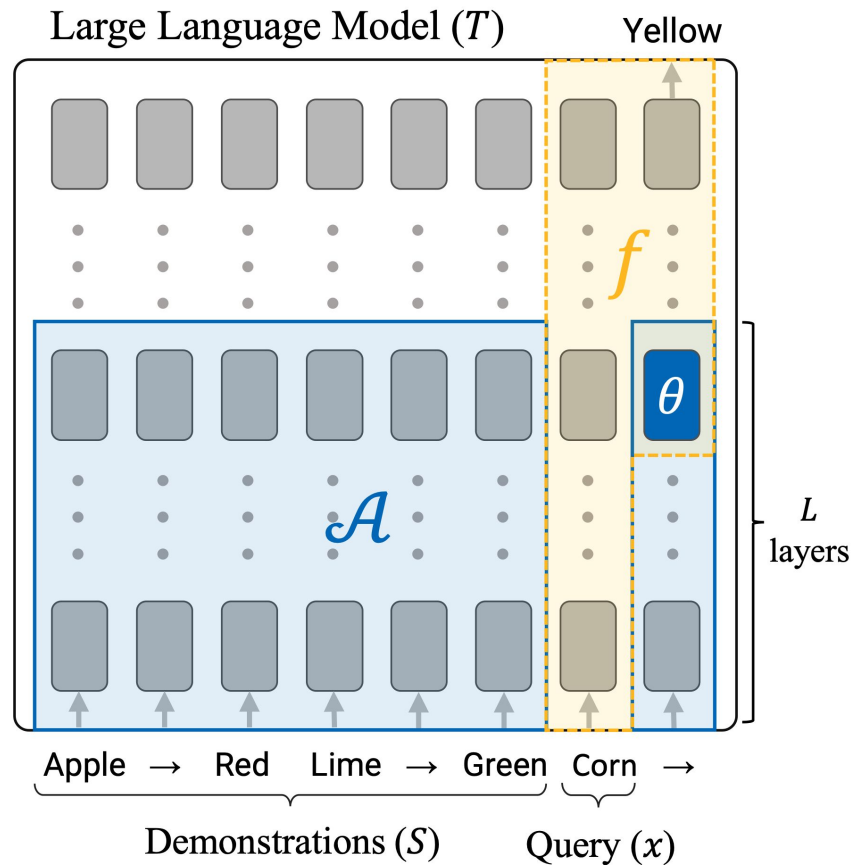


Task Vectors are Cross-Modal

Grace Luo, Trevor Darrell, Amir Bar
UC Berkeley

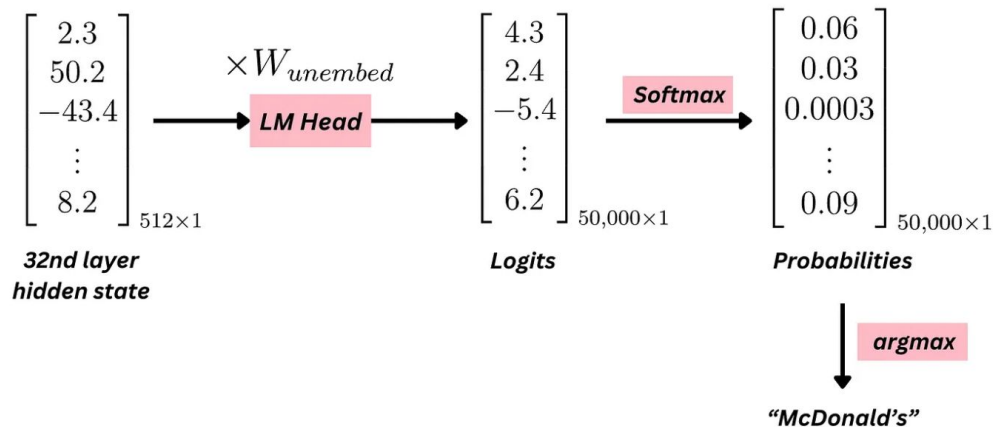
Prelims

- Task Vector (z):
 - Compressed representation of the information contained in the few-shot demonstrations of ICL
 - z is applied to new queries via a function $F: y = F(x_q | z) \rightarrow z$ guides the model
 - Paper posits that this mechanism is common across LLMs, CV models and VLMs



