2023 IEEE/CVF International Conference on Computer Vision (ICCV) | 979-8-3503-0718-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCV51070.2023.01844

Going Beyond Nouns With Vision & Language Models Using Synthetic Data

Paola Cascante-Bonilla*†[1,2] Khaled Shehada*[2,3] James Seale Smith[2,4] Sivan Doveh[6,7]
Donghyun Kim[2,7] Rameswar Panda[2,7] Gül Varol[5] Aude Oliva[2,3]
Vicente Ordonez[1] Rogerio Feris[2,7] Leonid Karlinsky[2,7]
[1]Rice University [2]MIT-IBM Watson AI Lab [3]MIT [4]Georgia Institute of Technology

Large-scale pre-traine
els have shown remarkab
tions, enabling replacing
with zero-shot open vocab
bitrary) natural language
have uncovered a fundan
For example, their difficul
Concepts (VLC) that go '
ing of non-object words (
states, etc.), or difficulty
soning such as understan
der of the words in a ser
gate to which extent pure
aged to teach these model
without compromising the
tribute ***Sy****nthetic* ***Vi****sual* ***C***
synthetic dataset and data
generate additional suita
standing and composition
tionally, we propose a ge

provements. Our extensi
VL-Checklist, Winogroun
strate that it is possible to adapt strong pre-trained VL models with synthetic data significantly enhancing their VLC understanding (e.g. by 9.9% on ARO and 4.3% on VL-Checklist) with under 1% *drop in their zero-shot accuracy.*

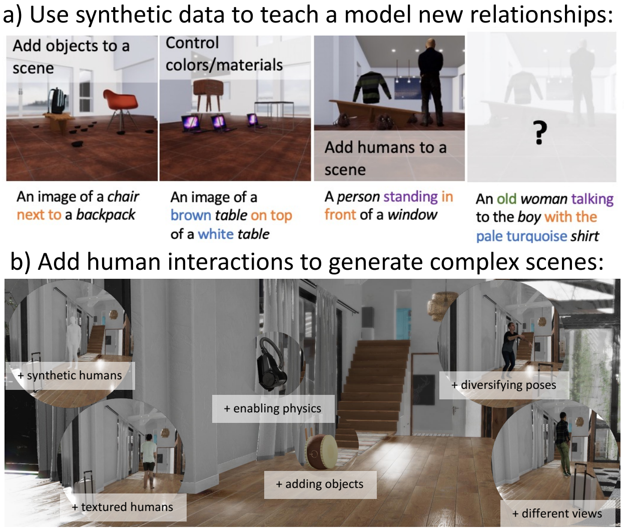
a) Use synthetic data to teach a model new relationships:
Add objects to a scene
Control colors/materials
Add humans to a scene
?
An image of a chair next to a backpack
An image of a brown table on top of a white table
A person standing in front of a window
An old woman talking to the boy with the pale turquoise shirt
b) Add human interactions to generate complex scenes:
+ synthetic humans
+ enabling physics
+ diversifying poses
+ adding objects
+ textured humans
+ different views

Figure 1. Overview of our proposed synthetic set: we place different objects in a scene and change their position, color, size, and material. We further emphasize on human-level interactions, sampling a wide set of body poses and behaviors to cover transitive and intransitive human actions.

1. Introduction

There have been impressive advances in the performance of zero-shot recognition through the use of large-scale pre-trained Vision & Language (VL) models [45, 20, 50, 28, 15, 57, 29, 13]. However, these VL models still face some important challenges in understanding Visual Language Concepts (VLC) beyond object nouns (e.g., recognizing attributes, relations, states) and in terms of compositional reasoning capabilities (i.e.., understanding subtle changes in meaning due to small changes in word order). Recently, several benchmark tests have been devised to demonstrate the extent to which these models lack these capabilities [51, 62, 59] [1]. As noted in several recent works [59, 62, 9], this behavior of VL models is likely due to the contrastive pre-training prevalent for all of them and likely inducing 'bag-of-objects' kind of representations (for both images and text alike). Indeed, for (even large) random batches of paired image-text samples, the collection of objects (nouns) in the image (or text) is likely to uniquely determine the image (or text) in the batch, making contrastive batch losses focus on the objects (nouns) while regarding

*Equal contribution. Project page: https://synthetic-vic.github.io/

†Work partially done while interning at the MIT-IBM Watson AI Lab.

[1] Please also see supplementary material for the expanded set of results of [62] including results for all the most recent open-sourced VL models, all exhibiting poor VLC understanding performance.

2380-7504/23/$31.00 ©2023 IEEE
DOI 10.1109/ICCV51070.2023.01844

other details (attributes, relations, states, word order, etc.) as unnecessary. Intuitively, this impairs VLC understanding and compositional reasoning of the resulting model.

Given the above, a natural question to ask is what is the most effective way to 'fix' the VL models to improve their VLC understanding and compositional reasoning performance? An approach proposed in concurrent works [9, 59], advocates for the use of text augmentation, using language tools to teach a model the importance of non-noun words by manipulating them (e.g., replacing them with incorrect alternatives) and adding the resulting texts to the same batch. Although effective, such augmentation techniques are only easy on the text side and are much harder and prohibitively expensive on the image side. Indeed, finding, collecting, or generating real image samples sharing the objects but differing in their composition, attributes, relations, or states is very difficult. Although significant progress has been achieved with text-based editing [16, 21, 37, 5, 23, 3, 38], these methods are relatively slow (leveraging diffusion) and not sufficiently stable to allow effective use for augmentation in training pipelines. In this work, therefore, we propose an orthogonal route – VL data synthesis for fixing VL models by targeted demonstration. Specifically, we propose enhancing the VLC and compositionality aspects of the generated *visual and text* data, in turn using this data for finetuning VL models teaching them to pay closer attention to these aspects. Moreover, besides being largely free and infinitely scalable, synthetic data has an additional advantage – it can also be free from privacy concerns always accompanying real data.

