ConCon-Chi: Concept-Context Chimera Benchmark for Personalized Vision-Language Tasks

Andrea Rosasco*[1,2] Stefano Berti*[1,2] Giulia Pasquale*[2] Damiano Malafronte[2]
Shogo Sato[3] Hiroyuki Segawa[3] Tetsugo Inada[3] Lorenzo Natale[2]
[1]University of Genoa, IT [2]Istituto Italiano di Tecnologia, IT [3]Sony Interactive Entertainment Inc., JP

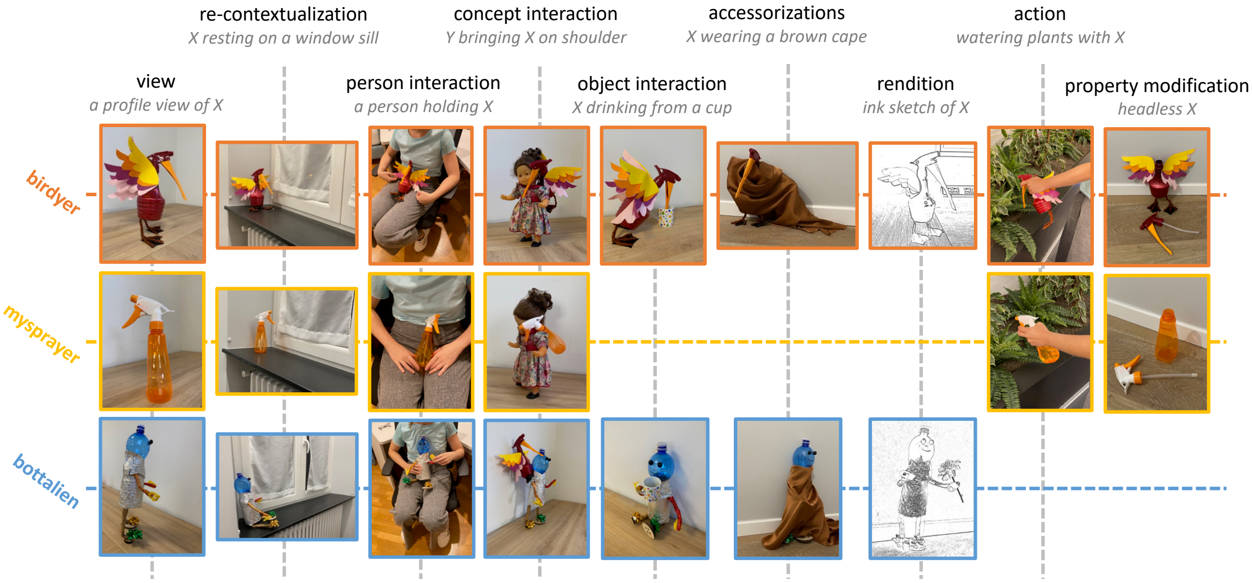

Figure 1. **Concept-context structure of the ConCon-Chi dataset.** Excerpt from the dataset structure (see Fig. 4). Each row represents images of a concept (concept name in Color) and each column images of a context (context kind in Black and context in Grey). The caption for an image is formed by the composition of the context with the concept (replacing X/Y with the concept textual identifier).

Abstract

While recent Vision-Language (VL) models excel at open-vocabulary tasks, it is unclear how to use them with specific or uncommon concepts. Personalized Text-to-Image Retrieval (TIR) or Generation (TIG) are recently introduced tasks that represent this challenge, where the VL model has to learn a concept from few images and respectively discriminate or generate images of the target concept in arbitrary contexts. We identify the ability to learn new meanings and their compositionality with known ones as two key properties of a personalized system. We show that the available benchmarks offer a limited validation of personalized textual concept learning from images with respect to the above properties and introduce ConCon-Chi as a benchmark for both personalized TIR and TIG, designed to fill this gap. We modelled the new-meaning concepts by crafting chimeric objects and formulating a large, varied set of contexts where we photographed each object. To promote the compositionality assessment of the learned concepts with known contexts, we combined different contexts with the same concept, and vice-versa. We carry out a thorough evaluation of state-of-the-art methods on the resulting dataset. Our study suggests that future work on personalized TIR and TIG methods should focus on the above key properties, and we propose principles and a dataset for their performance assessment. Dataset: `https://doi.org/10.48557/QJ1166` *and code:* `https://github.com/hsp-iit/concon-chi_benchmark`.

[*]Equal Contribution.

1. Introduction

Recent Vision-Language (VL) models for discriminative and generative tasks [18, 25, 26] excel at associating textual descriptions with images. This resulted in a paradigm shift from closed to open-vocabulary versions of several computer vision tasks. In these settings, the model is not bound to a closed set of predetermined classes but can operate on free-form textual descriptions. However, it can be difficult for a user to formulate a description such that the VL model returns the expected output [22, 32]. The problem boils down to finding a text input whose encoding is close to the visual embedding of the target concept. Since this mapping is typically learned from Web-scale data, crafting effective descriptions for uncommon, novel or specific concepts, is challenging, thus hampering performance of VL models.

To represent this challenge, the "personalized" versions of Text-to-Image Retrieval [5] (TIR) and Text-to-Image Generation [6] (TIG) have been recently proposed. The tasks consist of learning a user-specific *concept* using a few images and then performing retrieval (TIR) or generation (TIG) of such *concept* in a known *context*. Another similar task recently proposed is Zero-Shot Composed Image Retrieval (ZS-CIR) [2, 30], the retrieval of a reference image modified according to a relative caption [35].

Due to lack of benchmarks for personalization tasks, many works proposed their method alongside a new dataset. However, most of these benchmarks lack two key properties necessary for a thorough evaluation of personalized concept learning: **novel concepts** and **compositional** structure.