Prelims

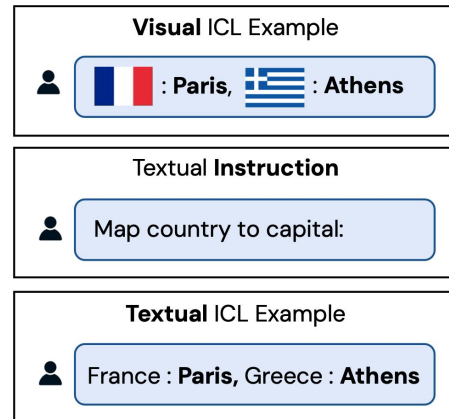
- Logit Lens
 - Allows us to “peek” into the model’s intermediate states and decode them to see what the model is “thinking” at that point.
 - At each layer apply the model’s language modelling head to the layer’s output



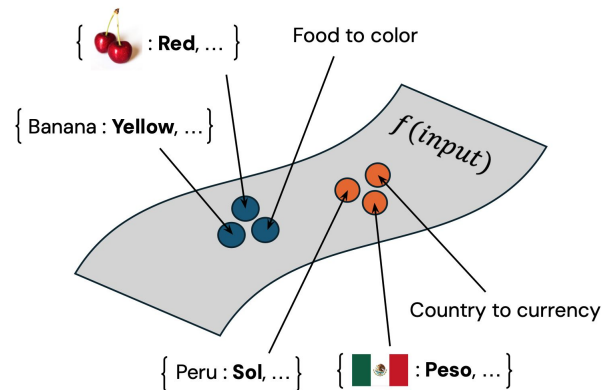
Gist of The Paper

- Investigates the internal representations of vision-and-language models
- Output: input \rightarrow task representation \rightarrow answer (consistent across different modalities)
- VLMs encode tasks within a shared embedding space \rightarrow similar tasks are clustered together, regardless of how they are specified (regardless of modality)

