),
29 as it is likely to garner attention from the computer vision community due to the utility for product
30 search in extensive catalogs.

31 In practice, object selection in complex scenes is classically tackled using object detection and
32 crops [21, 17, 12, 42]. Some recent approaches use categorical attributes [8] or text instead [6], and
33 automatically crop the image based on learned attention to input attributes. It is also possible to ask
34 the user to perform the crop himself, yet in all the situations the performance of the retrieval will be
35 sensitive to this extraction step making it costly to build a generic retrieval tool. Recently, Jiao et al.
36 [20] went a step further, incorporating prior knowledge about the taxonomy of fashion attributes and
37 classes without using crops. They use a multi-granularity loss and two sub-networks to learn attribute
38 and class-specific representations, resulting in improved robustness for fashion retrieval, yet without
39 providing any code.

40 In this work, we seek to support these efforts by providing a dataset dedicated to RVS. We extracted
41 a subset of LAION 5B [41] focused on pairs of images sharing a labeled similarity in the domain of
42 fashion, and propose a method to eliminate the need for explicit detection or segmentation, while still
43 producing similarities in the embedding space specific to the conditioning. We think that such end-to-
44 end approach has the potential to be more generalizable and robust, whereas localization-dependent
45 approaches hinge on multi-stage processing heuristics specific to the dataset.

46 This paper presents two contributions to the emerging field of Referred Visual Search, aiming at
47 defining image similarity based on conditioning information.

- 48 ✓ The introduction of a new dataset, referred to as LRVS-Fashion, which is derived from the
49 LAION-5B dataset and comprises 272k fashion products with nearly 842k images. This dataset
50 features a test set with an addition of more than 2M distractors, enabling the evaluation of method
51 robustness in relation to gallery size. The dataset’s pairs and additional metadata are designed to
52 necessitate the extraction of particular features from complex images.
- 53 ✓ An innovative method for learning to extract referred embeddings using weakly-supervised
54 training. Our approach demonstrates superior accuracy against a strong detection-based baseline
55 and existing published work. Furthermore, our method exhibits robustness against a large number
56 of distractors, maintaining high R@1 even when increasing the number of distractors to 2M.

57 2 Related Work

58 **Retrieval Datasets.** Standard datasets in metric learning literature consider that the images are
59 object-centric, and focus on single salient objects [49, 25, 45]. In the fashion domain there exist
60 multiple datasets dedicated to product retrieval, with paired images depicting the same product and
61 additional labeled attributes. A recurrent focus of such datasets is cross-domain retrieval, where the
62 goal is to retrieve images of a given product taken in different situations, for example consumer-to-
63 shop [31, 50, 32, 12], or studio-to-shop [32, 27]. The domain gap is in itself a challenge, with issues
64 stemming from irregular lighting, occlusions, viewpoints, or distracting backgrounds. However, the
65 query domain (consumer images for example) often contains scenes with multiple objects, making
66 queries ambiguous. This issue has been circumvented with the use of object detectors and landmarks
67 detectors [23, 18, 32, 12]. Some are not accessible anymore [23, 32, 50].

68 With more than 272k distinct training product identities captured in multi-instance scenes, our new
69 dataset proposes an exact matching task similar to the private Zalando dataset [27], while being larger
70 than existing fashion retrieval datasets and publicly available. We also create an opportunity for new
71 multi-modal approaches, with captions referring to the product of interest in each complex image,
72 and for robustness to gallery size with 2M added distractors at test time.

73 **Instance Retrieval.** In the last decade, content-based image retrieval has changed because of the
74 arrival of deep learning, which replaced many handcrafted heuristics (keypoint extraction, descriptors,
75 geometric matching, re-ranking. . .) [11]. In the industry this technology has been of interest to retail
76 companies and search engines to develop visual search solutions, with new challenges stemming from
77 the large scale of such databases. Initially using generic pretrained backbones to extract embeddings
78 with minimal retraining [53], methods have evolved toward domain-specific embeddings supervised
79 by semantic labels, and then multi-task domain-specific embeddings, leveraging additional product
80 informations [58, 3, 46]. The latest developments in the field incorporate multi-modal features for
81 text-image matching [59, 54, 62], with specific vision-language pretext tasks.

82 However, these methods often consider that the query image is unambiguous, and often rely on a
83 region proposal system to crop the initial image [21, 60, 17, 42, 3, 10]. In our work, we bypass this
84 step and propose an end-to-end framework, leveraging the Transformer architecture to implicitly
85 perform this detection step conditionally to the referring information.

86 **Referring Tasks.** Referring tasks are popular in vision-language processing, in particular Referring
87 Expression Comprehension and Segmentation where a sentence designates an object in a scene,
88 that the network has to localize. For the comprehension task (similar to open-vocabulary object
89 detection) the goal is to output a bounding box [34, 56, 57, 30]. The segmentation task aims at
90 producing an instance mask for images [61, 34, 19, 7, 24] and recently videos [52, 4]. In this paper,
91 we propose a referring expression task, where the goal is to embed the designated object of interest
92 into a representation that can be used for retrieval. We explore the use of Grounding DINO [30] and
93 Segment Anything [24] to create a strong baseline on our task.

94 **Conditional Embeddings.** Conditional similarity search has been studied through the retrieval
95 process and the embedding process. On one hand, for the retrieval process, Hamilton et al. [15]
96 propose to use a dynamically pruned random projection tree. On the other hand, previous work
97 in conditional visual similarity learning focused on attribute-specific retrieval, defining different
98 similarity spaces depending on chosen discriminative attributes [47, 36]. They use either a mask
99 applied on the features [47], or different projection heads [36], and require extensive data labeling.

100 In Fashion, ASEN [35] uses spatial and channel attention to an attribute embedding to extract specific
101 features in a global branch. Dong et al. [8] and Das et al. [6] build upon this model and add a local
102 branch working on an attention-based crop. Recently, Jiao et al. [20] incorporated prior knowledge
103 about fashion taxonomy in this process to create class-conditional embeddings based on known
104 fine-grained attributes, using multiple attribute-conditional attention modules. In a different domain,
105 Asai et al. [1] tackle a conditional document retrieval task, where the user intent is made explicit by
106 concatenating instructions to the query documents. In our work, we use Vision Transformers [9] to
107 implicitly pool features depending on the conditioning information, without relying on explicit ROI
108 cropping or labeled fine-grained attributes.

109 Composed Image Retrieval (CIR) [48] is another retrieval task where the embedding of an image must
110 be modified following a given instruction. Recent methods use a composer network after embedding
111 the image and the modifying text [28, 5, 2]. While CIR shares similarities with RVS in terms of inputs
112 and outputs, it differs conceptually. Our task focuses on retrieving items based on depicted attributes
113 and specifying a similarity computation method, rather than modifying the image. In Fashion, CIR
114 has been extended to dialog-based interactive retrieval, where an image query is iteratively refined
115 following user instructions [14, 51, 55, 16].