Existing datasets use, as concepts, instances of common objects (e.g., in PerVL, clothing items from DeepFashion2 [8]). However, these allow to evaluate the learning of new words for known concepts (closely related to synonym matching, or instance identification), rather than new concepts. To simulate a realistic personalization setting, inspired by [17], we introduce chimeric concepts: objects created by the union of two unrelated existing concepts (see Fig. 2).

The number and variability of contexts in which a concept appears is also underrepresented in personalized TIR/TIG benchmarks and this hampers the compositionality assessment of the newly learned concept with known contexts. Conversely, in a personalized TIR dataset multiple concepts should also appear in a same context, to prevent a method from attaining high performance by just attending to the context and disregarding the concept in the query ("context bias", pointed out also in ZS-CIR benchmarks [2, 30]). To avoid both these problems, we formulated a large, varied, set of contexts for each concept, while also ensuring that each context can be composed with several concepts. As a result, the dataset has a concept-context matrix structure (Fig. 1).

Our contributions are as follows:

- We highlight two key properties of personalized textual concept learning from images: learning new meanings and composing them with known ones; we propose the design of ConCon-Chi to model this problem.
- By evaluating on ConCon-Chi we show the limitations of state-of-the-art methods with respect to these properties.
- We release the dataset as a twofold benchmark for personalized TIR *and* TIG.

In the remainder of the paper, we compare related datasets in Sec. 2; we present ConCon-Chi in Sec. 3; we present our study for personalized TIR in Sec. 4 and for personalized TIG in Sec. 5. We report conclusions in Sec. 6.

2. Related Work

Learning Out Of Vocabulary (OOV) words. Learning textual concept representations from images can be seen as the multi-modal version of the NLP task of learning OOV word embeddings from a few examples [1] or the word definition [10, 11, 29]. We took inspiration from the Chimera dataset [17], which differentiates the problem of learning to associate new words with existing meanings (named entities, synonyms, aliases) from the one of learning new meanings, and models the latter with chimeric words that incorporate two unrelated concepts in a single one.

Personalized Text-to-Image Retrieval (TIR). Cohen *et al*. [5] introduced "Personalized Vision & Language" (PerVL), a setting where the vocabulary of a VL model is expanded with pseudo-tokens whose embeddings are learned from few images of user-specific concepts. The benchmark includes a retrieval and an instance segmentation dataset obtained by re-annotating images respectively from DeepFashion2 [8] and YouTube-VOS [38] with concept-context captions. PerVL includes many concept instances, which are however restricted to fashion or common items; moreover, it counts very few contexts per concept and a single concept per context, suffering from context bias. To tackle the proposed benchmark the authors propose PALAVRA, which we evaluate in this study.

The setting of personalized retrieval has been extended to videos [12, 39]. Recent work [24] proposed a similar approach to transfer CLIP [25] to a downstream image classification task by optimizing the class names.

Zero-Shot Composed Image Retrieval (ZS-CIR). This task can be seen as one-shot personalized TIR, where the reference image is the concept, and the relative caption the context. Thus personalization methods as PALAVRA have been evaluated for ZS-CIR [2]. Similarly, we consider methods as Pic2Word [30] and SEARLE [2] for evaluation on the proposed personalized TIR benchmark. However, we remark that the two settings are different. In recent ZS-CIR datasets (e.g., CIRR [20] and CIRCO [2]), images are drawn from an open domain (respectively NLVR^2 [33] and COCO [19]), thus concepts are defined at a semantic cat-

Task	Dataset	Concept Instances	Concept Type	Concepts /Query	Images /Query	Split	Queries	Pool	Concepts	Contexts	Contexts /Concept	Concepts /Context
pers TIR **pers TIG**	ConCon-Chi	**yes**	**chimeric** & common	**≤ 3**	**6.07**	val	42	165	3	20	14.34	2.15
						test-unseen	**986**	**4008**	**17**	**101**	**50.53**	**8.50**
						test	**1084**	**4008**	**20**	**101**	**46.65**	**9.24**
pers TIR	PerVL DF2	**yes**	fashion	1	1	val	229	229	**50**	**229**	5.58	1.00
						test	221	221	**50**	**221**	4.42	1.00
pers TIG	DreamBooth	**yes**	common	1	0	test	**750**	0	**30**	35	25	**21.43**
	CC101	**yes**	common	**≤ 2**	0	test	**3232**	0	**101**	**597**	**36.36**	**6.15**
ZS-CIR	FashionIQ	no	fashion	1	1	val (avg)	**2005**	**5179**	**1442**	**1994**	1.39	1.00
						test (avg)	**2039**	**5179**	**1454**	**2030**	1.40	1.00
	CIRR	no	open	1	1	val	**4184**	2297	**2165**	**4157**	1.93	1.00
						test	**4148**	2315	**2178**	**4135**	1.90	1.00
	CIRCO	no	open	1	**4.53**	val	**220**	**123403**	**220**	**220**	1	1
						test	**800**	**123403**	**798**	**796**	1.00	1.00

Table 1. Comparison of related datasets. PerVL DF2: PerVL DeepFashion2; CC101: CustomConcepts101. For FashionIQ we report the average number of images per split (shirt, dress, toptee). Dataset aspects that fulfill the criteria discussed in Sec. 3 are in **Bold**. ZS-CIR datasets reported to show the setting difference with personalized TIR/TIG.

egory level. Moreover, the reference image typically contains multiple elements. This aspect poses the additional challenge of understanding to which image element the relative caption refers to (see, e.g., [4]), and whether an image actually represents the same concept or not. Differently, in the considered personalization setting, concepts are typically instances, clearly identifiable in the provided image examples. FashionIQ [37] is another CIR dataset focusing on kinds of fashion items.

