① **Image-Language Pretraining**

I^{sim}

I^{real}

CNN

f_{img}^{sim} f_{lang}^{sim}

f_{img}^{real} f_{lang}^{real}

LLM

l^{sim} "gripper holding bread over square"

"gripper holding carrot over yellow mat" l^{real}

Language-regularized CNN pretraining variants

A. Regression-based

f_{img}^{sim}

Lang pred. $g(\cdot)$

$\hat{f}_{lang}^{sim}$ f_{lang}^{sim}

$\mathcal{L}_{cnn,reg}$

f_{img}^{real}

$\hat{f}_{lang}^{real}$ f_{lang}^{real}

B. Distance-based

f_{img}^{sim} f_{img}^{real}

d_{img}

f_{lang}^{sim}

d_{lang} f_{lang}^{real}

$\mathcal{L}_{cnn,dist}$

② **Multitask, Multidomain Behavioral Cloning**

Task Instruction: *"Place carrot on yellow mat."*

LLM ❄️

FiLM 🔥

CNN ❄️

Last Layer 🔥

Spatial Softmax

concat

+

FC Layers

1024 512 256

x y z

$[x, y, z, g]$

Robot state