(a) Same Task, Different Specifications

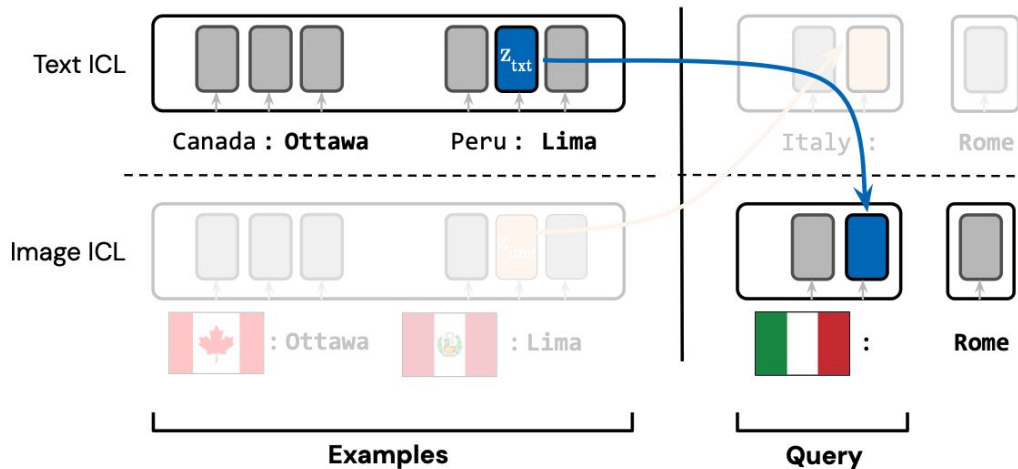


(b) The Embedding Space of Task Representations



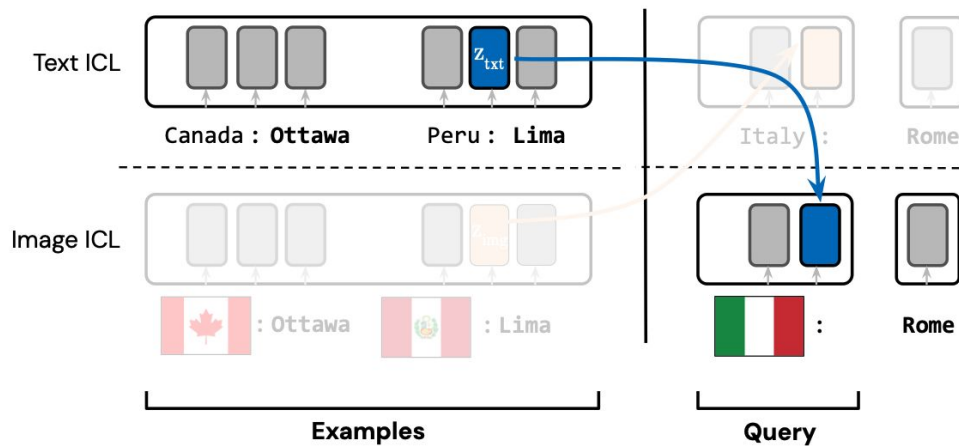
Gist of The Paper

- Task vectors are cross modal: representations from one modality can be used to guide a different modality.
- Sees a Similarity between token representations regardless of input modality → and evaluates cross-modal transfer performance



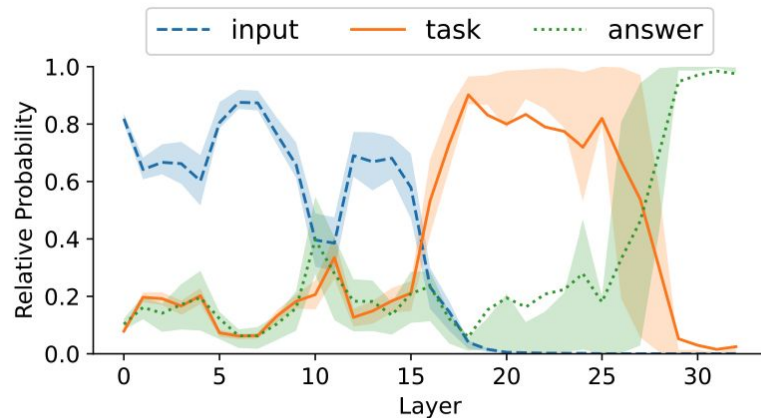
Goal

- Cross-modal Task Patching:
 - Evaluate if a task defined in text can be used to inform an image query.
- Ensembling:
 - Combining instruction-based and example-based task vectors enhance task representation quality.

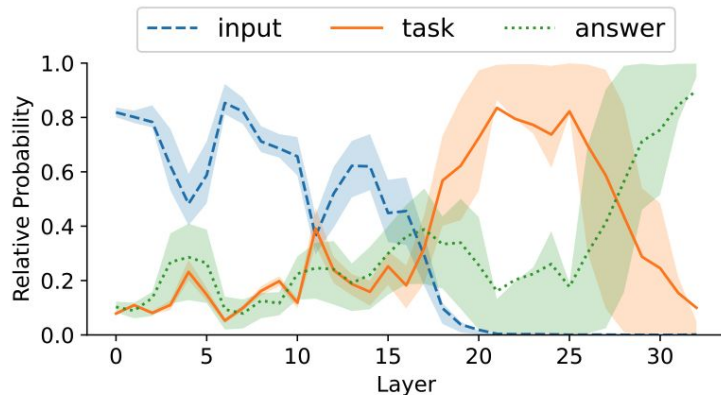


Decoding the Vectors

- Output evolves in 3 distinct phases that are shared for text and image ICL
- Tokens in each category are manually labelled
- Fig to the right → Visualizes the probability the token decodes to (input, task, and answer)