116 3 Dataset

117 Metric learning methods work by extracting features that pull together images labeled as similar [11].
118 In our case, we wanted to create a dataset where this embedding has to focus on a specific object
119 in a scene to succeed. We found such images in fashion, thanks to a standard practice in this field
120 consisting in taking pictures of the products alone on neutral backgrounds, and worn by models in
121 scenes involving other clothing items (see Fig. 3).

122 We created LAION-RVS-Fashion (abbreviated LRVS-F) from LAION-5B by collecting images of
123 products isolated and in context, which we respectively call *simple* and *complex*. We grouped them
124 using extracted product identifiers. We also gathered and created a set of metadata to be used as

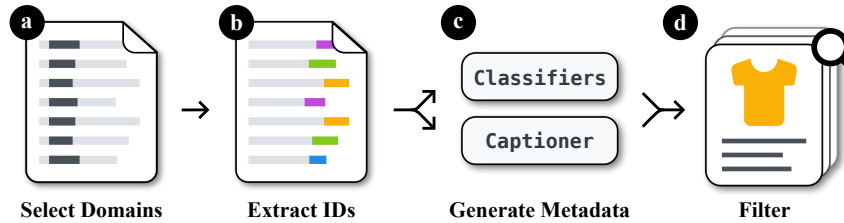


Figure 2: Overview of the data collection. *a)* Selection of a subset of domains belonging to known fashion retailers. *b)* Extraction of product identifiers in the URLs using domain-specific regular expressions. *c)* Generation of synthetic metadata for the products (categories, captions, ...) using both pretrained and finetuned models. *d)* Deduplication of the images, and assignment to subsets.

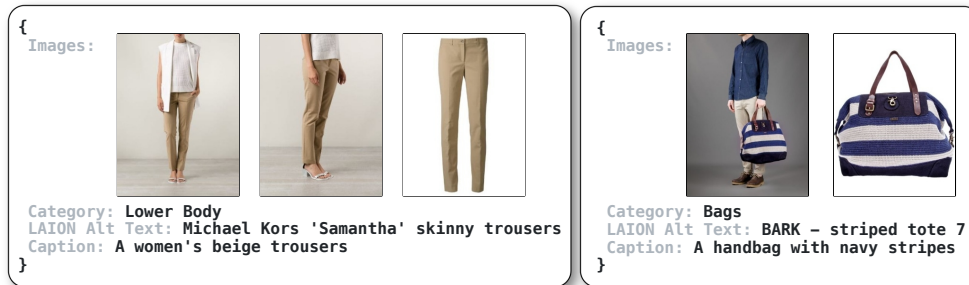


Figure 3: Samples from LRVS-F. Each product is represented on at least a simple and a complex image, and is associated with a category. The simple images are also described by captions from LAION and BLIP2. Please refer to Appendix [A.1](#) for more samples.

125 referring information, namely LAION captions, generated captions, and generated item categories.
 126 The process is depicted Fig. [2](#) presented in Section [3.1](#) with additional details in Appendix [A.3](#)

127 3.1 Construction

128 **Image Collection.** The URLs in LRVS-F are a subset of LAION-5B, curated from content delivery
 129 networks of fashion brands and retailers. By analyzing the URL structures we identified product
 130 identifiers, which we extracted with regular expressions to recreate groups of images depicting the
 131 same product. URLs without distinct identifiers or group membership were retained as distractors.

132 **Annotations.** We generated synthetic labels for the image complexity, the category of the product,
 133 and added new captions to replace the noisy LAION alt-texts. For the complexity labels, we
 134 employed active learning to incrementally train a classifier to discern between isolated objects on
 135 neutral backdrops and photoshoot scenes. The product categories were formed by aggregating various
 136 fine-grained apparel items into 10 coarse groupings. This categorization followed the same active
 137 learning protocol. Furthermore, the original LAION captions exhibited excessive noise, including
 138 partial translations or raw product identifiers. Therefore, we utilized BLIP-2 [\[29\]](#) to generate new,
 139 more descriptive captions.

140 **Dataset Split.** We grouped together images associated to the same product identifier and dropped
 141 the groups that did not have at least a simple and a complex image. We manually selected 400 of
 142 them for the validation set, and 2,000 for the test set. The distractors are all the images downloaded
 143 previously that were labeled as "simple" but not used in product groups. This mostly includes images
 144 for which it was impossible to extract any product identifier.

145 **Dataset Cleaning.** In order to mitigate false negatives in our results, we utilized Locality Sensitive
 146 Hashing and OpenCLIP ViT-B/16 embeddings to eliminate duplicates. Specifically, we removed
 147 duplicates between the test targets and test distractors, as well as between the validation targets and
 148 validation distractors. Throughout our experiments, we did not observe any false negatives in the
 149 results. However, there remains a small quantity of near-duplicates among the distractor images.

150 3.2 Composition

151 In total, we extracted 272,451 products for training, represented in 841,718 images. This represents
152 581,526 potential simple/complex positive pairs. We additionally extracted 400 products (800 images)
153 to create a validation set, and 2,000 products (4,000 images) for a test set. We added 99,541 simple
154 images in the validation gallery as distractors, and 2,000,014 in the test gallery.

155 We randomly sampled images and manually verified the quality of the labels. For the complexity
156 labels, we measured an empirical error rate of 1/1000 on the training set and 3/1000 for the distractors.
157 For the product categories, we measured a global empirical error rate of 1%, with confusions mostly
158 arising from semantically similar categories and images where object scale was ambiguous in isolated
159 settings (e.g. long shirt vs. short dress, wristband vs. hairband). The BLIP2 captions we provided
160 exhibit good quality, increasing the mean CLIP similarity with the image by +7.4%. However, as
161 synthetic captions, they are not perfect and may contain occasional hallucinations.

162 Please refer to Appendix [A.4](#) for metadata details, [A.5](#) for considerations regarding privacy and biases
163 and [C](#) for metadata details and a datasheet [\[13\]](#).

164 3.3 Benchmark

165 We define a benchmark on LRVS-F to evaluate different methods on a held-out test set with a large
166 number of distractors. The test set contains 2,000 unseen products, and up to 2M distractors. Each
167 product in the set is represented by a pair of images - a simple one and a complex one. The objective
168 of the retrieval task is to retrieve the simple image of each product from among a vast number of
169 distractors and other simple test images, given the complex image and conditioning information.

170 For this dataset, we propose to frame the benchmark as an asymmetric task : the representation of
171 simple images (the gallery) should not be computed conditionally. This choice is motivated by three
172 reasons. First, when using precise free-form conditioning (such as LAION texts, which contain
173 hashed product identifiers and product names) a symmetric encoding would enable a retrieval based
174 solely on this information, completely disregarding the image query. Second, for discrete (categorical)
175 conditioning it allows the presence of items of unknown category in the gallery, which is a situation
176 that may occur in distractors. Third, these images only depict a single object, thus making referring
177 information unnecessary. A similar setting is used by Asai et al. [\[1\]](#).