Besides the inherent challenges of realistic data simulation, building synthetic data that can be effectively used to improve VLC and compositionality aspects of VL models pre-trained on massive real data poses additional technical challenges. Unlike the majority of prior work focusing on synthetic visual data generation, we need not only to generate images, but also the text that describes compositional items in a scene. We generate synthetic videos that leverage realistic physical 3D simulation [11] including diverse 3D environments and different 3D objects, human motions, and actions assets [1, 35, 41, 44], added interaction with objects, and different camera viewpoints. Every frame of these videos is accompanied by rich metadata, allowing using language grammar for generating detailed descriptive captions of any instantaneous scene in each video. These captions, in turn, allow collecting diverse image-text pairs samples contrasting which one to another highlights to the model the importance of the compositional items in the text captions (e.g. different viewpoints or different frames in the same video share objects but may strongly differ in the VLC and other compositional items). While motion assets were used by previous works to generate synthetic data [55, 54], the visual data was not accompanied by textual captions and was not designed with the need to highlight compositionality in mind. We contribute **Sy**nthetic **Vi**sual **C**oncepts (**SyViC**) – a large (million-scale) generated synthetic VL dataset with rich textual captions, easily extensible through our data synthesis code together with all the already generated million-scale synthetic data used in this paper (Figure 1).

In addition to the data synthesis pipeline, we also offer a strategy for effectively leveraging the generated synthetic data, while avoiding forgetting real data alignment and losing the strong a-priori zero-shot capabilities of the model. We propose and extensively ablate a combination of domain adaptation by stylization [63], parameter efficient fine-tuning [17], long captions handling, and model averaging methods [56] to reduce forgetting, as well as examine the effect of different aspects of data synthesis and finetuning choices on the gains in VLC and compositionality understanding.

Our contributions can be summarized as follows: (i) we contribute **SyViC** – a million-scale synthetic dataset with rich textual captions, intended for improving VLC understanding and compositional reasoning in VL models, as well as the methodology and the generation codebase [2] for its synthesis and potentially extensibility; (ii) an effective general VL model finetuning strategy enabling effective leveraging of **SyViC** data for enhancing the aforementioned aspects of strong pre-trained VL models without sacrificing their zero-shot capabilities; (iii) experimental results and extensive ablation study showing significant (over 10% in some cases) improvement in VLC understanding and compositional reasoning respectively, measured on all the recent VL-Checklist, ARO, and Winoground benchmarks and validated on the most popular CLIP [45] model and its derivatives (e.g. the most recent CyCLIP [15]).

For supplemental materials, readers are referred to the associated arXiv document at [arXiv:2303.17590].

2. Related Work

Large-scale Vision&Language (VL) Models: Large-scale pre-trained VL models such as CLIP [45] or ALIGN [20] show remarkable success in many zero-shot recognition tasks such as image classification or detection [60]. Despite the continued advancements made in this direction [15, 57, 29, 13], recent studies (*e.g.*, [62, 51, 59]) show that existing VL models exhibit limited comprehension of structured vision language concepts (VLC). Yuksekgonul *et al.* [59] argue that contrastive learning for image-retrieval learns shortcuts and does not learn compositional information. To address this limitation, some approaches investigate how to augment the text captions or images in contrastive learning to enhance the ability of VLC [9, 59]. Smith *et al.* [48] learn VLC concepts with additional supervised datasets in a continual learning setup. In contrast, we use 3D graphic engines to generate realistic synthetic videos with different compositions and generate corresponding text captions,

[2] We release our code together with all million-scale synthetic data used in this paper here: https://github.com/uvavision/SyViC

which allows a VL model to learn compositionality and non-object words such as attributes, actions, relations, etc.

Learning from Synthetic Data. There has been a lot of work on learning from synthetic data in image classification [11, 40, 36], semantic segmentation [47, 46], human pose estimation [55, 22], action recognition [54], etc. Synthetic data is easy to generate and particularly useful for providing dense annotation such as semantic segmentation and depth estimation since these are prohibitively expensive to annotate manually. Some of the work relies on graphics engines to generate realistic data. Mishra *et al*. [36] propose a method to learn how to generate task-adaptive synthetic data with the 3D simulation engine. For human-related problems, parametric body models (*e.g*., SMPL [32]) can be leveraged, along with motion assets [53], to generate synthetic human videos for low-level body analysis tasks [55] or action recognition [7, 54]. Similar to our work, [7, 54] seek to associate semantic labels to synthetic images, but different from symbolic action categories, our focus is to assign rich textual descriptions to our generated images.

Since synthetic data suffers from a domain gap such as textures, visual styles, or colors from real images, domain adaptation, and generalization have been proposed to address this issue. Adversarial learning [52, 12, 49] can be used to generate real-like images or feature alignment between synthetic and real data. Additionally, stylization methods [64, 63] are proposed as a style augmentation to make a model robust to diverse styles. In contrast, we manually randomize the visual content including different 3D objects, materials, and color attributes in graphics engines. Then we generate realistic synthetic videos from different domains with corresponding text captions. The generated data can be served as a hard negative augmentation and enhance the ability of VLC.

3. Method

We first present our synthetic data generation pipeline (Sec. 3.1), then describe how we leverage it for significant gains in VLC understanding and compositional reasoning capabilities of strong pre-trained VL models (Sec. 3.2). Our entire approach is illustrated in detail in Fig. 2.

3.1. Synthetic Data Generation

In this section, we outline the components and the pipeline of our approach used to generate the proposed **Sy**nthetic **Vi**sual **C**oncepts (**SyViC**) synthetic VL dataset for improving VLC understanding and compositional reasoning of VL models. Our contributed dataset includes 767,738 image-text pairs, 598K sampled from 1,680 diverse synthetic videos, and the remaining 169K generated as individual static synthetic scenes. Example samples from **SyViC** are provided in Supplementary.