Personalized Text-to-Image Generation (TIG). This task was proposed in [6] and is the generative counterpart of personalized TIR. The authors presented Textual Inversion [6], a method that expands the vocabulary of a frozen text-to-image model (Latent Diffusion [26]) with user-specific concept embeddings learned from few images. In [27] the authors present DreamBooth, a method that selects a rarely-used token and binds it to the concept by fine-tuning the text-to-image model on the concept images. Subsequent methods propose improvements over these two [7, 14, 34]. The datasets introduced by these works (see, e.g., DreamBooth and CustomConcepts101 [14, 15]), are constituted by concept training images and a list of prompts for evaluation, but do not contain any real image representing such prompts. The validation is thus carried out by comparing a generated image with the training images (to measure concept fidelity) and with the context in the prompt (to measure context fidelity). Differently, typical validation metrics for generative models measure a distance between the population of real and generated samples (see, e.g., [9, 16, 23]. Since in ConCon-Chi a set of real image realizations is provided for each prompt, we show how the application of these distance measures can improve the validation of personalized TIG methods.

3. Concept-Context Chimera Benchmark

We present ConCon-Chi and compare with existing datasets in Sec. 3.1, then describe its acquisition process in Sec. 3.2.

3.1. Dataset overview

Since personalized TIR/TIG are few-shot tasks, we compare in Tab. 1 with related benchmarks in terms of validation/test splits.

ConCon-Chi test split includes 1084 queries and a pool of 4008 images which are treated as ground-truth images for TIR and image realizations for TIG (∼6 per query). Each query was generated by composing up to 3 concepts and a context from a set of 20 concepts and 101 contexts. The average number of contexts associated to each concept is indicated under Contexts/Concept and gives an idea of the variety of situations in which a concept is required to be retrieved or generated.

In Tab. 1 we also highlight the difference between personalization and ZS-CIR datasets. These latter do not explicitly deal with concepts and contexts, but each reference image and relative caption that compose a query are counted as a different concept and context. Thus their number of concepts and context is higher than in personalized tasks and the ratios Concepts/Context and Contexts/Concept are close to 1. Differently, to evaluate the capability of personalization methods to combine the learned concepts with contexts, Contexts/Concept should be high and, to avoid context bias, Concepts/Context should be high as well.

The most similar dataset to ours is CustomConcepts101, which however does not contain real images for evaluation (see Pool and Images/Query in Tab. 1).

Concepts. In Fig. 2 we show 10 of the 20 concepts in the dataset (the complete set is in supp. material Fig. 10). As

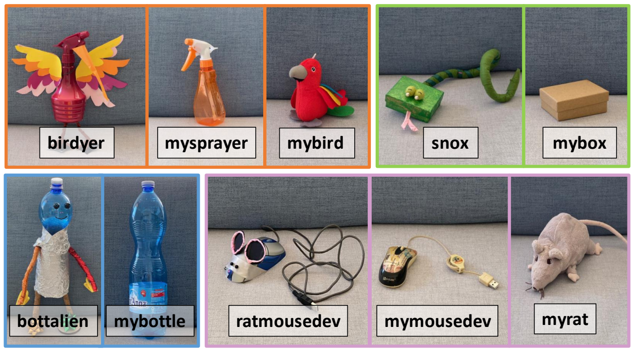

Figure 2. **Example concepts in ConCon-Chi**[1]**.** Four chimeric and six common concepts (hard negatives outlined with same color).

in [27, 34], we include animate and inanimate entities (puppets and tools, accessories, clothes) with the aim of creating rich interactions. There are 6 chimeric concepts, which are animal or alien puppets crafted out of tools or accessories, and 14 common concepts. Among these latter, we include instances of the same categories composing the chimeras, which are thus visually and semantically similar. These are hard negatives especially when appearing in the same context (e.g., BIRDYER and MYSPRAYER in Fig. 1) and enhance the compositionality assessment [21].

Concept names are invented and provided for completeness since are not used by the considered methods. Differently, in our evaluation (e.g., Tab. 2) we consider feeding a pre-trained VL model with a description of the concept as zero-shot baseline, thus we provide the adopted descriptions for reproducibility since words choice was empirical (supp. material Tab. 5). *Discriminative* descriptions were formulated to be a competitive baseline, by thinking to a minimal sentence discriminating the concept from others in the concept set. *Coarse* descriptions are category-level and model the typical case where the concepts to be discriminated from are unknown (e.g., "bag", but there are two bags in the concept set). *Rich* descriptions aim to verify whether enriching the *Discriminative* with visual details helps.

Contexts. Contexts are grouped into 9 kinds (one example per kind in Fig. 1). Inspired by [27], these are structured in concept modifications (accessorizations, property modifications, renditions) and relationships (actions, interactions, re-contextualizations). We describe them and present their distribution in supp. material Fig. 11a.

Similarly to [27, 34] we include general and specific contexts: general contexts are applicable to all or most concepts (e.g. re-contextualization in Fig. 1) and strongly contribute to avoiding context bias; specific contexts are typical of a concept category and aim to increase the difficulty of distinguishing between hard negatives (e.g. the action *containing* for bags-like objects). When a context specifies an interaction between a concept and an entity (object, person, another concept), a retrieval method could perform well simply by detecting their co-occurrence. To avoid this we included contexts specifying different interactions with the same entities (see Fig. 3). A particular interaction is the one between two learned concepts. Concept-concept interactions have been studied in personalized TIG [13, 34]. In Sec. 4 we evaluate this aspect in personalized TIR.