178 Additionally, we provide a list of subsets sampled with replacement to be used for bootstrapped
179 estimation of confidence intervals on the metrics. We created 10 subsets of 1000 test products, and
180 10 subsets of 10K, 100K and 1M distractors. We also propose a validation set of 400 products with
181 nearly 100K other distractors to monitor the training and for hyperparameter search.

182 4 Conditional Embedding

183 **Task Formulation.** Let x_q be a query image containing several objects of interest (e.g., a person
184 wearing many different clothes and items), and c_q the associated referring information that provides
185 cues about what aspect of x_q is relevant for the query (e.g., a text describing which garment is of
186 interest, or directly the class of the garment of interest). Similarly, let x_t be a target image, described
187 by the latent information c_t . The probability of x_t to be relevant for the query x_q is given by the
188 conditional probability $P(x_t, c_t | x_q, c_q)$. When working with categories for c_q and c_t , a filtering
189 strategy consists in assuming independence between the images and their category,

$$P(x_t, c_t | x_q, c_q) = P(x_t | x_q) P(c_t | c_q), \quad (1)$$

190 and further assuming that categories are uncorrelated (i.e., $P(c_t | c_q) = \delta_{c_q=c_t}$ with δ the Dirac
191 distribution). In this work, we remove those assumptions and instead assume that $P(x_t, c_t | x_q, c_q)$
192 can be directly inferred by a deep neural network model. More specifically, we propose to learn a
193 flexible embedding function ϕ such that

$$\langle \phi(x_q, c_q), \phi(x_t, c_t) \rangle \propto P(x_t, c_t | x_q, c_q). \quad (2)$$

194 Our approach offers a significant advantage by allowing the flexibility to change the conditioning
195 information (c_q) at query time, resulting in a different representation that focuses on different aspects
196 of the image. It is also *weakly supervised* in the sense that the referring information c_q is not required

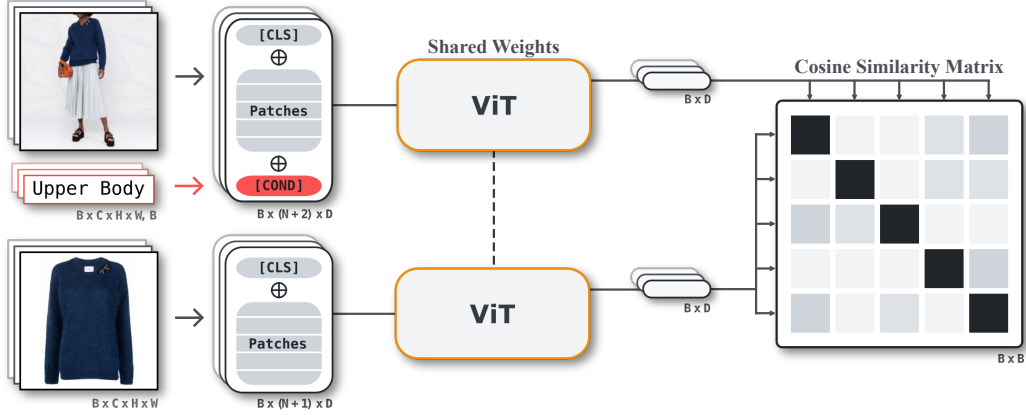


Figure 4: Overview of our method on LRV5-F. For each element in a batch, we embed the scene conditionally and the isolated item unconditionally. We optimize an InfoNCE loss over the cosine similarity matrix. \oplus denotes concatenation to the patch sequence.

197 to provide localized information about the content of interest (like a bounding box) and can be as
 198 imprecise as a free-form text, as shown in Fig. 1

199 **Method:** We implement ϕ by modifying the Vision Transformer (ViT) architecture [9]. The condi-
 200 tioning is an additional input token with an associated learnable positional encoding, concatenated
 201 to the sequence of image patches. The content of this token can either be learned directly (*e.g.* for
 202 discrete categorical conditioning), or be generated by another network (*e.g.* for textual conditioning).
 203 At the end of the network, we linearly project the [CLS] token to map the features to a metric
 204 space. We experimented with concatenating at different layers in the transformer, and found that
 205 concatenating before the first layer is the most sensible choice (see Appendix B.1).

206 We train the network with the InfoNCE loss [44, 38], following CLIP [40], which is detailed in the
 207 next paragraph. However, we hypothesize that even though our method relies on a contrastive loss,
 208 it does not explicitly require a specific formulation of it. We choose the InfoNCE loss because of
 209 its popularity and scalability. During training, given a batch of N pairs of images and conditioning
 210 $((x_i^A, c_i^A); (x_i^B, c_i^B))_{i=1..N}$, we compute their conditional embeddings $(z_i^A, z_i^B)_{i=1..N}$ with $z =$
 211 $\phi(x, c) \in \mathbb{R}^d$. We compute a similarity matrix S where $S_{ij} = s(z_i^A, z_j^B)$, with s the cosine similarity.
 212 We then optimize the similarity of the correct pair with a cross-entropy loss, effectively considering
 213 the $N - 1$ other products in the batch as negatives:

$$l(S) = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(S_{ii}\tau)}{\sum_{j=1}^N \exp(S_{ij}\tau)}, \quad (3)$$

214 with τ a learned temperature parameter, and the final loss is $\mathcal{L} = l(S)/2 + l(S^\top)/2$. Please refer to
 215 Fig. 4 for an overview of the method. The τ parameter is used to follow the initial formulation of
 216 CLIP [40] and is optimized by gradient during the training. At test time, we use FAISS [22] to create
 217 a unique index for the entire gallery and perform fast similarity search on GPUs.

218 5 Experiments

219 We compare our method to various baselines on LRV5-F, using both category- and caption-based
 220 settings. We report implementation details before analyzing the results.

221 5.1 Implementation details

222 All our models take as input images of size 224×224 , and output an embedding vector of 512
 223 dimensions. We use CLIP weights as initialization, and then train our models for 30 epochs with
 224 AdamW [33] and a maximum learning rate of 10^{-5} determined by a learning rate range test [43]. To
 225 avoid distorting pretrained features [26], we start by only training the final projection and new input

226 embeddings (conditioning and positional) for a single epoch, with a linear warm-up schedule. We
 227 then train all parameters for the rest of the epochs with a cosine schedule.

228 We pad the images to a square with white pixels, before resizing the largest side to 224 pixels. During
 229 training, we apply random horizontal flip, and random resized crops covering at least 80% of the
 230 image area. We evaluate the Recall at 1 (R@1) of the model on the validation set at each epoch, and
 231 report test metrics (recall and categorical accuracy) for the best performing validation checkpoint.

