# Synth$^2$: Boosting Visual-Language Models with Synthetic Captions and Image Embeddings

**Sahand Sharifzadeh**[*,1], **Christos Kaplanis**[1], **Shreya Pathak**[1], **Dharshan Kumaran**[1], **Anastasija Ilic**[1], **Jovana Mitrovic**[1], **Charles Blundell**[1] **and Andrea Banino**[*,1]
[*]Equal contributions, [1]Google DeepMind

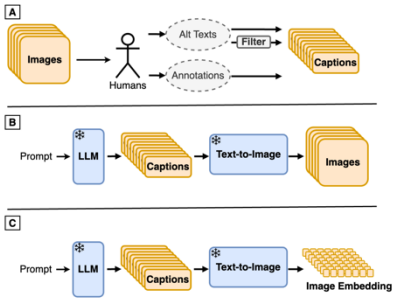


Figure 1 | (A) Traditional dataset curation pipelines require a human in the loop to collect and annotate images. (B) We study whether we can reverse this pipeline with generative models, i.e. by first sampling synthetic captions from an LLM and then synthetically generating images from those. (C) By operating in the image embedding space, we also propose to bypass computationally expensive encoder/decoder steps, optimizing and integrating the process within VLM training.


Figure 2 | Examples of synthetic captions and synthetic images generated by LLM and text-to-image generator.

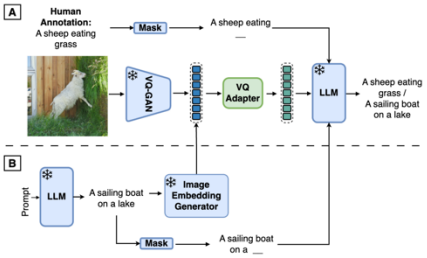


Figure 3 | We introduce a VLM framework that leverages LLMs and image generation models to create synthetic image-text pairs for efficient training. We can train a VLM from both non-synthetic (A) and synthetic (B) data as shown in this figure. Our model trained with the added synthetic pairs demonstrates impressive image captioning performance, significantly reducing the need for human annotated images.

Table 1 | A taxonomy of related work on synthetic data for training vision models.

| Generator | Method | Generator Model | Generated Set | Caption Class | Caption Type | Evaluation Setting |
|---|---|---|---|---|---|---|
| Canonical Concept Mapping | Sharifzadeh et al. (2021)<br>Sharifzadeh et al. (2022) | Linear | Scene Graph Embedding | Scene Graphs<br>Complex Text | Human Generated | SG Classification |
| Simulation/Rendering Engine | Mishra et al. (2022)<br>Greff et al. (2022)<br>Zheng et al. (2020)<br>de Melo et al. (2022) | Mix | Mixed pairs such as<br>(Segmentation, Images)<br>(Optimal Flow, Videos)<br>(Depth Maps, Images) | N/A | N/A | Mix |
| | Cascante-Bonilla et al. (2023) | Mix | (Captions, Images) | Complex Text | Rule-based | Vision Encoder |
| Off-the-shelf Image Generator | Azizi et al. (2023)<br>Fan et al. (2023) | SD<br>Imagen<br>MUSE | Images | Single Word | ImageNet Classes | Classifier |
| | Li et al. (2023c) | SD | Images | Complex Text | Human Generated | VLM |
| **Controlled Image Generator** | **Synth$^2$** | **MUSE** | **Text, Embeddings & Images** | **Complex Text** | **Human & LLM Generated** | **VLM** |