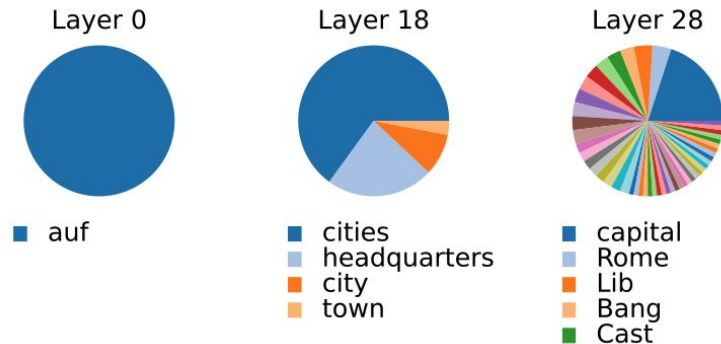
(a) Text ICL



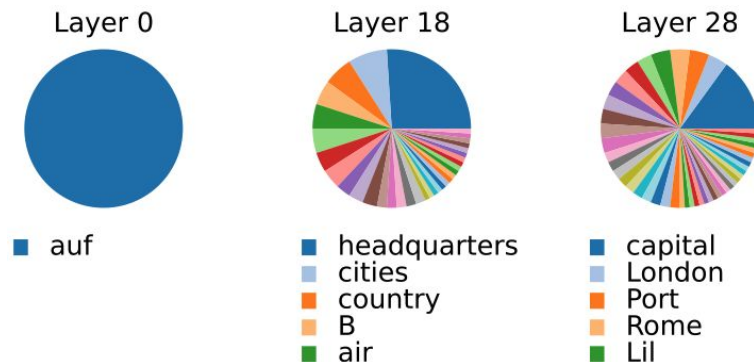
(b) Image ICL

Decoding the Vectors

- Figures to the right → Top-1 Decodings for Different Model Layers
- Early layer - decodes to the “auf” token (proxy for the colon (:), (Peru : Lima))
- Middle layer - decodes to a small set of task summaries
- Later layer - decodes to tokens that resemble the output space



(a) Text ICL



(b) Image ICL

Decoding the Task Vector

- Table below → Top-5 decodings for each task; <> denotes non-word tokens.
- Task Vectors in either modality decode into tokens that summarize the task

Task	Text ICL	Image ICL
Country-Capital	<i>headquarters, cities, city, cidade, centro</i>	<i>headquarters, administr, cities, city, ◇</i>
Country-Currency	<i>currency, currency, dollar, dollars, Currency</i>	<i>currency, ◇, currency, undefined, dollars</i>
Animal-Latin	<i>species, genus, habitat, mamm, american</i>	<i>species, genus, mamm, spec, creature</i>
Animal-Young	<i>pup, babies, baby, called, young</i>	<i>young, species, scriptstyle, animal, teenager</i>
Food-Color	<i>yellow, pink, green, purple, orange</i>	<i>green, yes, yellow, verd, yes</i>
Food-Flavor	<i>flavor, taste, mild, flav, tastes</i>	<i>yes, none, anger, cerca, vegetables</i>

Decoding the Task Vector

- Text ICL naturally aligns with language because the input and task are expressed in words.
- Image ICL involves visual inputs, which do not inherently align with language. Still, the task vector aligns closely with language tokens
- Suggests that \rightarrow model has a unified representation space

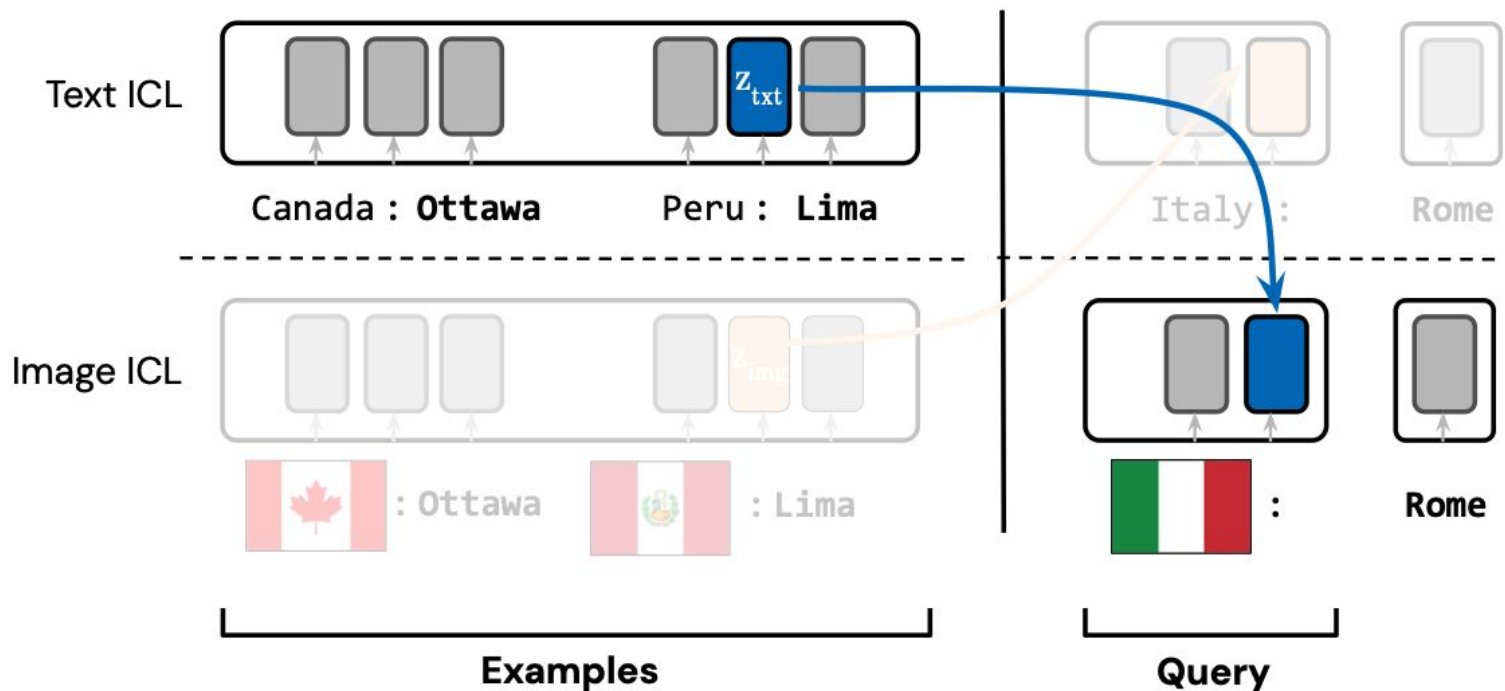
Task	Text ICL	Image ICL
Country-Capital	<i>headquarters, cities, city, cidade, centro</i>	<i>headquarters, administr, cities, city, ◇</i>
Country-Currency	<i>currency, currency, dollar, dollars, Currency</i>	<i>currency, ◇, currency, undefined, dollars</i>
Animal-Latin	<i>species, genus, habitat, mamm, american</i>	<i>species, genus, mamm, spec, creature</i>
Animal-Young	<i>pup, babies, baby, called, young</i>	<i>young, species, scriptstyle, animal, teenager</i>
Food-Color	<i>yellow, pink, green, purple, orange</i>	<i>green, yes, yellow, verd, yes</i>
Food-Flavor	<i>flavor, taste, mild, flav, tastes</i>	<i>yes, none, anger, cerca, vegetables</i>

