# SimVQA: Exploring Simulated Environments for Visual Question Answering

Paola Cascante-Bonilla[†*] Hui Wu[‡] Letao Wang[♮] Rogerio Feris[‡] Vicente Ordonez[†]
[†]Rice University [‡]MIT-IBM Watson AI Lab [♮]University of Virginia
{pc51, vicenteor}@rice.com, {wuhu, rsferis}@us.ibm.com, lw7jz@virginia.edu

## Abstract

*In this work we explore using synthetic computer-generated data to fully control the visual and language space, allowing us to provide more diverse scenarios. We quantify the effectiveness of leveraging synthetic data for real-world VQA. By exploiting 3D and physics simulation platforms, we generate synthetic data to expand and replace type-specific questions and answers without risking exposure of sensitive or personal data that might be present in real images. We offer a comprehensive analysis while expanding existing hyper-realistic datasets to be used for VQA. We also propose Feature Swapping (F-SWAP) – where we randomly switch object-level features during training to make a VQA model more domain invariant. We show that F-SWAP is effective for improving VQA models on real images without compromising on their accuracy to answer existing questions in the dataset.*

## 1. Introduction

Data augmentation is an effective way to achieve better generalization on several visual recognition and natural language understanding tasks. Existing work on Visual Question Answering (VQA) has explored augmenting the pool of questions and answers or by perturbing or masking some parts of the images [1, 5, 9]. Moreover, curating large-scale datasets is a laborious task and sourcing images is an expensive process that needs to account for practical issues such as copyright and privacy. Augmenting existing datasets with synthetically generated data offers a path to enhance our existing data-driven models at a lower cost.

Our work focuses on leveraging synthetically generated data through the use of modern 3D generated computer graphics using a couple of novel resources – Hypersim [7] and ThreeDWorld (TDW) [2]. We also propose feature swapping (F-SWAP), a simple yet effective method to augment a currently existing VQA dataset with computer graphics generated examples[1]. Existing methods for domain adaptation rely on the assumption that adaptation can be addressed by making the out-of-domain samples match the distribution of the in-domain samples. However current work often operationalizes this assumption by making the input images themselves look more like the real images e.g. [4, 8, 10]. Feature Swapping relies instead on swapping random object-level intermediate feature representations. We posit that unless realistic style-transfer is desired from the input domain to the target domain, as long as the two domains are matched at the feature level – domain adaptation can take place. We explain and compare our F-SWAP approach with other methods such as adversarial domain adaptation and demonstrate superior results.


Figure 1. Training samples for VQA from real and synthetic datasets. The first row shows existing examples from the VQA 2.0 dataset. The second row shows examples from Hypersim [7], a hyper-realistic synthetic dataset we extend for VQA. The third row shows some examples we generate using ThreeDWorld [2].

Our contributions can be summarized as follows:

- Synthetic datasets: We are providing an extension of the Hypersim dataset for VQA, and provide a synthetic VQA dataset using ThreeDWorld.
- Feature swapping: We are proposing a surprisingly simple yet effective new technique for incorporating

*Work partially done while interning at the MIT-IBM Watson AI Lab
[1]Project page: https://simvqa.github.io/

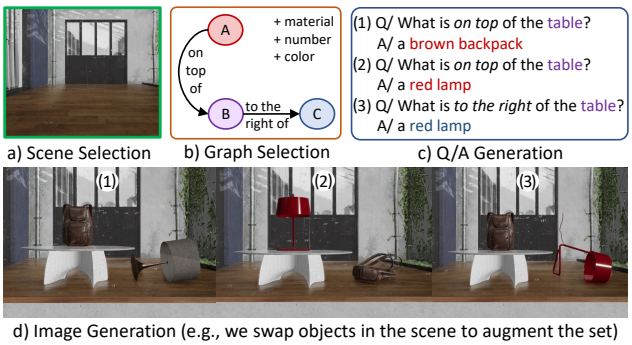


Figure 2. Sample pipeline for generating VQA data using ThreeDWorld. a) Manually select scenes from a set of random camera walks. b) Select one of the generated scene graphs containing object information such as positions, number, color, and materials. c) Generate question-answer pairs following a template based on the scene graph. d) Finally, generate images by placing objects and modifying characteristics of the scene based on steps b and c.

synthetic images in our training data while mitigating the domain shift.