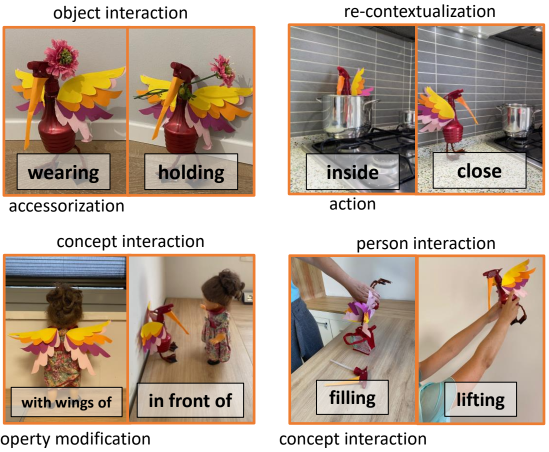

Figure 3. **Fine-grained contexts.** Examples queries for the concept BIRDYER, in which the recognition of the co-occurrence of elements does not suffice for retrieval.

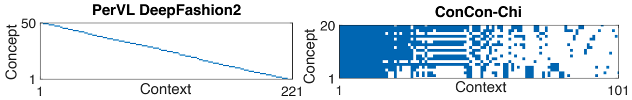

Figure 4. **Concept-context matrices.** Blue cells mark the concept-context combination appearing in the dataset (*test* split).

Concept-context structure. In Fig. 4 we compare the concept-context structure of our dataset with PerVL DF2. A matrix cell is Blue when a query formed by the corresponding context and concept exists in the dataset. While our dataset contains approximately half of the concepts and contexts, our concept-context matrix is denser, reflecting a higher number of queries (1084 vs. 221, see Tab. 1). Specifically, in PerVL DF2 each context is coupled with a single concept, while in our dataset this only occurs for 8 contexts, with 16 contexts combining with every concept. Moreover, in ConCon-Chi each concept is composed with 46.64 contexts on average vs. 4.42 in PerVL DF2 (few cells per row): similarly to context bias, concept bias limits the compositionality assessment since it makes it difficult to determine to what extent the retrieval method understands the context.

[1] Any representation of trademarks, trade names, logos, domain names such as any other distinguishing marks appearing in this dataset is purely random, and it is used exclusively for scientific and non-commercial purposes; therefore, the relevant representation cannot be understood as an expression of an opinion or an indication or a precondition for taking decisions.

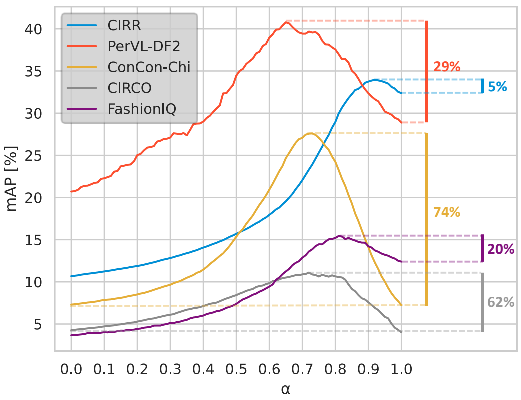

Figure 5. **Concept and context biases.** Retrieval performance by varying the relative weight (α) of the concept and the context in the query. Lateral bars: performance drop when using either only the concept (α=0) or the context (α=1), relative to peak performance.

3.2. Acquisition and Annotation

Acquisition. First we gathered the objects and crafted the chimeric concepts; then we moved to photograph each concept in the designed contexts. We arranged the scene to represent the query and shot multiple pictures by varying the viewpoint. Any private and sensitive information was removed from the scene; images containing humans were cropped to remove the face and make the subject unidentifiable. We provide the distribution of environments in supp. material Fig. 11b and details about the creation of renditions in supp. material Sec. 7.3.

Annotation. Annotations consist of query-image associations. A first step consisted of labelling each image with the represented concepts and context. A second step was carried out to remove false negatives in the annotations. These happen whenever the ground-truth (GT) images of a query that is more specific, are not included among the GT of a query that is more generic (query overlap). A similar problem happens when GT images of a query contain content that also corresponds to other queries (image overlap). In benchmarks created by re-annotating existing datasets of images “in the wild” it is not possible to control overlaps and an exhaustive check is unfeasible. To this end, CIRR [20] and CIRCO [2] designed approximated procedures to ensure that false negatives are respectively absent in sub-pools of images or below an estimated percentage. We accounted for the problem since the dataset design by controlling and minimizing overlaps. The GT images were then assigned following concept and context overlap and manually checking those of queries where we were aware of possible image overlap. As a result, to the best of our knowledge, the set of 1084 queries is free from false negatives with respect to the pool of 4008 images. We report the number of GT images per query in supp. material Fig. 12.

4. Personalized Text-to-Image Retrieval

We define the TIR benchmark task on ConCon-Chi and compare it with related benchmarks in Sec. 4.1. We present and analyse results respectively in Sec. 4.2 and Sec. 4.3.

4.1. Benchmark Task

Train and test splits. Each concept in the set of 20 is trained independently on 1 to 5 images where the concept is standing in front of some background (*train* split, examples in Fig. 2). At test time, for each query in the *test* split we rank the images in the pool according to their similarity with the query. Training backgrounds do not appear in the *test* split. Since the concepts are not trained jointly, similar concepts play the role of hard negatives. In supp. material Sec. 7.4 we introduce other splits not used in this paper.

Metrics. For each query we evaluate the rank of the first GT image (mean Reciprocal Rank, mRR) and whether this is among the top-k (recall rate, R@k); then we evaluate the rank of all GT images (mean Average Precision [3], mAP) and of GT images up to the top-k, mAP@k [2]. See supp. material Sec. 8.1 for definitions.

