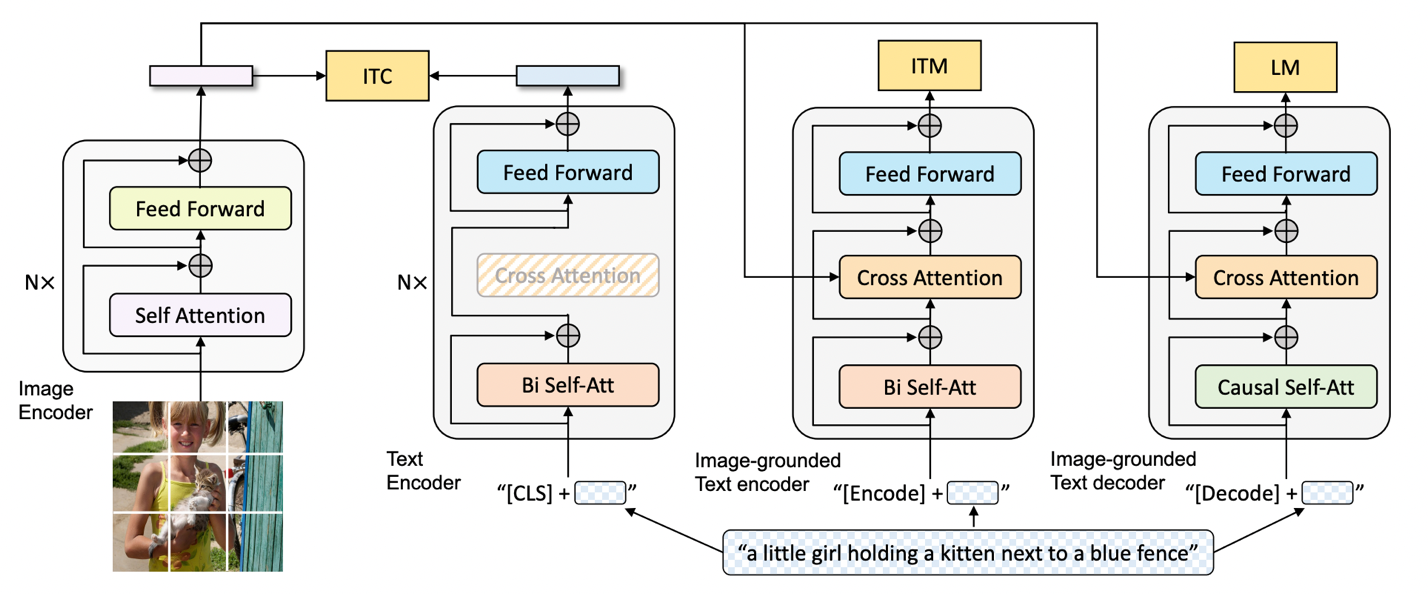
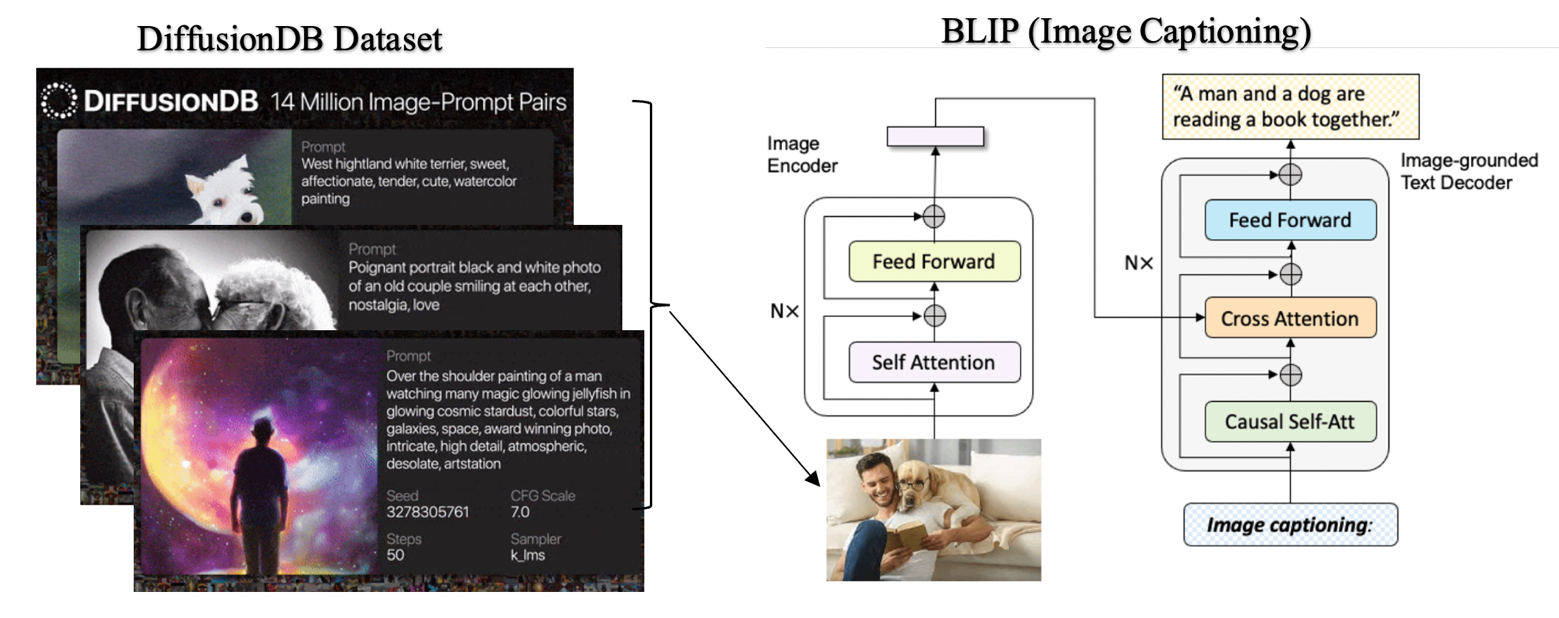
BLIP (Bootstrapping Language Image Pre-training for unified vision-language understanding and generation) is a new VLP (Vision-Language Pre-training) framework that has shown excellent results in vision-language tasks, including understanding-based and generation-based tasks. The BLIP framework is inspired by the bottom-up learning and inference principles of human vision and language processing. The paradigm believes that visual and linguistic information are processed in the human brain in parallel and interactively, and that semantic representations of objects and qualities are critical in vision-language interpretation and production.