- Experimental results: We are providing an empirical analysis of well known techniques vs our proposed approach to alleviate the visual domain gap.

## 2. Synthetic Datasets

First, we describe the generation of a VQA dataset by extending the existing Hypersim dataset [7] (section 2.1). We name this dataset Hypersim-VQA, or H-VQA, for short. Then we explore the automatic creation of a VQA Dataset using ThreeDWorld [2] (section 2.2). We name this dataset ThreeDWorld-VQA, or W-VQA, for short.

### 2.1. Extending Hypersim for VQA

Hypersim [7] is an existing 3D graphics generated dataset with a high image quality and displays a diverse array of scenes and objects. Hypersim metadata includes the complete geometry information per scene, dense per-pixel semantic instance segmentations for every image, and instance-level NYU40 labels annotations. We extend these data by manually annotating objects on all images given their dimensions and positions in the scene. We generate questions and answer pairs based in the visibility of an object in a scene frame. We call this new set Hypersim-VQA.

### 2.2. Automatic VQA Generation

TDW [2], is a platform for interactive multi-modal physical simulation that we use to generate images. We follow the steps shown in Figure 2 to generate the image $I$, question $Q$ and answer $A$ triplets for our W-VQA dataset. Questions and answers are generated following a template based grammar associated with a predefined scene graph and it's

| Feature backbone | Feature size | Training data Real | Training data Synthetic | | R-VQA Accuracy |
|---|---|---|---|---|---|
| | | R | H | W | *Numeric* |
| FastRCNN – RN101 | 100×2048 | ✓ | | | 42.73 |
| | | ✓ | ✓ | | 44.70 +1.97 |
| | | ✓ | | ✓ | 42.86 +0.13 |
| CLIP - RN50 | 558×2048 | ✓ | | | 42.83 |
| | | ✓ | ✓ | | 43.61 +0.78 |
| | | ✓ | | ✓ | 42.91 +0.08 |
| CLIP – ViT-B | 85×512 | ✓ | | | 41.93 |
| | | ✓ | ✓ | | 43.98 +2.05 |
| | | ✓ | | ✓ | 41.35 −0.58 |

Table 1. Data augmentation using synthetic data improves Real-VQA performance (R-VQA) on numeric questions, especially when using Hypersim-VQA (H). In all these experiments only counting questions were used for training from both the existing Real-VQA dataset, $\text{VQA}_C$ (**R**) and our synthetic dataset variants: Hypersim-VQA (**H**) and TDW-VQA (**W**).

corresponding image. In our setup, a *noun* is directly associated with the model object label category from the TDW asset, *position* is taken from the relationship between objects from the scene graph, and *number*, *adjective_color* and *adjective_material* are taken from the attributes selected when generating the graph and the synthetic image.

## 3. Feature Swapping

Given a triplet of images $I$, questions $Q$ and answers $A$, we have access to three datasets from different domains, where $(I_R, Q_R, A_R) \in R$ correspond to a Real-VQA dataset consisting of real images (we use VQA 2.0 [3]), $(I_H, Q_H, A_H) \in H$ correspond to the Hypersim-VQA dataset, and $(I_W, Q_W, A_W) \in W$ correspond to the TDW-VQA dataset. We assume that the images and their corresponding questions are inputs to a VQA model, and the objective is to predict as output the corresponding ground-truth answers. Given an image $I$, we use a pre-trained model $G$ to extract the image region features $G_f(I) = \{f_1, f_2, ..., f_n\}$ along with their corresponding pseudo-labels $G_{sl}(I) = \{sl_1, sl_2, ..., sl_n\}$ which are assigned based on the attribute annotations from Visual Genome [6]. Since we have access to all images from the three sets, we create a dictionary $D_{type}$ per dataset with $type = R \vee H \vee W$, where the $[key, value]$ of a dictionary $D_{type}$ corresponds to the pseudo-label $(sl_i)_{type}$, and all the region features $[(f_i)_{type}, ..., (f_m)_{type}]$ that the model $G$ assign as $(sl_i)_{type}$ respectively. Once we retrieve the information for all the dictionaries $D_R, D_H, D_W$, we use them to swap features from one dataset to the other. While training, when sampling datapoints from $R$, we randomly select an Image $I_R$ and get all the region features $G_f(I_R)$ and its corresponding pseudo-labels $G_{sl}(I_R)$. Since we have access