3D physics-based simulation platform: ThreeDWorld (TDW) [11], which is built on top of Unity3D, is a multi-modal simulation platform that enables realistic physical interactions. TDW contains 2304 objects, 585 unique materials subdivided in metal, cardboard, wood, ceramic, and glass, and over 30 indoor and 9 outdoor scenes (3D environments). For generating synthetic VL data, we start with placing random objects in a scene following the workflow proposed by [6]. We also use their camera positions and configurations to place objects visible inside good empty room perspectives. We group the available 3D object models by assigning dimension-related labels to each object and use the ImageNet category labels available for each object model as its associated text for later caption synthesis.

Camera Viewpoints: To further augment the set of plausible object placements and relations, we simultaneously place 4 to 12 cameras around a specific point of an empty room, and randomly place $n \geq 1$ objects in the scene, allowing us to render images from different views of the same scene further strengthening the compositional aspects of the data as discussed in the introduction (Sec. 1). For each scene (frame), TDW cameras are able to capture RGB images, the corresponding object instance and category semantic segmentation masks, and a depth map. We use these, as well as a range of sensor and physics data representing the state of the world returned by TDW's API, to enable dense annotations and supervision for each scene (frame) as part of our metadata generation process. We collect all of this information in our metadata and use it to estimate the position of the objects in the scene instead of relying on the 3D coordinates of each object and the camera position.

Digital humans: As we focus on compositionality aspects of images and text pairs, having people in our images is important. However, people models (especially animatable ones) are usually not present in common collections of 3D assets. Existing large-scale synthetic datasets often focus on realistically placing objects in a scene, but typically humans and animals are not included. We first inspected what libraries were available for realistic human synthesis. PeopleSansPeople [10], a library with 28 human 3D models and 39 unique actions, allows only random human placement, not allowing for humanoid customization or integration of human-object interactions. We leverage TDW support for Skinned Multi-Person Linear Model [33] (SMPL) humanoids. SMPL is a parametric body model that enables a realistic representation of the shape and pose of arbitrary (non-clothed) 3D human bodies with diverse genders and shapes. SMPL models can be easily animated using motion capture data. The pose of the SMPL model is defined by a set of joint angles that determine the position of the corresponding body parts in the 3D space. The extended SMPL-X [39] additionally allows controlling hand articulation and face expressions. Given the available library asset in TDW that enables placing these SMPLs in a scene, we create a stand-alone module to automatically incorporate arbitrary custom animations and 514 unique human textures from the SURREAL [55] and Multi-Garment [4] datasets for clothing the synthetic human models for further enhancing the diversity and compositional features of our data.

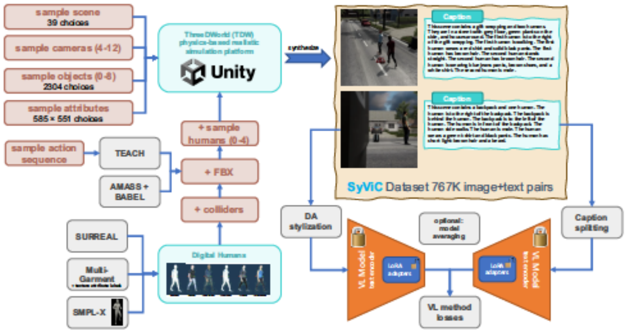

Figure 2. Summarizing the entire flow of the proposed approach including components and choices of SyViC data synthesis pipeline (**left**) and proposed effective finetuning technique (**right**) attaining significant gains detailed below.

Human motion synthesis and handling interactions: In Unity, SMPLs have skeletons and can be driven by motion-capture animations but they don't have mass or colliders, this means that they are not physics assets since they can walk through other objects without interacting with them. To solve this issue, we add colliders to each body part of the SMPL model and create three asset templates (i.e., male, female, neutral) that contain all mesh configurations. Figure 2 shows some examples of our rigged asset templates with colliders. All other existing 3D models in TDW have colliders and track collisions at runtime through TDW physics-based simulation. Therefore, our collider-enhanced SMPLs are pulled downward by gravity, as well as simulate interaction by reacting naturally to collisions with other objects in the scene during motion simulation. For human actions, we first synthesize a diverse set of human actions from random language descriptions using TEACH [2], a Transformer-based model that generates a continuous sequence of SMPL motions from a sequence of action text labels. TEACH was trained on AMASS [35], a large-scale motion-capture (mocap) collection, and BABEL [44], a dataset that provides textual descriptions for AMASS, including per-frame unique actions annotations. Second, we extend our set of SMPL motions by directly sampling unique human motions from BABEL and AMASS. We export the corresponding mocaps to FBX files, and extract the animations in Unity, enabling them as asset bundles for use with TDW. FBX is a common format that facilitates data exchange between different 3D simulation platforms.

Domain randomization: One of the key qualities of the generated synthetic data, is its ability to highlight the importance of VLCs and compositional properties of the scene (e.g., objects attributes, relations, and more) in the contrastive learning objectives guiding the VL model finetuning. As opposed to methods based on text augmentation [9, 59] that can only enhance those on the text part of the VL image-text pairs, in SyViC construction we can easily manipulate also the visual content. We randomly place 1 to 8 3D object models in the scene, randomizing their material and color attributes ($2304 \times 585 \times 551$ choices for each placed object). We randomly place 0 to 4 human avatars in the scene, randomizing their gender and clothing. We set camera poses as explained above, keeping scene ID shared for all the cameras of the same scene, and randomly sample 4 viewpoints of each scene. We use human motion assets (as explained above) and randomly sample a motion sequence for each human avatar out of 1898 imported or generated mocaps. Finally, we sample an average of 1500 frames from each resulting synthetic video sequence generating image-text pairs describing a scene with the same objects and attributes, but in different arrangements and different corresponding captions thus enhancing the importance of the compositional aspects of the scene in the contrastive loss. We explore the importance of different simulation aspects in our ablation studies in Sec. 4.4.