232 We used mixed precision and sharded loss to run our experiments on multiple GPUs. B/32 models
 233 were trained for 6 hours on 2 V100 GPUs, with a total batch size of 360. B/16 were trained for 9
 234 hours on 12 V100, with a batch size of 420. Batch sizes were chosen to maximize GPU memory use.

235 5.2 Results

236 **Detection-based Baseline** We leveraged the recent Grounding DINO [30] and Segment Anything
 237 [24] to create a baseline approach based on object detection and segmentation. In this setting, we
 238 feed the model the query image and conditioning information, which can be either the name of the
 239 category or a caption. Subsequently, we use the output crops or masks to train a ViT following the
 240 aforementioned procedure. Please refer to Tab. 1 for the results.

241 Initial experiments conducted with pretrained CLIP features showed a slight preference toward
 242 segmenting the object. However, training the image encoder revealed that superior performances
 243 can be attained by training the network on crops. Our supposition is that segmentation errors lead to
 244 definitive loss of information, whereas the network’s capacity is sufficient for it to learn to disregard
 245 irrelevant information and recover from a badly cropped image.

246 Overall, using Grounding DINO makes for a strong baseline. However, it is worth highlighting that
 247 the inherent imprecision of category names frequently results in overly large bounding boxes, which
 248 in turn limits the performances of the models. Indeed, adding more information into the dataset such
 249 as bounding boxes with precise categories would help, yet this would compromise the scalability
 250 of the model as such data is costly to obtain. Conversely, the more precise boxes produced by the
 251 caption-based model reach 67.8%R@1 against 2M distractors.

Table 1: Comparisons of results on LRVS-F for localization-based models. For 0, 10K, 100K and 1M distractors, we report bootstrapped means and standards deviations estimated from 10 randomly sampled sets. We observe superior performances from the caption-based models, due to the precision of the caption which leads to better detections.

		Distractors →		+10K		+100K		+1M		+2M	
Condi.	Preprocessing	Embedding	%R@1	%Cat@1	%R@1	%Cat@1	%R@1	%Cat@1	%R@1	%Cat@1	
Category	Gr. DINO-T + SAM-B	CLIP ViT-B/32	16.9 ± 1.45	67.4 ± 1.70	8.9 ± 0.79	65.6 ± 1.93	4.4 ± 0.44	64.5 ± 1.48	2.9	64.0	
	Gr. DINO-T + SAM-B	ViT-B/32	83.0 ± 1.06	94.6 ± 0.75	69.4 ± 1.36	92.0 ± 0.67	53.1 ± 1.63	90.0 ± 0.77	46.4	89.2	
	Gr. DINO-T	ViT-B/32	88.7 ± 0.74	96.4 ± 0.55	77.0 ± 1.79	94.3 ± 0.82	62.8 ± 1.92	92.2 ± 1.26	56.0	91.8	
	Gr. DINO-B	ViT-B/16	89.9 ± 0.87	96.2 ± 0.77	80.8 ± 1.35	94.5 ± 0.73	68.8 ± 2.17	93.2 ± 0.90	62.9	92.5	
Caption	Gr. DINO-T + SAM-B	CLIP ViT-B/32	27.3 ± 1.29	72.9 ± 1.68	16.3 ± 0.86	71.1 ± 1.17	9.1 ± 0.73	70.1 ± 1.56	6.2	69.8	
	Gr. DINO-T + SAM-B	ViT-B/32	83.5 ± 1.56	94.6 ± 0.39	72.2 ± 1.59	93.0 ± 0.42	56.5 ± 1.61	90.9 ± 0.74	50.8	90.2	
	Gr. DINO-T	ViT-B/32	89.7 ± 0.76	96.7 ± 0.74	79.0 ± 0.82	95.1 ± 0.74	65.4 ± 2.03	93.1 ± 1.14	59.0	92.0	
	Gr. DINO-B	ViT-B/16	91.6 ± 0.77	97.6 ± 0.31	83.6 ± 0.93	96.1 ± 0.60	73.6 ± 1.49	94.7 ± 0.64	67.8	94.3	

252 **Categorical Conditioning** We compare our method with categorical detection-based approaches,
 253 and unconditional ViTs finetuned on our dataset. To account for the extra conditioning information
 254 used in our method, we evaluated the latter on filtered indexes, with only products belonging to the
 255 correct category. We did not try to predict the item of interest from the input picture, and instead
 256 consider it as a part of the query. We also report unfiltered metrics for reference. Results are in Tab. 2.

257 Training the ViTs on our dataset greatly improves their performances, both in terms of R@1 and
 258 categorical accuracy. Filtering the gallery brings a modest mean gain of 2 – 4%R@1 across all
 259 quantities of distractors (Fig. 4b), reaching 62.4%R@1 for 2M distractors with a ViT-B/16 architecture.
 260 In practice, this approach is impractical as it necessitates computing and storing an index for each
 261 category to guarantee a consistent quantity of retrieved items. Moreover, a qualitative evaluation of
 262 the filtered results reveals undesirable behaviors. When filtering on a category divergent from the
 263 network’s intrinsic focus, we observe the results displaying colors and textures associated with the
 264 automatically focused object rather than the requested one.

Table 2: Comparisons of results on LRVS-F for unconditional, category-based and caption-based models. For 0, 10K, 100K and 1M distractors, we report bootstrapped means and standards deviations from 10 randomly sampled sets. Our CondViT-B/16 outperforms other methods for both groups.