| Data | Method | +0% R-VQA$_C$ | | | +1% R-VQA$_C$ | | | +10% R-VQA$_C$ | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | *Numeric* | *Others* | *Overall* | *Numeric* | *Others* | *Overall* | *Numeric* | *Others* | *Overall* |
| H-VQA$_C$ | Simple Augmentation | 15.99 | 68.97 | 62.02 | 29.64 | 68.45 | 63.34 | 35.72 | 68.61 | 64.29 |
| H-VQA$_C$ | Adversarial | 16.07$_{+0.08}$ | 66.01$_{-2.96}$ | 59.46$_{-2.56}$ | 28.31$_{-1.33}$ | 66.89$_{-1.56}$ | 61.83$_{-1.51}$ | 35.01$_{-0.71}$ | 66.91$_{-1.7}$ | 62.71$_{-1.58}$ |
| H-VQA$_C$ | MMD | 24.79$_{+8.80}$ | 67.13$_{-1.84}$ | 61.58$_{-0.44}$ | 31.61$_{+1.97}$ | 67.78$_{-0.67}$ | 63.04$_{-0.30}$ | 38.87$_{+3.15}$ | 68.36$_{-0.25}$ | 64.49$_{+0.2}$ |
| H-VQA$_C$ | Domain Independent | 22.87$_{+6.88}$ | 68.65$_{-0.32}$ | 62.64$_{+0.62}$ | 29.05$_{-0.59}$ | 68.73$_{+0.28}$ | 63.52$_{+0.18}$ | 37.67$_{+1.95}$ | 69.34$_{+0.73}$ | 65.17$_{+0.88}$ |
| H-VQA$_C$ | Feature Swapping (F-SWAP) | 23.38$_{+7.39}$ | 69.07$_{+0.10}$ | 63.07$_{+1.05}$ | 31.64$_{+2.00}$ | 69.08$_{+0.63}$ | 64.15$_{+0.81}$ | 39.71$_{+3.99}$ | 69.13$_{+0.52}$ | 65.26$_{+0.97}$ |
| W-VQA$_C$ | Simple Augmentation | 21.18 | 68.91 | 62.65 | 31.18 | 68.97 | 64.01 | 38.47 | 68.86 | 64.87 |
| W-VQA$_C$ | Feature Swapping (F-SWAP) | 26.84$_{+5.66}$ | 68.89$_{-0.02}$ | 63.67$_{+1.02}$ | 31.21$_{+0.03}$ | 68.82$_{-0.15}$ | 63.89$_{-0.12}$ | 38.54$_{+0.07}$ | 68.97$_{+0.11}$ | 64.97$_{+0.10}$ |

Table 2. Counting skill learning under different low-regime settings for Real VQA counting questions (R-VQA$_C$). All models share the basic training set: VQA$_{NC}$ (the non-counting subset of VQA v2 training data).

to all dictionaries, we lookup for the pseudo-labels that also exist in $D_H \vee D_W$, for simplicity, $D_S = D_H \vee D_W$, thus, after obtaining $G_{sl}(I_R) \in D_S$ we proceed to randomly select a portion $\lambda|G_f(I_R)|$ of the corresponding pseudo-labeled features in $D_S$ and replace them with the matching features in $I_R$. In all of our experiments, $\lambda = 0.2$.