Metadata-driven caption text synthesis: In addition to RGB frames, we obtain a large collection of rich metadata from the simulation platform, containing information on objects, humanoids, and the scene setting. For each frame, the metadata includes: (i) The world coordinates of each object and humanoid in the scene, including the camera position and viewing direction. (ii) The physical attributes of each object and humanoid in the scene (object physical attributes include color, size, and material; human attributes include the per-frame action label that changes over time and clothing description). (iii) Rendered depth images, instance segmentation masks, and category segmentation masks. Using the metadata, we compute the positional relations between each pair of objects and/or humans by comparing the pixels covered by their segmentation masks as well as their camera coordinates. Then, we use a simple grammar that deterministically maps positional relationships, object attributes, human attributes and action descriptions, and scene descriptions to a well-formed caption. More details on the grammar are provided in Supplementary.

3.2. Finetuning large-scale pre-trained VL models using synthetic data

In this section, we propose a methodology for effectively leveraging the SyViC synthetic VL data produced as explained in Sec. 3.1. We will use the following notation. Let (T, I) be the text & image pair admitted by a VL model. The model (e.g., CLIP [45], CyCLIP [15]) components are denoted as: (i) image encoder $e_I = \mathcal{E}_I(I)$; (ii) text encoder $e_T = \mathcal{E}_T(T)$. In this notation, the text-to-image similarity score is computed as:

$$\mathcal{S}(T, I) = cos(\mathcal{E}_T(T), \mathcal{E}_I(I))) = cos(e_T, e_I), \quad (1)$$

where cos is the cosine similarity (inner product of normalized vectors). We next describe in detail the components of our finetuning strategy. Their merit and tradeoffs are thoroughly investigated in Sec. 4.4, arriving at the conclusion that parameter efficient finetuning + domain adaptive stylization + proposed caption splitting technique are the most effective combination. We also confirm in Sec. 4.4, that model averaging can provide expected trade-offs between VLC understanding and compositional reasoning gains and maintaining zero-shot performance.

Avoiding forgetting through parameter efficient fine-tuning: Inspired by [9, 48], we use LoRA [17] for VL fine-tuning with reduced forgetting of base model performance. We apply LoRA [17] to adapt the encoders $(\mathcal{E}_T, \mathcal{E}_I)$ of a pre-trained VL model by parameterizing the adapted weights $\mathcal{W}_k^*$ corresponding to the original model weights $\mathcal{W}_k$ for each layer k as:

$$\mathcal{W}_k^* = \mathcal{W}_k + \mathcal{A}_k \cdot \mathcal{B}_k \tag{2}$$

where for $\mathcal{W}_k$ of size $m \times l$, $\mathcal{A}_k$ and $\mathcal{B}_k$ are rank-r matrices of sizes $m \times r$ and $r \times l$ respectively. These low-rank residual adapters can be applied efficiently during training and collapsed at inference time resulting in zero cost in terms of inference speeds or parameter counts [17]. During finetuning all the base model parameters $\forall k, \{\mathcal{W}_k\}$ are frozen and only the LoRA adapters $\forall k, \{(\mathcal{A}_k, \mathcal{B}_k)\}$ are being learned. Keeping rank r low, the number of extra parameters added by all the LoRA adapters is low, consequently leading to significantly reduced forgetting in terms of largely maintaining the zero-shot performance of the original VL model.

Further reducing forgetting via model averaging: [56] introduced an elegant technique to mitigate forgetting in finetuned models. All the parameters of the source model (before finetune) and the final model (after finetune) are averaged between the two models (typically with $\alpha = 0.5$ weight). We evaluate the effect of this on **SyViC** finetuned models in our ablation Sec. 4.4.

Domain adaption using style transfer: In addition, to mitigate the domain gap introduced by the use of synthetic data, we experiment with two style transfer techniques that align the content and feature statistics of the input frames with randomly-selected real-life images. A pre-trained Adaptive Instance Normalization (AdaIN) [18] enabled encoder-decoder model was used to align the channel-wise statistics of each synthetic frame with a randomly-sampled image from the Human Motion Database (HMDB51) [26] dataset thus generating a stylized synthetic image. We use AdaIN with an interpolation factor $\alpha = 0.5$. In addition, in order to preserve the color information in the synthetic frames, we first match the color distribution of the sampled style image to that of the synthetic frame [14]. We additionally experimented with MixStyle [65] (using ImageNet as a source of real style images) as an extension of the DA stylization pipeline without observing significant gains over AdaIN.

Handling arbitrary caption length with caption splitting: The captions generated for SyViC are comprehensive: they contain descriptions of every object and/or humanoid visible in the frame as well as the pairwise positional relationship between objects. Intuitively, including these more elaborate (dense) descriptions in our captions gives a clear advantage in terms of promoting VLC understanding and compositionality following the finetuning of a VL model on **SyViC**. Hence, captions need to be sufficiently long texts that cannot be fully processed by common VL models (e.g. CLIP) text encoders $(\mathcal{E}_T)$ during training, as those are caped by relatively short max sequence context length (e.g. 77 for CLIP). Therefore, inspired by CLIP multi-caption strategy for inference [45], during training, we handle arbitrary caption lengths by splitting a given caption into sub-captions that can each be encoded separately and averaging the text features obtained from each sub-caption. In particular, the features of a caption of arbitrary length text T is:

$$\mathcal{E}_T(T) = \frac{1}{n}\sum_{i}^{n} \mathcal{E}_T(T_i) \tag{3}$$

where T_i is a sub-caption comprised of one or more sentences that fit into the text encoder max context size.

Losses: We employ the original models (e.g. CLIP [45] and CyCLIP [15]) contrastive and other losses when training on **SyViC** with the aforementioned architectural and training protocol changes as explained above.

4. Experiments

4.1. Implementation details

For CLIP, we use the original OpenAI CLIP implementation and checkpoints. We modify their codebase to include LoRA adapters (Sec. 3.2), and use rank 16 in all our experiments. For CyCLIP, we adapt the implementation used in [9] [3]. For both CLIP and CyCLIP, we use a 5e-7 initial learning rate for finetuning and follow a cosine annealing learning rate schedule [34] using an Adam [24] optimizer. For all experiments, we use ViT/32-B as the model architecture and fine-tune it for six epochs on one A100 GPU with a total batch size of 400 image-caption pairs. In addition to the original CLIP data augmentation transforms, we apply heavy random augmentation policies including manipulations in image inversion, contrast, sharpness, equalization, posterization, colorization, brightness, and solarization.