**

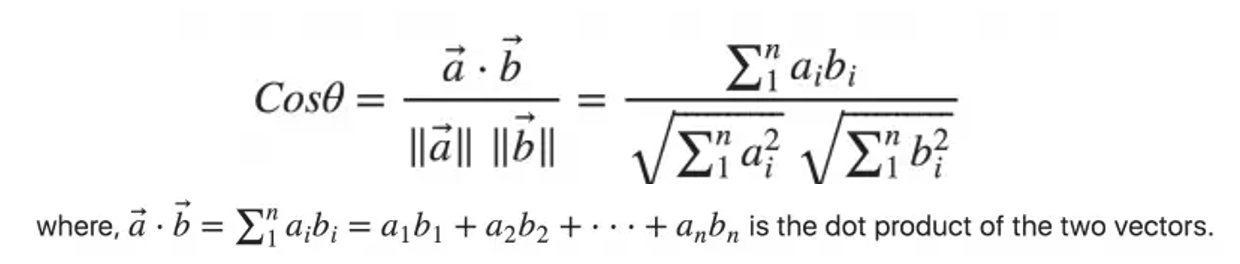
*Figure1. BLIP Model Architecture*

BLIP uses a dataset bootstrapped from large-scale noisy image-text pairings to pre-train a multimodal mixture of encoder-decoder (MED) models by inserting different synthetic captions and deleting noisy captions. The MED is a multitask model that can operate in the following three functionalities: a unimodal encoder that is able to encode image and text, an image-grounded text encoder that injects visual information, and an image-grounded text decoder. The unimodal encoder is trained with an image-text contrastive (ITC) loss to align the vision and language representations. The image-grounded text encoder is trained with an image-text matching (ITM) loss to discriminate between positive and negative image-text pairs, and it makes use of extra cross-attention layers to describe interactions between vision and language. As for the image-grounded text decoder, the bi-directional self-attention layers is replaced by the causal self-attention layers, and the cross-attention layer is the same as the one in the image-grounded text encoder. To create captions for provided photos, the decoder is trained using a language modeling (LM) loss.

With excellent performance for image captioning on the dataset NoCaps [1] and COCO that outperforms state-of-the-art image captioning methods like Enc-Dec and LEMON, BLIP shows its strength in dealing with image-to-prompt tasks. Therefore, we decided to use BLIP and fine-tune it on the DiffusionDB dataset, which is the first extensive dataset for text-to-image prompts. It has 14 million Stable Diffusion-generated photos that were created using prompts and hyperparameters provided by actual users. (Need to add some descriptions about the subdataset that we used: size&why, maybe mention the computational limitations).

*Figure2. Work Flow: (1) Select data from DiffusionDB. (2) Fine-tuned the model and used the model to generate image captioning.*

To evaluate our results, we split our dataset into training, validation, and test sets and use cosine similarity to quantify similarities between our predictions and the actual prompts’ text vectors. Cosine similarity is a measure of similarity between two non-zero vectors defined in an inner product space [2]. Mathematically, it is the result of dividing the vectors' dot products by the sum of their lengths. Hence, regardless of the length of the documents, cosine similarity provides a good indicator of how similar two texts are likely to be in terms of their subject matter. [3] The resulting similarity ranges from 0 indicating orthogonality or decorrelation, to 1 meaning exactly the same, while in between values indicating intermediate similarity.

*Figure3. Cosine Similarity Formula [4]*

[1]H. Agrawal *et al.*, “nocaps: novel object captioning at scale,” *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8947–8956, Oct. 2019, doi: <https://doi.org/10.1109/ICCV.2019.00904>.

[2]Wikipedia Contributors, “Cosine similarity,” *Wikipedia*, Mar. 03, 2019. <https://en.wikipedia.org/wiki/Cosine_similarity>

[3]  [Singhal, Amit](https://en.wikipedia.org/wiki/Amit_Singhal), "[Modern Information Retrieval: A Brief Overview](http://singhal.info/ieee2001.pdf)". *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* 24 (4): 35–43.

<http://singhal.info/ieee2001.pdf>

[4]S. Prabhakaran, “Cosine Similarity – Understanding the math and how it works (with python codes),” *Machine Learning Plus*, Oct. 22, 2018. <https://www.machinelearningplus.com/nlp/cosine-similarity/>