Concept and context biases. We first aim to compare the presented benchmark with existing ones in terms of context and concept biases. To quantify the importance of attending the concept and the context for correct retrieval, we adopt the experimental setup proposed in [30] and model the concept-context interaction by computing a query embedding q as weighted sum of c, the CLIP [25] embedding of the context (output of CLIP text encoder, ViT-L14 backbone), and i, the average CLIP embedding of the concept training images (output of the vision encoder): $q = (1-\alpha)\cdot i + \alpha\cdot c$ with $\alpha \in [0, 1]$. In Fig. 5 we compare the retrieval performance by varying α on ConCon-Chi, PerVL-DF2, CIRCO, CIRR and FashionIQ. For PerVL DF2 and ConCon-Chi we consider the available training images and for each dataset we use the *test* split if available, the *validation* otherwise (see Sec. 8.2 and Tab. 1 for details). When in ConCon-Chi multiple concepts appear in the query we average their embeddings.

For three datasets the performance achieved by just putting the context in the query (α=1) is higher than 70% of the peak performance, indicating that the probability of retrieving the correct context-concept combination is relatively high even when ignoring the concept (context bias). Differently, in ConCon-Chi we observe low performance for α=1, and a drop of 74% from the peak performance with α=0 (only the concept in the query). Only CIRCO exhibits a similar trend, with lower peak performance possibly because of the intrinsic difficulty of the proposed ZS-CIR task and the large image pool.

	Method	mAP [%]	mRR [%]	R@1 [%]
k=0	*Coarse*	16.83	24.21	14.48
	Discriminative	*30.16*	*43.16*	*31.92*
	Rich	*27.65*	*40.58*	*29.98*
k=1	PALAVRA	22.56 ± 1.29	34.39 ± 1.68	24.59 ± 1.94
	Pic2Word	25.23 ± 1.20	37.16 ± 1.76	26.35 ± 1.85
	SEARLE	28.16 ± 0.55	41.07 ± 0.92	31.16 ± 0.94
k=5	PALAVRA	23.59	35.99	26.75
	Pic2Word	26.39	38.62	27.68
	SEARLE	**30.74**	**43.83**	**33.49**

Table 2. **Personalized TIR benchmark.** Performance of CLIP baselines (k=0) and personalized TIR (PALAVRA) and ZS-CIR methods (Pic2Word, SEARLE) on the retrieval task in ConCon-Chi. The best method per metric is in **Bold**; *Discriminative* and *Rich* baselines (oracles) are highlighted with *asterisks*.

4.2. Benchmark Results

Methods. We compare the personalized TIR method PALAVRA [5] with two ZS-CIR methods, Pic2Word [30] and SEARLE [2], on the defined TIR benchmark. Since, differently from ZS-CIR, in our setting multiple concept example images are available, for the two latter methods we average the generated token embeddings to create the concept embedding. All methods rely on CLIP ViT-L14 backbone with same input pre-processing. The methods learn a textual token embedding that expands CLIP vocabulary and assign it an arbitrary textual identifier such that at inference they replace the learned embedding whenever this is encountered in an input query. We used the code released by the authors (see supp. material Sec. 8.3 for details).

Results. In Tab. 2 we evaluate each method with k=1 and k=5 training images per concept (for k=1 we report mean and standard deviation over the 5 images). We also report, as baselines that do not use any image (k=0), the performance achieved by feeding CLIP with queries where the concept identifier is replaced with the *Coarse*, *Discriminative* or *Rich* descriptions introduced in Sec. 3.1.

As expected, *Coarse* descriptions provide a lower bound, since they are shared among the concepts of same category to simulate descriptions that are not tailored to the discriminative task at hand. Differently, *Discriminative* and *Rich* descriptions provide a competitive baseline (higher than PALAVRA and Pic2Word). They represent an oracle, since were formulated by discriminating every concept from the others, thus accessing information which is unavailable to other methods that learn each concept independently. In this respect, we note that adding visual details degrades performance. Interestingly, SEARLE outperforms the *Discriminative* baseline thus being the best method, also because it exhibits a smaller standard deviation in the 1-shot scenario and a larger gain when more images are provided. This confirms that SEARLE outperforms the other two methods not only in ZS-CIR [2] but also in the personalized TIR setting. We report more metrics in supp. material Tab. 7.

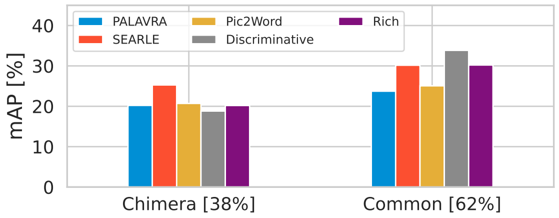

Figure 6. **New-meaning learning.** Retrieval performance on queries containing at least one chimeric concept vs. common concepts (percentage of each query kind in the dataset in brackets).

Method	All Concepts	Chimeric Concepts	Common Concepts	Contexts
Coarse	29.38	09.51	37.89	42.11
Discriminative	*66.48*	*60.61*	*68.99*	40.92
Rich	*84.54*	*85.54*	*84.11*	29.81
PALAVRA	**91.52**	**91.15**	**91.67**	28.56
Pic2Word	50.04	51.32	49.49	**50.12**
SEARLE	76.58	82.25	74.15	40.47

Table 3. **Concept-context compositionality.** Performance (F1 score [%]) of recognition of concepts and contexts in the retrieval task of Tab. 2.