4.2. Datasets

To test the effectiveness of our proposed **SyViC** synthetic dataset and the accompanying finetuning approach for improving VL models' VLC understanding and compositional reasoning capabilities we have evaluated on 3 benchmarks (Winoground [51], VL-Checklist [62], and ARO [59]) consisted of 7 datasets total.

VL-Checklist [62] – is a large-scale dataset comprised of: Visual Genome [25], SWiG [43], VAW [42], and HAKE [30]. Each image of these datasets is associated with two captions, a positive and a negative. The positive caption corresponds to the image and is taken from the source

[3] Code and checkpoints kindly shared by the authors.

	VL Checklist			ARO					Zero-Short
	Relation	Attribute	**Average**	VG-Rel.	VG-Att.	Flickr30k	COCO	**Average**	(21 tasks)
CLIP	63.57	67.51	65.54	58.84	63.19	47.20	59.46	57.17	56.07
CyCLIP	61.15	66.96	64.06	59.12	65.41	20.82	29.54	43.72	55.99
syn-CLIP	69.39 **(+5.82)**	70.37 **(+2.86)**	69.88 **(+4.34)**	71.40 **(+12.56)**	66.94 **(+3.75)**	59.06 **(+11.86)**	70.96 **(+11.5)**	67.09 **(+9.9)**	55.27 **(-0.8)**
syn-CyCLIP	65.73 **(+4.58)**	68.06 **(+1.1)**	66.89 **(+2.83)**	69.02 **(+9.9)**	63.65 **(-1.76)**	49.17 **(+28.35)**	59.36 **(+29.82)**	60.30 **(+16.58)**	55.40 **(-0.6)**

Table 1. Performance of syn-<*model*>s – finetuned on **SyViC** using our proposed recipe, measured on VL-Checklist [62] and ARO [59]. Gains and losses are highlighted in green and red respectively.

	Winoground			Winoground†		
	Text	Image	**Group**	Text	Image	**Group**
CLIP	31.25	10.50	8.00	31.58	10.53	8.19
syn-CLIP	30.00	11.50	9.50 (+1.50)	29.82	12.28	9.94 (+1.75)

Table 2. Winoground [51] performance of syn-CLIP – finetuned on **SyViC**. The syn-CyCLIP results on Winoground are provided in the Supplementary. † 'clean' (no-tag) subset of valid Winoground samples from [8]

objects with attr. randomization	humans	VL Checklist		
		Relation	Attribute	Average
CLIP		63.57	67.51	65.54
✗	✓	64.03	67.09	65.56
✓	✗	65.00	68.15	65.95
✓	✓	**69.39**	**70.37**	**69.88**

Table 3. Importance of human avatars, objects, and object attribute variations, evaluated on VL-Checklist and CLIP

dataset. The negative caption is made from the positive caption by changing one word, so the resulting sentence no longer corresponds to the image. Depending on the word that was changed, VL-Checklist evaluates 7 types of VLC that can be divided into two main groups: (1) Attributes – color, material, size, state, and action, and (2) Relations – spatial or action relation between two objects and/or humans. In the following, we report average results for each of the main (Rel. and Attr.) groups on the combined VL-Checklist dataset. We also detail the individual improvements on all 7 VLC types in Fig. 3 (left).

Winoground [51] – is a small dataset that evaluates the ability of VL models for compositional reasoning, specifically understanding the meaning of the sentence after changing the order of its words. The dataset has 400 samples, each comprised of two images and two texts. The texts have the same words in a different order, each text corresponding to one image in the sample. The Winoground metrics include (a) image score - percent of samples where the model picks the correct text for each image; (b) text score - percent of samples where the model picks the correct image for each text; (c) group score - percent of samples where both text and image score conditions are satisfied jointly. Recently, [8] has analyzed Winoground for the source of its difficulty and found that only 171 of its 400 samples are a valid subset. Other samples are not compositional, ambiguous, related to invisible details, have highly uncommon images or text, or require complex reasoning beyond compositionality. We report results on both the full Winoground and the 'clean' 171 images subset from [8].

ARO [59] – or the Attribution, Relation, and Order benchmark, is a large dataset designed to evaluate the ability of VL models to understand four different types of skills. It consists of Visual Genome Attribution and Visual Genome Relation, which leverages the Visual Genome [25] dataset along with the GQA [19] annotations to test the understanding of properties and relational understanding of objects in complex natural scenes. VG-Relation includes 48 distinct relations with 23937 test cases, and VG-Attribution includes 117 unique attribute pairs with 28748 test cases. It also leverages the COCO [31] and Flickr30k [58] datasets to evaluate the model sensitivity to select the right caption after applying four different shuffling perturbations (e.g., exchanging nouns and adjectives, or by shuffling trigrams). These tests are performed on the 5000 and the 1000 images from the respective COCO and Flickr30k test splits.

4.3. Results

The main results of finetuning CLIP [45], CyCLIP [15] – one of CLIP's most recent improvements are summarized in Tables 1 and 2. All models were finetuned using our proposed approach and **SyViC** synthetic data to obtain their syn-<*model*> variants. Each model is compared to its respective source model pre-trained on large-scale real data before finetuning on **SyViC**. As we can observe, our **SyViC** synthetic data and the proposed finetuning recipe on this data demonstrate significant improvements over their source baselines. E.g. for CLIP obtaining 1.75%, 4.34%, and 9.9% average absolute improvement in Winoground group score (most difficult average metric), VL-Checklist and ARO respectively. In addition, we illustrate the individual VLC metrics improvements obtained for CLIP in VL-Checklist and ARO benchmarks in Fig. 3 showing up to 9.1% and 12.6% respective absolute improvements. This underlines the effectiveness and promise of our method and **SyViC** synthetic data towards improving VLC understanding and compositional reasoning in VL models. Importantly, as we can see from Table 1, these strong gains come at a very small (under 1%) cost in the zero-shot performance of the respective VL models measured using the standard Elevater [27] benchmark using 21 diverse zero-shot tasks.

