NEW YORK UNIVERSITY

Patch and Control - Steering Behavior of Large Vision-Language Models via Latent Activations

Aditeya Baral Rijul Dahiya Dilip Venkatesh

Motivation

Current LVLM behavior can be conditioned through supervised fine-tuning or explicit prompting.

Leverage activation patching to steer LVLM behavior in the **latent space**, enabling inference-time control **without** fine-tuning or prompts.

Understand how conditioning and multi-modal information propagate through layers and provide insights into **model interpretability**.

Experiments

- Evaluate LVLM steering on Flickr30K images
- Explore multi-modal prompt patching strategies
  - Patch the [control] tokens
  - Patch the [image] tokens
  - Simultaneously patch the [control] and [image]
- Investigate the effect of layer choice and linear combination of original and patched signal

Extraction and Patching

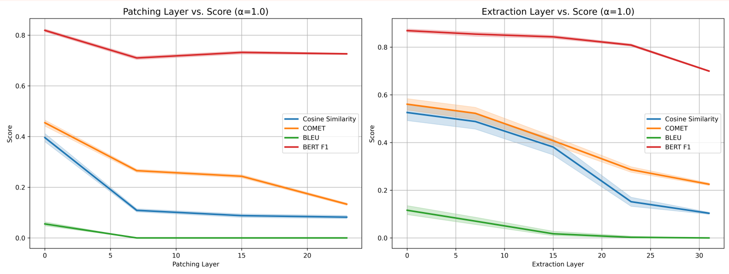


Figure 1. Early layers are best at capturing the most information

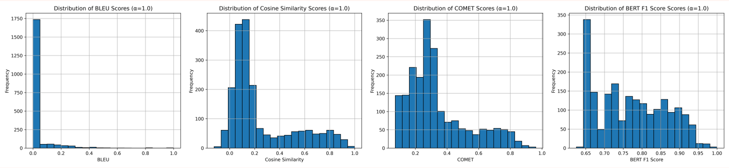


Figure 2. We show satisfactory overlap with gold captions

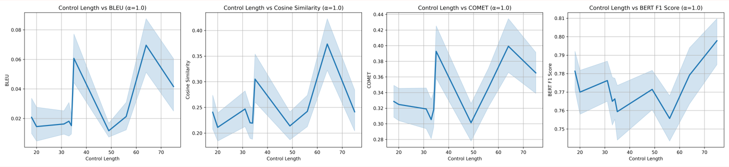


Figure 3. Patching is agnostic of the length of the [control]

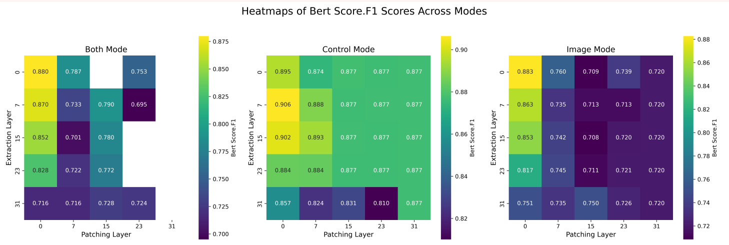
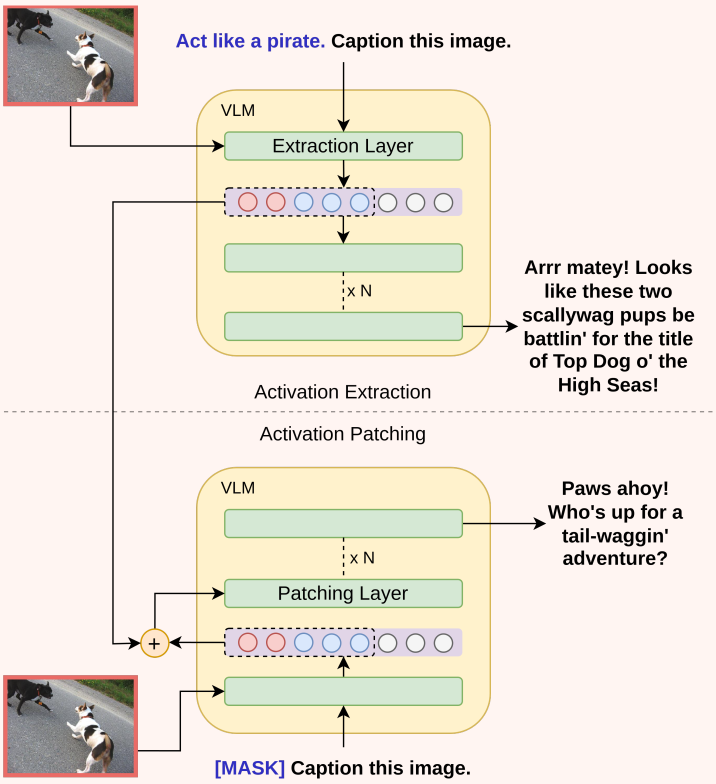


Figure 4. Optimal performance: closely-spaced early layers

Overview



Act like a pirate. Caption this image.

VLM

Extraction Layer

x N

Arrr matey! Looks like these two scallywag pups be battlin' for the title of Top Dog o' the High Seas!

Activation Extraction

Activation Patching

VLM

x N

Paws ahoy! Who's up for a tail-waggin' adventure?

Patching Layer

[MASK] Caption this image.

Figure 5. Architecture of Patch and Control. We first extract the activations of the [control] and [image] tokens from the extraction layer and patch them into the chosen patching layer during inference. A weighted parameter α is used to linearly combine signals from the original and patched activations.

Impact of Patching

Zero-shot performance on different comparison metrics.

Mode	α	BLEU	COMET	Cosine	BERT
[control]	0.25	0.0854	0.6144	0.5570	0.8792
	0.5	0.0959	0.6206	0.5675	0.8811
	0.75	0.0913	0.6018	0.5432	0.8699
	1.0	0.1035	0.6000	0.5441	0.8655
[image]	0.25	0.0065	0.2103	0.1648	0.7593
	0.5	0.0090	0.2304	0.1638	0.7579
	0.75	0.0108	0.2290	0.1627	0.7517
	1.0	0.0095	0.2455	0.1669	0.7503
Both	0.25	0.0110	0.2894	0.2354	0.7917
	0.5	0.0156	0.2904	0.2314	0.7847
	0.75	0.0209	0.2900	0.2104	0.7669
	1.0	0.0288	0.3400	0.2508	0.7726

Table 1. **Performance of Patch and Control across different metrics**. BLEU and COMET evaluate syntactic alignment, while cosine similarity and BERTScore evaluate semantic alignment. **Our results suggest we can transfer conditioning and steer LVLM behavior without prompting or fine-tuning.**

Github: aditeyabaral/activation-patching CSCI-GA 3033 Emerging Topics in NLP ab12057, rd3629, dv2469 @nyu.edu