Model	+10K		+100K		+1M		+2M	
	%R@1	%Cat@1	%R@1	%Cat@1	%R@1	%Cat@1	%R@1	%Cat@1
ViT-B/32	85.6 \pm 1.08	93.7 \pm 0.31	73.4 \pm 1.35	90.9 \pm 0.78	58.5 \pm 1.37	87.8 \pm 0.86	51.7	86.9
ViT-B/16	88.4 \pm 0.88	94.8 \pm 0.52	79.0 \pm 1.02	92.3 \pm 0.73	66.1 \pm 1.21	90.2 \pm 0.92	59.4	88.8
ASEN _g [8]	63.1 \pm 1.50	76.3 \pm 1.26	46.1 \pm 1.21	68.5 \pm 0.84	29.8 \pm 1.86	62.9 \pm 1.27	24.1	62.0
ViT-B/32 + Filt.	88.9 \pm 1.01	—	76.8 \pm 1.24	—	62.0 \pm 1.31	—	55.1	—
CondViT-B/32 - Category	90.9 \pm 0.98	99.2 \pm 0.31	80.2 \pm 1.55	98.8 \pm 0.39	65.8 \pm 1.42	98.4 \pm 0.65	59.0	98.0
ViT-B/16 + Filt.	90.9 \pm 0.88	—	81.9 \pm 0.87	—	68.9 \pm 1.11	—	62.4	—
CondViT-B/16 - Category	93.3 \pm 1.04	99.5 \pm 0.25	85.6 \pm 1.06	99.2 \pm 0.35	74.2 \pm 1.82	99.0 \pm 0.42	68.4	98.8
CoSMo [28]	88.3 \pm 1.30	97.6 \pm 0.45	76.1 \pm 1.85	96.0 \pm 0.32	59.1 \pm 1.42	94.7 \pm 0.40	52.1	94.8
CLIP4CIR [2]	92.9 \pm 0.64	99.0 \pm 0.33	81.9 \pm 1.63	98.1 \pm 0.68	66.9 \pm 2.05	96.5 \pm 0.67	59.1	95.5
CondViT-B/32 - Caption	92.7 \pm 0.77	99.1 \pm 0.30	82.8 \pm 1.22	98.7 \pm 0.40	68.4 \pm 1.50	98.1 \pm 0.43	62.1	98.0
CondViT-B/16 - Caption	94.2 \pm 0.90	99.4 \pm 0.37	86.4 \pm 1.13	98.9 \pm 0.49	74.6 \pm 1.65	98.4 \pm 0.58	69.3	98.2

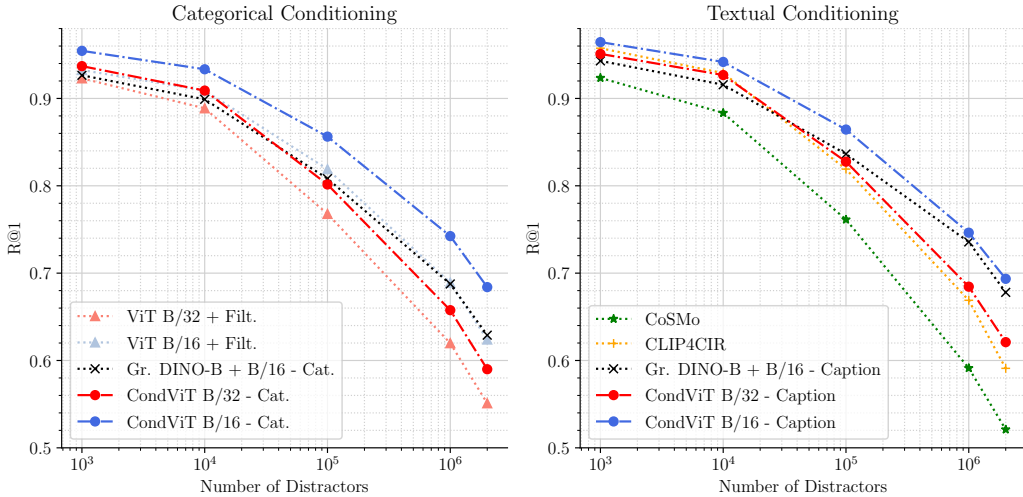


Figure 5: R@1 with respects to number of added distractors, evaluated on the entire test set. Please refer to Tab. 1 and 2 for bootstrapped metrics and confidence intervals. Our categorical CondViT-B/16 reaches the performances of the best caption-based models, while using a sparser conditioning.

265 We also compare with ASEN [8] trained on our dataset using the authors’ released code. This
 266 conditional architecture uses a global and a local branch with conditional spatial attention modules,
 267 respectively based on ResNet50 and ResNet34 backbones, with explicit ROI cropping. However
 268 in our experiments the performances decrease with the addition of the local branch in the second
 269 training stage, even after tuning the hyperparameters. We report results for the global branch.

270 We train our CondViT using the categories provided in our dataset, learning an embedding vector
 271 for each of the 10 clothing categories. For the i -th product in the batch, we randomly select in the
 272 associated data a simple image x_s and its category c_s , and a complex image x_c . We then compute
 273 their embeddings $z_i^A = \phi(x_c, c_s)$, $z_i^B = \phi(x_s)$. We also experimented with symmetric conditioning,
 274 using a learned token for the gallery side (see Appendix B.1).

275 Our categorical CondViT-B/16, with 68.4%R@1 against 2M distractors significantly outperforms
 276 all other category-based approaches (see Fig. 5, left) and maintains a higher categorical accuracy.
 277 Furthermore, it performs similarly to the detection-based method conditioned on richer captions,
 278 while requiring easy-to-acquire coarse categories. It does so without making any assumption on the
 279 semantic nature of these categories, and adding only a few embedding weights (7.7K parameters) to
 280 the network, against 233M parameters for Grounding DINO-B. We confirm in Appendix B.2 that its
 281 attention is localized on different objects depending on the conditioning.



Figure 6: Qualitative results for our categorical (first 2 rows) and textual (last 2 rows) CondViT-B/16. We use free-form textual queries instead of BLIP2 captions to illustrate realistic user behavior, and retrieve from the whole test gallery. See Fig. [13](#) and [14](#) in the Appendix for more qualitative results.

282 **Textual Conditioning** To further validate our approach, we replaced the categorical conditioning
 283 with referring expressions, using our generated BLIP2 captions embedded by a Sentence T5-XL
 284 model [\[37\]](#). We chose this model because it embeds the sentences in a 768-dimensional vector,
 285 allowing us to simply replace the categorical token. We pre-computed the caption embeddings, and
 286 randomly used one of them instead of the product category at training time. At test time, we used the
 287 first caption.

288 In Tab. [2](#), we observe a gain of 3.1%R@1 for the CondViT-B/32 architecture, and 0.9%R@1 for
 289 CondViT-B/16, compared to categorical conditioning against 2M distractors, most likely due to the
 290 additional details in the conditioning sentences. When faced with users, this method allows for more
 291 natural querying, with free-form referring expressions. See Figure [6](#) for qualitative results.

292 We compare these models with CIR methods: CoSMo [\[28\]](#) and CLIP4CIR [\[2\]](#). Both use a compositor
 293 network to fuse features extracted from the image and accompanying text. CoSMo reaches perfor-
 294 mances similar to an unconditional ViT-B/32, while CLIP4CIR performs similarly to our textual
 295 CondViT-B/32. We hypothesize that for our conditional feature extraction task, early condition-
 296 ing is more effective than modifying embeddings through a compositor at the network’s end. Our
 297 CondViT-B/16 model significantly outperforms all other models and achieves results comparable to
 298 our caption-based approach using Grounding DINO-B (see Fig. [5](#), right). As the RVS task differs
 299 from CIR, despite both utilizing identical inputs, this was anticipated. Importantly, CondViT-B/16
 300 accomplishes this without the need for explicit detection steps or dataset-specific preprocessing.
 301 Notably, we observe that our models achieve a categorical accuracy of 98% against 2M distractors,
 302 surpassing the accuracy of the best corresponding detection-based model, which stands at 94.3%.