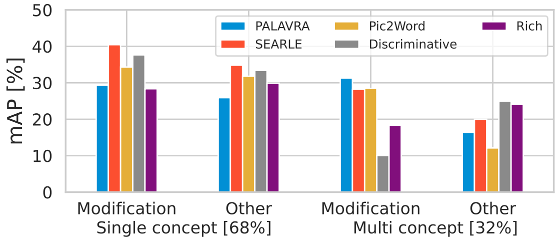

Figure 7. **Concept-concept compositionality.** Performance on single- and multi-concept queries, comparing "accessorization" and "property modification" ("modification") kinds versus the rest of the kinds ("other").

4.3. Analysis of Results

We analyse the performance achieved in the benchmark in terms of new-meaning learning and compositionality. For this study we consider the methods trained on 5 images.

New-meaning learning. In Fig. 6 we report the mAP of Tab. 2 separately on queries containing at least one chimeric concept and common ones. We see that all methods and baselines achieve a lower performance when retrieving queries containing a chimeric concept. We then inspected whether this can be explained with a misclassification of the concept or the context, or their combination. For each query therefore we considered the first N ranked images (N equal to the number of GT images) and marked the concept in the query as ground-truth and the concept in each retrieved image as a prediction. We computed the F1 score for each concept by accumulating the

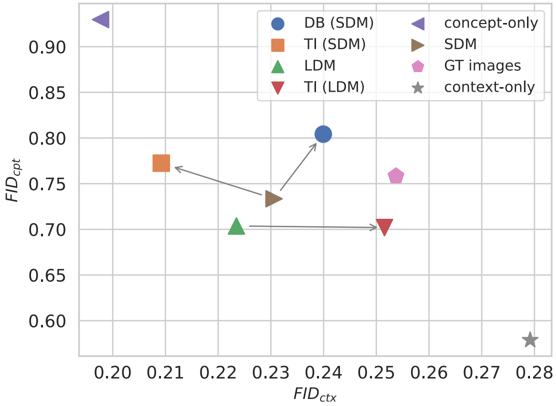

Figure 8. **Personalized TIG benchmark.** Scatter plot of the FID_{cpt}/FID_{ctx} trade-off for the considered methods, baselines and upper bounds on the generation task ConCon-Chi.

		k=3		k=10			
		Density ↑	Coverage ↑	Density ↑	Coverage ↑	FID_{cpt}	FID_{ctx}
SDM	Common	2.66	0.12	8.84	1.33	0.74	0.23
	Chimeric	0.78	0.02	2.40	0.21	0.73	0.23
TI	Common	3.43	0.23	11.56	**2.93**	0.71	**0.25**
	Chimeric	2.71	0.15	5.43	0.90	0.69	**0.25**
DB	Common	**6.55**	**0.28**	**18.98**	2.65	**0.80**	0.24
	Chimeric	4.81	0.14	11.49	1.09	0.79	**0.25**

Table 4. **Comparing generated and GT images.** We consider TI and DB with SDM. The best for each metric is in **Bold**. Coverage and Density are in [%] (Density not upper bounded by 100).

predictions from the queries containing that concept and we report in Tab. 3 the average scores over all, chimeric and common concepts. Please note that we restricted this analysis to single-concept queries and discarded ranked images containing multiple concepts.

In Tab. 3 we observe that personalization methods achieve equal or better classification performance for chimeric with respect to common concepts. We then conclude that, while these methods learn to recognize the visual appearance of the chimeras, they struggle more at retrieving these new meanings in composition with known contexts.

Concept-context compositionality. As for the concepts, we computed the classification performance in a similar way for the contexts, and report it in Tab. 3. We see that PALAVRA achieves the best concept and the lowest context classification performance. We deduce that its low retrieval performance in the benchmark must be due to the poor compositionality properties of the learned tokens. Conversely, Pic2Word excels at recognizing the contexts, but cannot learn discriminative tokens. Thus, the good retrieval performance of the *Discriminative* baseline and of SEARLE in Tab. 2 seems explained by their capability to trade-off and combine context and concept recognition. This seems confirmed by the fact that the *Rich* descriptions improve concept recognition at the expense of the context, probably since longer descriptions out-weight the rest of the query.

Concept-concept compositionality. We finally investigated whether performance also depends on the context kind. In Fig. 7 we report the mAP of Tab. 2, separately on queries containing a single and multiple concepts and, in each group, containing a concept modification ("property modification" or "accessorization") or another kind of context. As expected, multi-concept queries are more challenging. On these we observe an interesting trend: description-based baselines drop performance on concept modifications, while the personalization methods drop performance on the other kinds. Multi-concept queries with concept modifications are mostly the ones where a part of a concept is worn or applied to another one (an example in Fig. 3 where the wings of BIRDYER are applied to MYDOLL). These queries require knowing visual information on the concept parts, thus explaining why the baselines fail in these cases. Conversely, the other kinds of multi-concept queries typically require the detection of the co-occurrence of the concepts and their relationship (e.g., in Fig. 3, "MYDOLL in front of BIRDYER"). To achieve this, the concept textual representations must exhibit robust compositionality properties. Thus, the tokens learned by the methods do not seem to retain the same compositionality properties of the tokens in the original vocabulary (which form the descriptions).

5. Personalized Text-to-Image Generation

We define the TIG benchmark task and present results in Sec. 5.1. We analyse them in more detail in Sec. 5.2.

5.1. Benchmark Task and Results

Train and test splits. We train each concept on the same 5 images used for TIR and at test time we generate 4 images per prompt. We restrict the evaluation to the single-concept prompts in the *test* split (which are 735).

Metrics. We report the two metrics introduced in [6] and adopted by following works. These rely on CLIP and compute, for each prompt, the average pairwise cosine similarity between the visual embeddings of the generated images and of, respectively, the visual embeddings of the training images (fidelity-concept or FID_{cpt} as in [27]) and the textual embedding of the context in the prompt (fidelity-context or FID_{ctx}). Personalized TIG methods are expected to trade-off the two metrics by learning to represent the concept appearance while retaining the capability to represent contexts.