# 4. Experimental Settings

**Real-VQA Dataset.** Following Whitehead et al [11]'s skill-concept separation for compositional analysis, we take the VQA 2.0 dataset [3], and separate the counting questions for a detailed analysis on how synthetic data may affect a model performance. For training, we create two different splits: R-VQA$_C$ that corresponds to the training set with only counting questions, and R-VQA$_{NC}$ which corresponds to the VQA 2.0 training set without counting questions. R-VQA$_C$ contains $48,431$ datapoints, and R-VQA$_{NC}$ contains $378,018$ datapoints. For testing, we use the standard VQA 2.0 validation set and report our results on *Numeric* questions, where $\sim$85% of the questions correspond to counting questions, *Others*, and *Overall* for the general accuracy. **Hypersim-VQA.** We generate $254,174$ counting questions for $41,551$ images. In our experiments, we use a subset of $20,000$ questions that only containe NYU40 labels (excluding *otherstructure*, *otherfurniture* and *otherprop*) and include $10,000$ randomly selected from the extra annotated labels. We also generate $40,000$ yes/no questions probing whether an object is present in an image. **TDW-VQA.** We generate $33,264$ counting related datapoints and add $30,000$ yes/no questions to the same images. Additionally, we generate $12,000$ extra images and add color and material questions, for a total of $87,264$ automatically generated datapoints using the ThreeDWorld simulation platform. **Base VQA model.** We select the top-performing model without large-scale pretraining [12] as our base model. Our base code follows the hyper-parameter selection included in their publicly available implementation[2].

## 4.1. Data augmentation experiments.

We evaluate the effect of augmenting Real-VQA data with the proposed synthetic datasets. We are interested to test if the ability of VQA models to answer counting questions on synthetic data could improve the counting performance on real VQA data. Table 1 shows that, under different feature backbones, the performance of counting questions on real data is improved when R-VQA$_C$ is augmented with the proposed H-VQA dataset.

## 4.2. Domain alignment experiments.

We explore to what extent skill learning using synthetic data can be helped by explicit alignment of visual features between two domains. The real data used in this experiment includes R-VQA$_{NC}$, as well as R-VQA$_C$ under three different regimes ($0\%$, $1\%$, $10\%$). Table 2 summarizes the experimental results when using different domain alignment approaches. The results suggest that Feature Swapping outperforms the baseline and other domain alignment methods, and produces consistent gains on counting questions as well as the overall accuracy, across different regimes of VQA$_C$.

# 5. Conclusion

In this work we explored the efficacy of VQA datasets generated using 3D computer graphics to incorporate new skills into existing VQA models trained on real data. We particularly showed that we can teach a VQA model how to count objects in the real world by using only synthetic data while not decreasing the model performance on other types of questions. This is challenging since real and synthetic datasets often exhibit a large domain gap. We further proposed F-SWAP as a simple yet effective technique for domain adaptation that is competitive and surpasses previous methods in our experiments.

[2]https://github.com/MILVLG/mcan-vqa

## References


[1] Vedika Agarwal, Rakshith Shetty, and Mario Fritz. Towards causal vqa: Revealing and reducing spurious correlations by invariant and covariant semantic editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9690–9698, 2020. 1

[2] 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. 1, 2

[3] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017. 2, 3

[4] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. Cycada: Cycle-consistent adversarial domain adaptation. In *International conference on machine learning*, pages 1989–1998. PMLR, 2018. 1

[5] Kushal Kafle, Mohammed Yousefhussien, and Christopher Kanan. Data augmentation for visual question answering. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 198–202, 2017. 1

[6] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73, 2016. 2

[7] Mike Roberts and Nathan Paczan. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. *ArXiv*, abs/2011.02523, 2020. 1, 2

[8] Adrian Lopez Rodriguez and Krystian Mikolajczyk. Domain adaptation for object detection via style consistency. *British Machine Vision Conference (BMVC)*, 2019. 1

[9] Ruixue Tang, Chao Ma, Wei Emma Zhang, Qi Wu, and Xiaokang Yang. Semantic equivalent adversarial data augmentation for visual question answering. In *European Conference on Computer Vision*, pages 437–453. Springer, 2020. 1

[10] 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. 1

[11] Spencer Whitehead, Hui Wu, Heng Ji, Rogério Schmidt Feris, Kate Saenko, and Uiuc MIT-IBM. Separating skills and concepts for novel visual question answering. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5628–5637, 2021. 3

[12] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *CVPR*, pages 6274–6283, 2019. 3