303 6 Conclusion & Limitations

304 We studied an approach to image similarity in fashion called Referred Visual Search (RVS), which
 305 introduces two significant contributions. Firstly, we introduced the LAION-RVS-Fashion dataset,
 306 comprising 272K fashion products and 842K images. Secondly, we proposed a simple weakly-
 307 supervised learning method for extracting referred embeddings. Our approach outperforms strong
 308 detection-based baselines. These contributions offer valuable resources and techniques for advancing
 309 image retrieval systems in the fashion industry and beyond.

310 However, one limitation of our approach is that modifying the text description to refer to something
 311 not present or not easily identifiable in the image does not work effectively. For instance, if the
 312 image shows a person carrying a green handbag, a refined search with "red handbag" as a condition
 313 would only retrieve a green handbag. The system may also ignore the conditioning if the desired
 314 item is small or absent in the database. Examples of such failures are illustrated in Appendix [B.3](#).
 315 Additionally, extending the approach to more verticals would be relevant.

316 **References**

- 317 [1] Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh
318 Hajishirzi, and Wen-tau Yih. Task-aware retrieval with instructions. *arXiv preprint arXiv:2211.09260*,
319 2022. [3](#) [5](#)
- 320 [2] Alberto Baldrati, Marco Bertini, Tiberio Uricchio, and Alberto Del Bimbo. Composed image retrieval using
321 contrastive learning and task-oriented clip-based features. *ACM Transactions on Multimedia Computing,*
322 *Communications and Applications*, 2023. [3](#) [8](#) [9](#)
- 323 [3] Sean Bell, Yiqun Liu, Sami Alsheikh, Yina Tang, Edward Pizzi, M. Henning, Karun Singh, Omkar Parkhi,
324 and Fedor Borisyuk. GrokNet: Unified Computer Vision Model Trunk and Embeddings For Commerce. In
325 *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*.
326 ACM, 2020. ISBN 978-1-4503-7998-4. doi: 10.1145/3394486.3403311. [3](#)
- 327 [4] Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-end referring video object segmentation
328 with multimodal transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and*
329 *Pattern Recognition (CVPR)*, 2022. [3](#)
- 330 [5] Yiyang Chen, Zhedong Zheng, Wei Ji, Leigang Qu, and Tat-Seng Chua. Composed image retrieval with
331 text feedback via multi-grained uncertainty regularization, 2022. [3](#)
- 332 [6] Nilotpal Das, Aniket Joshi, Promod Yenigalla, and Gourav Agrwal. MAPS: Multimodal Attention for
333 Product Similarity. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*
334 *(WACV)*, 2022. [2](#) [3](#)
- 335 [7] Henghui Ding, Chang Liu, Suchen Wang, and Xudong Jiang. Vision-Language Transformer and Query
336 Generation for Referring Segmentation. In *2021 IEEE/CVF International Conference on Computer Vision*
337 *(ICCV)*. IEEE, 2021. ISBN 978-1-66542-812-5. doi: 10.1109/ICCV48922.2021.01601. [3](#)
- 338 [8] Jianfeng Dong, Zhe Ma, Xiaofeng Mao, Xun Yang, Yuan He, Richang Hong, and Shouling Ji. Fine-Grained
339 Fashion Similarity Prediction by Attribute-Specific Embedding Learning. *IEEE Transactions on Image*
340 *Processing*, 2021. ISSN 1057-7149, 1941-0042. doi: 10.1109/TIP.2021.3115658. [2](#) [3](#) [8](#)
- 341 [9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas
342 Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit,
343 and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In
344 *International Conference on Learning Representations*, 2021. [3](#) [6](#)
- 345 [10] Ming Du, Arnau Ramisa, Amit Kumar K C, Sampath Chanda, Mengjiao Wang, Neelakandan Rajesh,
346 Shasha Li, Yingchuan Hu, Tao Zhou, Nagashri Lakshminarayana, Son Tran, and Doug Gray. Amazon
347 Shop the Look: A Visual Search System for Fashion and Home. In *Proceedings of the 28th ACM SIGKDD*
348 *Conference on Knowledge Discovery and Data Mining*. ACM, 2022. ISBN 978-1-4503-9385-0. doi:
349 10.1145/3534678.3539071. [3](#)
- 350 [11] Shiv Ram Dubey. A Decade Survey of Content Based Image Retrieval using Deep Learning. *IEEE*
351 *Transactions on Circuits and Systems for Video Technology*, 2022. ISSN 1051-8215, 1558-2205. doi:
352 10.1109/TCSVT.2021.3080920. [3](#)
- 353 [12] Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaoou Tang, and Ping Luo. DeepFashion2: A Versatile
354 Benchmark for Detection, Pose Estimation, Segmentation and Re-Identification of Clothing Images. In
355 *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2019. ISBN
356 978-1-72813-293-8. doi: 10.1109/CVPR.2019.00548. [2](#)
- 357 [13] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal
358 Daumé III, and Kate Crawford. Datasheets for datasets, 2021. [5](#)
- 359 [14] Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. Dialog-based
360 Interactive Image Retrieval. In *Advances in Neural Information Processing Systems*. Curran Associates,
361 Inc., 2018. [3](#)
- 362 [15] Mark Hamilton, Stephanie Fu, Mindren Lu, Johnny Bui, Darius Bopp, Zhenbang Chen, Felix Tran,
363 Margaret Wang, Marina Rogers, Lei Zhang, Chris Hoder, and William T. Freeman. MosAlc: Finding
364 Artistic Connections across Culture with Conditional Image Retrieval. In *Proceedings of the NeurIPS 2020*
365 *Competition and Demonstration Track*. PMLR, 2021. [3](#)
- 366 [16] Xiao Han, Sen He, Li Zhang, Yi-Zhe Song, and Tao Xiang. UIGR: Unified Interactive Garment Retrieval.
367 In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. [3](#)