Methods. As in personalized TIR, the considered methods assign a textual identifier to the concept and when this is encountered in the prompt they load a corresponding learned token embedding (Textual Inversion, TI [6]) or model (DreamBooth, DB [27]). For DB we adopted the Diffusers library [36] with a Stable Diffusion Model (SDM) [26] as pre-trained text-to-image model (called DB(SDM)). For TI we adopted the code by the authors with either a Latent Diffusion Model (LDM) [26] or, by following [34], the same SDM as for DB (called TI(LDM) and TI(SDM)). Details in supp. material Sec. 9.2.

Results. In Fig. 8 we show the fidelity-concept/context trade-off (numbers in supp. material Tab. 8). To quantify these metrics, as in [6] we compare with upper bounds. For the fidelity-concept, the upper bound returns always the 5 training images irrespective of the prompt (concept-only). For the fidelity-context, the upper bound are images generated by the pre-trained SDM when fed as input only with the context of the prompt (context-only). We also evaluate the baseline performance of the LDM or SDM pre-trained models by replacing the concept identifier in the prompt with its *Rich* description. As reference, we report also the metrics for the GT images available in ConCon-Chi.

We first observe that TI(LDM) does not improve the fidelity-concept over LDM, but improves the fidelity-context. This can be explained by considering that the *Rich* description in LDM tends to out-weight the context. Conversely, TI(SDM) improves the fidelity-concept of SDM, while degrading the fidelity-context. This behaviour has also been observed in recent work [34] and may be due to the learned token over-fitting the context of the training images. A better learning behaviour is provided by DB(SDM), which improves both metrics over SDM. Notably, compared with the GT images, DB shows a higher fidelity-concept and lower fidelity-context.

5.2. Analysis of Results

We leverage the availability of GT images to inspect the performance of DB, TI and their pre-trained model SDM with two metrics proposed for evaluation of text-to-image generation [23]: Density as measure of realism (fidelity) and Coverage as measure of diversity (how well generated images span the real images manifold). We created one real and one generated manifold per concept as the union of all prompts for that concept (similarly to [9] Appendix L). In Tab. 4 we report the metrics for two values of the parameter k, averaged over chimeric and common concepts. For all methods, Density and Coverage are higher on common concepts than on chimeric ones. This is in line with what has been observed in [27, 34] that more common objects are easier to generate since these methods leverage the pre-trained knowledge about the concept category. Interestingly, such difference is not evident from the $\text{FID}_{\text{cpt/ctx}}$ metrics typically adopted in personalised TIG. We observe that DB has best Density, while in terms of Coverage there is a smaller gap with TI. Thus DB achieves relatively better realism, but diversity remains low. We conclude that leveraging the availability of a real population of images allows for a more thorough evaluation of personalised TIG: there seems to be large room for improvement, with new-meaning concepts posing more challenges.

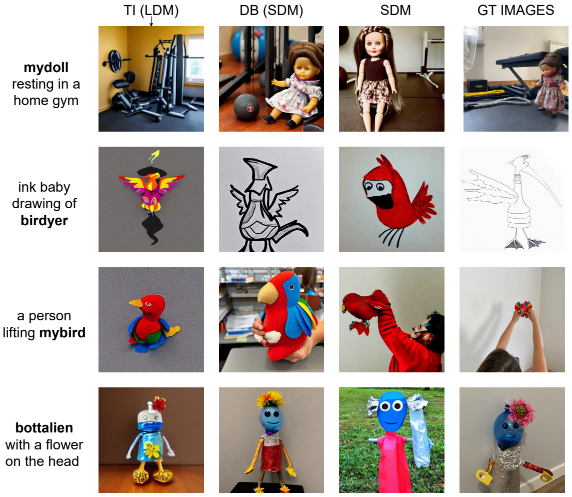
TI (LDM)
DB (SDM)
SDM
GT IMAGES
mydoll resting in a home gym
ink baby drawing of birdyer
a person lifting mybird
bottalien with a flower on the head

Figure 9. **Qualitative results for personalized TIG.**

Qualitative examples. We report some cherry-picked examples in Fig. 9. We observe that DB represents the concepts more accurately and combines them more nicely with the context than TI, which sometimes forgets the concept (First Row) and sometimes the context (Third Row). It can also be noticed how common concepts are learned very accurately by DB (First and Third Row), while chimeric ones are not. As expected, using a visually *Rich* description (Third Column) is not enough for personalized generation, especially on chimeric concepts (Second and Fourth Row), which lack more distinctive features than the common ones.

6. Conclusion

We present a new dataset called ConCon-Chi for the evaluation of personalized TIR and TIG. The dataset models novel concepts as chimeric objects and by adopting a concept-context matrix structure allows to study the learning of new meanings in terms of their compositionality properties with known ones. Our analysis on retrieval showed that current methods struggle at composing these new concepts with known contexts, and also together. A similar issue was observed when evaluating the generated images with respect to real examples by adopting image generation metrics. We hope that this study and the dataset released will help improving current personalized VL methods.

Acknowledgments

We express our sincere gratitude to the members of Sony Interactive Entertainment Inc. teams for their invaluable support and insightful feedback related to our project.