SURREAL	Multi-Garment	VL Checklist		
		Relation	Attribute	Average
CLIP		63.57	67.51	65.54
✗	✗	67.56	68.73	68.15
✓	✗	67.56	67.26	67.41
✗	✓	**69.39**	**70.37**	**69.88**

Table 4. Importance of human avatar clothing choices between SURREAL, Multi-Garment, and simple color textures (corresponding to none), evaluated on VL-Checklist and CLIP

4.4. Ablations

We extensively ablate our **SyViC** synthetic data and the proposed VL models finetuning approach on this data according to the following points. We use the most popular CLIP model finetuned on our **SyViC** synthetic dataset evaluated on the largest of the benchmarks - the VL-Checklist to perform our ablations.

SyViC - objects, humans, object attribute randomization – we evaluate the major components that comprise our **SyViC** synthetic data, namely the importance of the synthetic data to contain humans performing various motions and actions, the importance of having objects with randomized attributes (Sec. 3.1), and the final result of having all types of data combined. The results of this ablation are summarized in Tab. 3. As expected, humans alone cannot teach the model the needed skills only improving relations VLC by a small margin. Additionally, having only objects with randomized attributes improves attribute VLC, yet only improves relations by 1.4% which is also expected, as many of the relations involve human actions. The best result is observed on the combined dataset with all the components.

SyViC - human clothing – we evaluate the diversity of human clothing comparing 3 levels of diversity: (i) none - using only a uniform color for human models; (ii) basic - using less diverse texture maps from SURREAL [55]; and (iii) most diverse - using texture maps from Multi-Garment [4], enriched with clothing colors, human age, and hair color annotations (manually done by us for the textures) which increase captions' expressivity. Results are presented in Table 4. As expected, the most diverse human textures deliver the best result underlining the importance of this factor. Surprisingly, better human textures improve VL-Checklist Relations metric performance, likely due to the significantly better realism of the Multi-Garment textures.

SyViC - types of object attributes to randomize – Table 5 examines how randomizing different object attributes affects performance. Specifically, we evaluate the randomization of size, material, and color. Interestingly, we find that the best performance is achieved without color randomization. We suspect it is due to unnatural color-object combinations that arise under such randomization, which teach the model wrong beliefs on real objects' color distributions and go against true object-color associations existing in the VL model following pre-training on the original VL data.

color	size	material	VL Checklist		
			Relation	Attribute	Average
CLIP			63.57	67.51	65.54
✓	✗	✗	67.71	64.61	66.16
✗	✓	✗	68.58	68.23	68.40
✗	✗	✓	65.23	67.01	66.12
✓	✓	✓	66.67	65.97	66.32
✗	✓	✓	**69.39**	**70.37**	**69.88**

Table 5. Importance of different kinds of object attributes randomization, evaluated on VL-Checklist and CLIP

SyViC - types of captioning – we have investigated several variants of ways to obtain textual captions from **SyViC** metadata (Sec. 3.1). Results are summarized the Supplementary. We compared our proposed metadata grammar-based approach to two cascade methods that paraphrase the captions resulting from the grammar using zero-shot LLM inference (in-context learning). The paraphrased caption is then appended to the original grammar-based caption and consumed through our caption-splitting module (as standalone, open LLM-based paraphrasing is not very high quality). As can be seen, currently paraphrasing has minimal effect, but we posit it will become an important tool as stronger LLMs will become openly available.

SyViC - importance of physics and number of humans in the scene – we also looked into to which extent reliable physical simulation (made available in **SyViC** through TDW and Unity capabilities) and human-human positional and other relations are important for the observed VL model improvements. In Table 6 we evaluate the effects of removing the physics (resulting in humans or objects floating in space) or removing the multi-human scenes (thus preventing all human-human relations from appearing in the data). As expected, both reliable physics simulation and human-human relations (interactions of sorts) are mostly important to the gains in the Relations metric.

SyViC - number of models and number of samples – for lack of space, this is explored in the Supplementary.

SyViC - finetuning recipe components – in Table 7 we extensively evaluate the different components of the proposed finetuning approach on **SyViC** that leads to significant improvements on VL-Checklist, ARO, and Winoground VLC and compositional reasoning metrics. We start with vanilla CLIP finetuning on **SyViC** (row #1), already showing some improvement in VL-Checklist relations metrics, and on the

physics	multi-human	VL Checklist		
		Relation	Attribute	Average
CLIP		63.57	67.51	65.54
✓	✗	67.66	69.24	68.45
✗	✓	65.91	69.01	67.46
✓	✓	69.39	70.37	69.88

Table 6. Importance of physics simulation and multi-human relations in **SyViC**, evaluated on VL-Checklist and CLIP

#	LoRA	freezing $\mathcal{E}_I$	model averaging	DA styl.	Caption split emb.	VL Checklist Rel.	VL Checklist Attr.	VL Checklist Avg.	ARO Avg.	ZS (21 tasks)
			CLIP			63.57	67.51	65.54	57.17	56.07
1	✗	✗	✗	✗	✗	67.10	65.45	66.28	62.83	53.84
2	✗	✗	✗	✗	✓	67.76	67.51	67.64	59.27	53.87
3	✓	✗	✗	✗	✓	69.32	69.46	69.39	64.34	53.16
4	✓	✗	✗	✓	✓	**69.39**	**70.37**	**69.88**	**67.09**	55.27
5	✓	✓	✗	✓	✓	63.54	68.20	65.87	60.29	54.54
6	✓	✗	✓	✓	✓	66.70	69.62	68.16	63.76	**55.72**
7	✓	✓	✗	✓	✗	65.16	66.88	66.02	57.55	52.69