- 368 [17] Houdong Hu, Yan Wang, Linjun Yang, Pavel Komlev, Li Huang, Xi Chen, Jiawei Huang, Ye Wu, Meenaz
369 Merchant, and Arun Sacheti. Web-scale responsive visual search at bing. In *Proceedings of the 24th ACM*
370 *SIGKDD international conference on knowledge discovery & data mining*, 2018. [2](#) [3](#)
- 371 [18] Junshi Huang, Rogerio Feris, Qiang Chen, and Shuicheng Yan. Cross-Domain Image Retrieval with a
372 Dual Attribute-Aware Ranking Network. In *2015 IEEE International Conference on Computer Vision*
373 *(ICCV)*. IEEE, 2015. ISBN 978-1-4673-8391-2. doi: 10.1109/ICCV.2015.127. [2](#)
- 374 [19] Shaofei Huang, Tianrui Hui, Si Liu, Guanbin Li, Yunchao Wei, Jizhong Han, Luoqi Liu, and Bo Li.
375 Referring Image Segmentation via Cross-Modal Progressive Comprehension. In *2020 IEEE/CVF Confer-*
376 *ence on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020. ISBN 978-1-72817-168-5. doi:
377 10.1109/CVPR42600.2020.01050. [3](#)
- 378 [20] Yang (Andrew) Jiao, Yan Gao, Jingjing Meng, Jin Shang, and Yi Sun. Learning attribute and class-specific
379 representation duet for fine-grained fashion analysis. In *CVPR 2023*, 2023. [2](#) [3](#)
- 380 [21] Yushi Jing, David Liu, Dmitry Kislyuk, Andrew Zhai, Jiajing Xu, Jeff Donahue, and Sarah Tavel. Visual
381 search at pinterest. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge*
382 *Discovery and Data Mining*, 2015. [2](#) [3](#)
- 383 [22] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *IEEE Transac-*
384 *tions on Big Data*, 2019. [1](#) [6](#)
- 385 [23] M. Hadi Kiapour, Xufeng Han, Svetlana Lazebnik, Alexander C. Berg, and Tamara L. Berg. Where to
386 Buy It: Matching Street Clothing Photos in Online Shops. In *2015 IEEE International Conference on*
387 *Computer Vision (ICCV)*. IEEE, 2015. ISBN 978-1-4673-8391-2. doi: 10.1109/ICCV.2015.382. [2](#)
- 388 [24] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao,
389 Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything.
390 *arXiv:2304.02643*, 2023. [3](#) [7](#)
- 391 [25] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3D Object Representations for Fine-Grained
392 Categorization. In *2013 IEEE International Conference on Computer Vision Workshops*. IEEE, 2013.
393 ISBN 978-1-4799-3022-7. doi: 10.1109/ICCVW.2013.77. [2](#)
- 394 [26] Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-tuning can
395 distort pretrained features and underperform out-of-distribution. In *International Conference on Learning*
396 *Representations*, 2022. [6](#)
- 397 [27] Julia Lasserre, Katharina Rasch, and Roland Vollgraf. Studio2Shop: from studio photo shoots to fashion
398 articles. In *Proceedings of the 7th International Conference on Pattern Recognition Applications and*
399 *Methods*, 2018. doi: 10.5220/0006544500370048. [2](#)
- 400 [28] Seungmin Lee, Dongwan Kim, and Bohyung Han. Cosmo: Content-style modulation for image retrieval
401 with text feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern*
402 *Recognition (CVPR)*, 2021. [3](#) [8](#) [9](#)
- 403 [29] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training
404 with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. [4](#) [16](#) [20](#)
- 405 [30] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang,
406 Hang Su, Jun Zhu, et al. Grounding DINO: Marrying DINO with grounded pre-training for open-set object
407 detection. *arXiv preprint arXiv:2303.05499*, 2023. [3](#) [7](#)
- 408 [31] Si Liu, Zheng Song, Guangcan Liu, Changsheng Xu, Hanqing Lu, and Shuicheng Yan. Street-to-shop:
409 Cross-scenario clothing retrieval via parts alignment and auxiliary set. In *2012 IEEE Conference on*
410 *Computer Vision and Pattern Recognition*, 2012. [2](#)
- 411 [32] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. DeepFashion: Powering Robust Clothes
412 Recognition and Retrieval with Rich Annotations. In *2016 IEEE Conference on Computer Vision and*
413 *Pattern Recognition (CVPR)*. IEEE, 2016. ISBN 978-1-4673-8851-1. doi: 10.1109/CVPR.2016.124. [2](#)
- 414 [33] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on*
415 *Learning Representations*, 2019. [6](#)
- 416 [34] Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. Multi-
417 Task Collaborative Network for Joint Referring Expression Comprehension and Segmentation. In *2020*
418 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2020. ISBN 978-1-
419 72817-168-5. doi: 10.1109/CVPR42600.2020.01005. [3](#)

- 420 [35] Zhe Ma, Jianfeng Dong, Zhongzi Long, Yao Zhang, Yuan He, Hui Xue, and Shouling Ji. Fine-grained
421 fashion similarity learning by attribute-specific embedding network. In *Proceedings of the AAAI Conference*
422 *on Artificial Intelligence*, 2020. [3](#)
- 423 [36] Emily Mu and John Guttag. Conditional Contrastive Networks. In *NeurIPS 2022 First Table Representation*
424 *Workshop*, 2022. [3](#)
- 425 [37] Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei
426 Yang. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. In *Findings of the*
427 *Association for Computational Linguistics: ACL 2022*. Association for Computational Linguistics, 2022.
428 doi: 10.18653/v1/2022.findings-acl.146. [9](#)
- 429 [38] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive
430 coding. *arXiv preprint arXiv:1807.03748*, 2018. [6](#)
- 431 [39] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre
432 Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu,
433 Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel
434 Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski.
435 Dinov2: Learning robust visual features without supervision, 2023. [19](#)
- 436 [40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish
437 Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from
438 natural language supervision. In *International conference on machine learning*. PMLR, 2021. [6](#)
- 439 [41] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti,
440 Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale
441 dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. [2](#)
- 442 [42] Raymond Shiao, Hao-Yu Wu, Eric Kim, Yue Li Du, Anqi Guo, Zhiyuan Zhang, Eileen Li, Kunlong Gu,
443 Charles Rosenberg, and Andrew Zhai. Shop The Look: Building a Large Scale Visual Shopping System at
444 Pinterest. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery &*
445 *Data Mining*, 2020. doi: 10.1145/3394486.3403372. [2](#) [3](#)
- 446 [43] Leslie N Smith. Cyclical learning rates for training neural networks. In *2017 IEEE winter conference on*
447 *applications of computer vision (WACV)*. IEEE, 2017. [6](#)
- 448 [44] Kihyuk Sohn. Improved Deep Metric Learning with Multi-class N-pair Loss Objective. In *Advances in*
449 *Neural Information Processing Systems*. Curran Associates, Inc., 2016. [6](#)
- 450 [45] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep Metric Learning via Lifted
451 Structured Feature Embedding. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*
452 *(CVPR)*. IEEE, 2016. ISBN 978-1-4673-8851-1. doi: 10.1109/CVPR.2016.434. [2](#)
- 453 [46] Son Tran, R. Manmatha, and C. J. Taylor. Searching for fashion products from images in the wild. In *KDD*
454 *2019 Workshop on AI for Fashion*, 2019. [3](#)
- 455 [47] Andreas Veit, Serge Belongie, and Theofanis Karaletsos. Conditional similarity networks. In *Proceedings*
456 *of the IEEE conference on computer vision and pattern recognition*, 2017. [3](#)
- 457 [48] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing text and
458 image for image retrieval-an empirical odyssey. In *CVPR*, 2019. [3](#)
- 459 [49] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD
460 Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. [2](#)
- 461 [50] Xi Wang, Zhenfeng Sun, Wenqiang Zhang, Yu Zhou, and Yu-Gang Jiang. Matching User Photos to Online
462 Products with Robust Deep Features. In *Proceedings of the 2016 ACM on International Conference on*
463 *Multimedia Retrieval*. ACM, 2016. ISBN 978-1-4503-4359-6. doi: 10.1145/2911996.2912002. [2](#)
- 464 [51] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris.
465 Fashion IQ: A New Dataset Towards Retrieving Images by Natural Language Feedback. In *CVPR*, 2019. [3](#)
- 466 [52] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as Queries for Referring Video
467 Object Segmentation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
468 IEEE, 2022. ISBN 978-1-66546-946-3. doi: 10.1109/CVPR52688.2022.00492. [3](#)
- 469 [53] Fan Yang, Ajinkya Kale, Yury Bubnov, Leon Stein, Qiaosong Wang, Hadi Kiapour, and Robinson
470 Piramuthu. Visual Search at eBay. In *Proceedings of the 23rd ACM SIGKDD International Conference on*
471 *Knowledge Discovery and Data Mining*, 2017. doi: 10.1145/3097983.3098162. [3](#)