References

[1] Dzmitry Bahdanau, Tom Bosc, Stanisław Jastrzębski, Edward Grefenstette, Pascal Vincent, and Yoshua Bengio. Learning to Compute Word Embeddings On the Fly. *arXiv preprint arXiv:1706.00286*, 2017. 2

[2] Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-Shot Composed Image Retrieval with Textual Inversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15338–15347, 2023. 2, 5, 6

[3] Steven M. Beitzel, Eric C. Jensen, and Ophir Frieder. *MAP*, pages 1691–1692. Springer US, Boston, MA, 2009. 5

[4] Wentao Chen, Chenyang Si, Zhang Zhang, Liang Wang, Zilei Wang, and Tieniu Tan. Semantic Prompt for Few-Shot Image Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 23581–23591, 2023. 3

[5] Niv Cohen, Rinon Gal, Eli A. Meirom, Gal Chechik, and Yuval Atzmon. "This Is My Unicorn, Fluffy": Personalizing Frozen Vision-Language Representations". In *Computer Vision – ECCV 2022*, pages 558–577, Cham, 2022. Springer Nature Switzerland. 2, 6

[6] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 3, 7, 8

[7] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. Encoder-based Domain Tuning for Fast Personalization of Text-to-Image Models. *ACM Trans. Graph.*, 42(4), 2023. 3

[8] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5337–5345, 2019. 2

[9] Jiyeon Han, Hwanil Choi, Yunjey Choi, Junho Kim, Jung-Woo Ha, and Jaesik Choi. Rarity Score: A New Metric to Evaluate the Uncommonness of Synthesized Images. In *The Eleventh International Conference on Learning Representations*, 2023. 3, 8

[10] Aurélie Herbelot and Marco Baroni. High-risk learning: acquiring new word vectors from tiny data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 304–309, Copenhagen, Denmark, 2017. Association for Computational Linguistics. 2

[11] Felix Hill, Kyunghyun Cho, Anna Korhonen, and Yoshua Bengio. Learning to Understand Phrases by Embedding the Dictionary. *Transactions of the Association for Computational Linguistics*, 4:17–30, 2016. 2

[12] Bruno Korbar and Andrew Zisserman. Personalised CLIP or: how to find your vacation videos. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*. BMVA Press, 2022. 2

[13] Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution. In *International Conference on Learning Representations*, 2022. 4

[14] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-Concept Customization of Text-to-Image Diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1931–1941, 2023. 3

[15] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-Concept Customization of Text-to-Image Diffusion. *arXiv preprint arXiv:2212.04488*, 2023. 3

[16] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved Precision and Recall Metric for Assessing Generative Models. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2019. 3

[17] Angeliki Lazaridou, Marco Marelli, and Marco Baroni. Multimodal Word Meaning Induction From Minimal Exposure to Natural Text. *Cognitive Science*, 41(S4):677–705, 2017. 2

[18] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *Proceedings of the 39th International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 2

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 2, 6

[20] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image Retrieval on Real-life Images with Pre-trained Vision-and-Language Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2125–2134, 2021. 2, 5

[21] Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. CREPE: Can Vision-Language Foundation Models Reason Compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10910–10921, 2023. 4

[22] Sachit Menon and Carl Vondrick. Visual Classification via Description from Large Language Models. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[23] Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. Reliable Fidelity and Diversity Metrics for Generative Models. In *Proceedings of the 37th International Conference on Machine Learning*, pages 7176–7185. PMLR, 2020. 3, 8, 7

[24] Sarah Parisot, Yongxin Yang, and Steven McDonagh. Learning To Name Classes for Vision and Language Models. In

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 23477–23486, 2023. 2

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 5

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 2, 3, 8

[27] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510, 2023. 3, 4, 7, 8

[28] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115:211–252, 2015. 6

[29] Elena Sofia Ruzzetti, Leonardo Ranaldi, Michele Mastromattei, Francesca Fallucchi, Noemi Scarpato, and Fabio Massimo Zanzotto. Lacking the Embedding of a Word? Look it up into a Traditional Dictionary. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2651–2662, Dublin, Ireland, 2022. Association for Computational Linguistics. 2

[30] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2Word: Mapping pictures to words for zero-shot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19305–19314, 2023. 2, 5, 6

[31] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, 2018. Association for Computational Linguistics. 6

[32] Otilia Stretcu, Edward Vendrow, Kenji Hata, Krishnamurthy Viswanathan, Vittorio Ferrari, Sasan Tavakkol, Wenlei Zhou, Aditya Avinash, Emming Luo, Neil Gordon Alldrin, MohammadHossein Bateni, Gabriel Berger, Andrew Bunner, Chun-Ta Lu, Javier Rey, Giulia DeSalvo, Ranjay Krishna, and Ariel Fuxman. Agile Modeling: From Concept to Classifier in Minutes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22323–22334, 2023. 2

[33] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A Corpus for Reasoning about Natural Language Grounded in Photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6418–6428, Florence, Italy, 2019. Association for Computational Linguistics. 2

[34] Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-Locked Rank One Editing for Text-to-Image Personalization. In *ACM SIGGRAPH 2023 Conference Proceedings*, New York, NY, USA, 2023. Association for Computing Machinery. 3, 4, 8, 9

[35] Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. Composing Text and Image for Image Retrieval - an Empirical Odyssey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6439–6448, 2019. 2

[36] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022. 8

[37] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. Fashion IQ: A New Dataset Towards Retrieving Images by Natural Language Feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11307–11317, 2021. 3

[38] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. YouTube-VOS: Sequence-to-Sequence Video Object Segmentation. In *Computer Vision – ECCV 2018*, pages 603–619, Cham, 2018. Springer International Publishing. 2

[39] Chun-Hsiao Yeh, Bryan Russell, Josef Sivic, Fabian Caba Heilbron, and Simon Jenni. Meta-Personalizing Vision-Language Models To Find Named Instances in Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19123–19132, 2023. 2

[40] Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep Sets. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 6