Table 7. Importance of the finetuning recipe components, evaluated on VL-Checklist and CLIP

ARO benchmark, yet losing to base CLIP on VL-Checklist attributes metrics. Adding our caption splitting module (Sec. 3.2) allows handling long (arbitrary size) texts outputted by our metadata-driven grammar and consequently utilizes all the caption information re-gaining the attributes performance (row #2). Adding parameter-efficient finetuning (LoRA, Sec. 3.2) regularizes finetuning by forcing smaller (low-rank, low-parameters) updates of the large-scale pre-trained CLIP model, consequently somewhat handling the expected domain gap between the synthetic data of **SyViC** and the downstream evaluation (real data) tasks. Notably, LoRA does not add any additional parameters to the model, all LoRA adapters are collapsed into the model weights after finetuning. Consequently, we observed significant improvements from adding LoRA in all metrics (row #3) with only minor degradation (0.7%) in ZS evaluation. With adding domain stylization (Sec. 3.2) we observe the best results in all VLC and compositional reasoning metrics improving ARO by 2.8% and keeping (even slightly improving) the advantages on VL-Checklist. Next, we investigate the variations of our best approach (LoRA + domain stylization + caption splitting). First, we investigate a strategy inspired by the LiT [61] approach (row #5). Freezing the visual encoder $\mathcal{E}_I$ as expected provides a (small) boost in ZS performance, but the reduced plasticity of the model comes at the price of observing smaller (only 3.1%) improvements on ARO and almost no improvements on the VL-Checklist. This leads us to conclude, that freezing the visual encoder is not a good strategy for **SyViC** finetuning. Next, we check the model averaging strategy (Sec. 3.2) inspired by [56] (row #6). This does a better job of mitigating ZS forgetting, while at the same time keeping more of the gains on VL-Checklist and ARO. We conclude that model averaging is a good strategy to complement **SyViC** finetuning, allowing a soft trade-off between mitigating ZS forgetting and VLC and compositionality metrics gains. Finally, we again explore the importance of caption splitting for the best finetuning configuration of **SyViC** (row #7) and re-confirm its significance as performance drops without it.

5. Summary & Conclusions

Large vision and language models have dictated the status quo in computer vision and multimodal perception, achieving state-of-the-art results in a number of challenging benchmarks. However, existing models struggle with compositional reasoning and understanding concepts beyond object nouns, such as attributes and relationships. Our work has investigated, for the first time, whether synthetic data can be leveraged to mitigate these shortcomings. We proposed a data generation pipeline, used to create a million-scale dataset of synthetic images and accompanying captions, and an effective fine-tuning strategy with comprehensive analysis to enhance the compositional and concept understanding capabilities of multimodal models, without compromising their zero-shot classification performance.

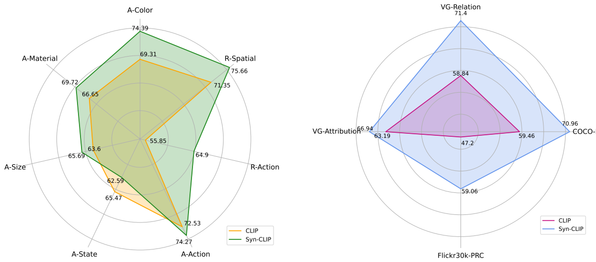

Figure 3. **(left)** detailed evaluation of syn-CLIP on the 7 separate VL-Checklist [62] metrics; **(right)** detailed evaluation of syn-CLIP on all the Compositional Tasks proposed in ARO [59].

Limitations. While we have achieved quite promising results in three different benchmarks, our work has limitations. As an example, our graphics simulator has a simplified model of lighting, sensor noise, and reflectance functions compared to the real world, which may impact robustness to color constancy. We believe more advanced domain adaptation and rendering techniques are likely needed to further improve our results. We also think a more detailed study of the scaling laws for synthetic data is a great research direction to fully unlock the potential of our work.

Acknowledgements. This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. FA8750-19-C-1001. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Defense Advanced Research Projects Agency (DARPA). This project was also partially supported by NSF Award IIS-2221943, ANR CorVis ANR-21-CE23-0003-01, and the MIT-IBM Watson AI Lab. We thank Jeremy Schwartz and Ester Alter for their helpful discussions and assistance with the TDW package.

References

[1] Nikos Athanasiou, Mathis Petrovich, Michael J. Black, and Gül Varol. Teach: Temporal action composition for 3d humans. In *3DV*, 2022.

[2] Nikos Athanasiou, Mathis Petrovich, Michael J. Black, and Gül Varol. Teach: Temporal action composition for 3d humans. *2022 International Conference on 3D Vision (3DV)*, pages 414–423, 2022.

[3] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18208–18218, 2022.

[4] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In *IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2019.

[5] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022.

[6] Paola Cascante-Bonilla, Hui Wu, Letao Wang, Rogerio Feris, and Vicente Ordonez. Sim vqa: Exploring simulated environments for visual question answering. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5046–5056, 2022.

[7] César Roberto De Souza, Adrien Gaidon, Yohann Cabon, and Antonio M. López Peña. Procedural generation of videos to train deep action recognition networks. In *CVPR*, 2017.

[8] Anuj Diwan, Layne Berry, Eunsol Choi, David Harwath, and Kyle Mahowald. Why is winoground hard? investigating failures in visuolinguistic compositionality. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2236–2250, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics.

[9] Sivan Doveh, Assaf Arbelle, Sivan Harary, Rameswar Panda, Roei Herzig, Eli Schwartz, Donghyun Kim, Raja Giryes, Rogerio Feris, Shimon Ullman, et al. Teaching structured vision&language concepts to vision&language models. *arXiv preprint arXiv:2211.11733*, 2022.

[10] Salehe Erfanian Ebadi, You-Cyuan Jhang, Alex Zook, Saurav Dhakad, Adam Crespi, Pete Parisi, Steve Borkman, Jonathan Hogins, and Sujoy Ganguly. Peoplesanspeople: A synthetic data generator for human-centric computer vision. 2021.