- 472 [54] Licheng Yu, Jun Chen, Animesh Sinha, Mengjiao Wang, Yu Chen, Tamara L Berg, and Ning Zhang. Com-
 473 mmercemm: Large-scale commerce multimodal representation learning with omni retrieval. In *Proceedings*
 474 *of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022. [3](#)
- 475 [55] Yifei Yuan and Wai Lam. Conversational Fashion Image Retrieval via Multiturn Natural Language
 476 Feedback. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development*
 477 *in Information Retrieval (SIGIR '21)*, 2021. [3](#)
- 478 [56] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-Grained Vision Language Pre-Training: Aligning Texts
 479 with Visual Concepts. In *Proceedings of the 39th International Conference on Machine Learning*. PMLR,
 480 2022. [3](#)
- 481 [57] Yan Zeng, Xinsong Zhang, Hang Li, Jiawei Wang, Jipeng Zhang, and Wangchunshu Zhou. X²-VLM:
 482 All-in-one pre-trained model for vision-language tasks. *arXiv preprint arXiv:2211.12402*, 2022. [3](#)
- 483 [58] Andrew Zhai, Hao-Yu Wu, Eric Tzeng, Dong Huk Park, and Charles Rosenberg. Learning a unified em-
 484 bedding for visual search at pinterest. In *Proceedings of the 25th ACM SIGKDD International Conference*
 485 *on Knowledge Discovery & Data Mining*, 2019. [3](#)
- 486 [59] Xunlin Zhan, Yangxin Wu, Xiao Dong, Yunchao Wei, Minlong Lu, Yichi Zhang, Hang Xu, and Xiaodan
 487 Liang. Product1M: Towards Weakly Supervised Instance-Level Product Retrieval via Cross-Modal
 488 Pretraining. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2021. ISBN
 489 978-1-66542-812-5. doi: 10.1109/ICCV48922.2021.01157. [3](#)
- 490 [60] Yanhao Zhang, Pan Pan, Yun Zheng, Kang Zhao, Yingya Zhang, Xiaofeng Ren, and Rong Jin. Visual
 491 Search at Alibaba. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge*
 492 *Discovery & Data Mining*, 2018. doi: 10.1145/3219819.3219820. [3](#)
- 493 [61] Yuting Zhang, Luyao Yuan, Yijie Guo, Zhiyuan He, I-An Huang, and Honglak Lee. Discriminative
 494 Bimodal Networks for Visual Localization and Detection with Natural Language Queries. In *2017 IEEE*
 495 *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017. ISBN 978-1-5386-0457-1.
 496 doi: 10.1109/CVPR.2017.122. [3](#)
- 497 [62] Xiaoyang Zheng, Zilong Wang, Ke Xu, Sen Li, Tao Zhuang, Qingwen Liu, and Xiaoyi Zeng. MAKE:
 498 Vision-Language Pre-training based Product Retrieval in Taobao Search. In *Companion Proceedings of the*
 499 *ACM Web Conference 2023*, 2023. doi: 10.1145/3543873.3584627. [3](#)

500 Checklist

- 501 1. For all authors...
- 502 (a) Do the main claims made in the abstract and introduction accurately reflect the paper’s
 503 contributions and scope? [\[Yes\]](#)
- 504 (b) Did you describe the limitations of your work? [\[Yes\]](#)
- 505 (c) Did you discuss any potential negative societal impacts of your work? [\[Yes\]](#) See
 506 Appendix [A.5](#)
- 507 (d) Have you read the ethics review guidelines and ensured that your paper conforms to
 508 them? [\[Yes\]](#) See Appendix [A.5](#)
- 509 2. If you are including theoretical results...
- 510 (a) Did you state the full set of assumptions of all theoretical results? [\[N/A\]](#)
- 511 (b) Did you include complete proofs of all theoretical results? [\[N/A\]](#)
- 512 3. If you ran experiments (e.g. for benchmarks)...
- 513 (a) Did you include the code, data, and instructions needed to reproduce the main ex-
 514 perimental results (either in the supplemental material or as a URL)? [\[Yes\]](#) See Ap-
 515 pendix [A.2](#)
- 516 (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they
 517 were chosen)? [\[Yes\]](#) See Section [5.1](#)
- 518 (c) Did you report error bars (e.g., with respect to the random seed after running experi-
 519 ments multiple times)? [\[Yes\]](#) See Table [1](#) and [2](#)
- 520 (d) Did you include the total amount of compute and the type of resources used (e.g., type
 521 of GPUs, internal cluster, or cloud provider)? [\[Yes\]](#) See Section [5.1](#)

- 522 4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
- 523 (a) If your work uses existing assets, did you cite the creators? [Yes]
- 524 (b) Did you mention the license of the assets? [Yes]
- 525 (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
- 526 We provide the URLs to the assets in Appendix [A.2](#)
- 527 (d) Did you discuss whether and how consent was obtained from people whose data you're
- 528 using/curating? [Yes]
- 529 (e) Did you discuss whether the data you are using/curating contains personally identifiable
- 530 information or offensive content? [Yes]
- 531 5. If you used crowdsourcing or conducted research with human subjects...
- 532 (a) Did you include the full text of instructions given to participants and screenshots, if
- 533 applicable? [N/A]
- 534 (b) Did you describe any potential participant risks, with links to Institutional Review
- 535 Board (IRB) approvals, if applicable? [N/A]
- 536 (c) Did you include the estimated hourly wage paid to participants and the total amount
- 537 spent on participant compensation? [N/A]