[11] Chuang Gan, Jeremy Schwartz, Seth Alter, Martin Schrimpf, James Traer, Julian De Freitas, Jonas Kubilius, Abhishek Bhandwaldar, Nick Haber, Megumi Sano, Kuno Kim, Elias Wang, Damian Mrowca, Michael Lingelbach, Aidan Curtis, Kevin T. Feigelis, Daniel Bear, Dan Gutfreund, David Cox, James J. DiCarlo, Josh H. McDermott, Joshua B. Tenenbaum, and Daniel L. K. Yamins. Threedworld: A platform for interactive multi-modal physical simulation. *ArXiv*, abs/2007.04954, 2020.

[12] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030, 2016.

[13] Yuting Gao, Jinfeng Liu, Zihan Xu, Jun Zhang, Ke Li, and Chunhua Shen. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining. *arXiv preprint arXiv:2204.14095*, 2022.

[14] L. A. Gatys, A. S. Ecker, M. Bethge, A. Hertzmann, and E. Shechtman. Controlling perceptual factors in neural style transfer. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3730–3738, Los Alamitos, CA, USA, jul 2017. IEEE Computer Society.

[15] Shashank Goel, Hritik Bansal, Sumit Bhatia, Ryan A Rossi, Vishwa Vinay, and Aditya Grover. Cyclip: Cyclic contrastive language-image pretraining. *arXiv preprint arXiv:2205.14459*, 2022.

[16] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

[17] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[18] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017.

[19] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019.

[20] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 2021.

[21] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-based real image editing with diffusion models. *arXiv preprint arXiv:2210.09276*, 2022.

[22] Donghyun Kim, Kaihong Wang, Kate Saenko, Margrit Betke, and Stan Sclaroff. A unified framework for domain adaptive pose estimation. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIII*, pages 603–620. Springer, 2022.

[23] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022.

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[25] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017.

[26] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. HMDB: a large video database for human motion recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2011.

[27] Chunyuan Li, Haotian Liu, Liunian Harold Li, Pengchuan Zhang, Jyoti Aneja, Jianwei Yang, Ping Jin, Yong Jae Lee, Houdong Hu, Zicheng Liu, and Jianfeng Gao. Elevater: A benchmark and toolkit for evaluating language-augmented visual models. *Neural Information Processing Systems*, 2022.

[28] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022.

[29] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*, 2021.

[30] Yong-Lu Li, Liang Xu, Xinpeng Liu, Xijie Huang, Yue Xu, Mingyang Chen, Ze Ma, Shiyi Wang, Hao-Shu Fang, and Cewu Lu. Hake: Human activity knowledge engine. *arXiv preprint arXiv:1904.06539*, 2019.

[31] 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: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[32] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015.

[33] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015.

[34] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations*, 2017.

[35] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, Oct. 2019.

[36] Samarth Mishra, Rameswar Panda, Cheng Perng Phoo, Chun-Fu Richard Chen, Leonid Karlinsky, Kate Saenko, Venkatesh Saligrama, and Rogerio S Feris. Task2sim: Towards effective pre-training and transfer from synthetic data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9194–9204, 2022.

[37] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. *arXiv preprint arXiv:2211.09794*, 2022.

[38] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. *arXiv preprint arXiv:2302.03027*, 2023.

[39] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019.

[40] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*, 2017.

[41] Mathis Petrovich, Michael J. Black, and Gül Varol. TEMOS: Generating diverse human motions from textual descriptions. 2022.

[42] Khoi Pham, Kushal Kafle, Zhe Lin, Zhihong Ding, Scott Cohen, Quan Tran, and Abhinav Shrivastava. Learning to predict visual attributes in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13018–13028, 2021.

[43] Sarah Pratt, Mark Yatskar, Luca Weihs, Ali Farhadi, and Aniruddha Kembhavi. Grounded situation recognition. In *European Conference on Computer Vision*, pages 314–332. Springer, 2020.

[44] Abhinanda R. Punnakkal, Arjun Chandrasekaran, Nikos Athanasiou, Alejandra Quiros-Ramirez, and Michael J. Black. BABEL: Bodies, action and behavior with English labels. 2021.

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[46] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 102–118. Springer, 2016.

[47] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3234–3243, 2016.

[48] James Seale Smith, Paola Cascante-Bonilla, Assaf Arbelle, Donghyun Kim, Rameswar Panda, David Cox, Diyi Yang, Zsolt Kira, Rogerio Feris, and Leonid Karlinsky. Construct-vl: Data-free continual structured vl concepts learning. *arXiv e-prints*, pages arXiv–2211, 2022.

[49] Ashish Shrivastava, Tomas Pfister, Oncel Tuzel, Joshua Susskind, Wenda Wang, and Russell Webb. Learning from simulated and unsupervised images through adversarial training. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2107–2116, 2017.

[50] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15638–15650, 2022.

[51] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross.

Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022.

[52] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017.

[53] Carnegie Mellon University. CMU graphics lab motion capture database. http://mocap.cs.cmu.edu/.

[54] Gül Varol, Ivan Laptev, Cordelia Schmid, and Andrew Zisserman. Synthetic humans for action recognition from unseen viewpoints. *IJCV*, 2021.

[55] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *CVPR*, 2017.

[56] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7959–7971, 2022.

[57] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*, 2021.

[58] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 02 2014.

[59] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2023.

[60] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021.

[61] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022.

[62] Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *arXiv preprint arXiv:2207.00221*, 2022.

[63] Yuyang Zhao, Zhun Zhong, Na Zhao, Nicu Sebe, and Gim Hee Lee. Style-hallucinated dual consistency learning: A unified framework for visual domain generalization. *arXiv preprint arXiv:2212.09068*, 2022.

[64] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. *arXiv preprint arXiv:2104.02008*, 2021.

[65] Kaiyang Zhou, Yongxin Yang, Yu Qiao, and Tao Xiang. Domain generalization with mixstyle. In *ICLR*, 2021.