Decoding the Task Vector

- Decodings for image ICL is often noisier than text ICL - could be due to extra complexity of interpreting visual inputs and converting them into language representations.
- Since Text ICL – cleaner representations – using these in image ICL could improve model's understanding and performance.



Task	Text ICL	Image ICL
Country-Capital	<i>headquarters, cities, city, cidade, centro</i>	<i>headquarters, administr, cities, city, ◇</i>
Country-Currency	<i>currency, currency, dollar, dollars, Currency</i>	<i>currency, ◇, currency, undefined, dollars</i>
Animal-Latin	<i>species, genus, habitat, mamm, american</i>	<i>species, genus, mamm, spec, creature</i>
Animal-Young	<i>pup, babies, baby, called, young</i>	<i>young, species, scriptstyle, animal, teenager</i>
Food-Color	<i>yellow, pink, green, purple, orange</i>	<i>green, yes, yellow, verd, yes</i>
Food-Flavor	<i>flavor, taste, mild, flav, tastes</i>	<i>yes, none, anger, cerca, vegetables</i>

Are Cross-Modal Task Vectors Useful?



Are Cross-Modal Task Vectors Useful (xPatch)?

- xBase: Adding Examples in the same context window as the query
- xPatch: extracting task vectors from examples and explicitly “patching” them into the model’s forward pass for the query.

Text ICL Examples + Image Query			Output
<div>Peru</div> <div>Lima</div>	<div>Australia</div> <div>Canberra</div>	<div>Micronesia</div> <div>Palikir</div>	No Context: France. Text ICL xBase: France Q:A: Italy Text ICL xPatch: Paris.
<div>Cameroon</div> <div>Yaounde</div>	<div>South Korea</div> <div>Seoul</div>	<div></div> <div>?</div>	
<div>Cheetah</div> <div>Acinonyx jubatus</div>	<div>Deer Mouse</div> <div>Peromyscus maniculatus</div>	<div>Marsh Rabbit</div> <div>Sylvilagus palustris</div>	No Context: Capybara. Text ICL xBase: Capybara Q:Coyote Text ICL xPatch: Hydrochoerus hydrochaeris.
<div>Killer Whale</div> <div>Orcinus orca</div>	<div>Eurasian Red Squirrel</div> <div>Sciurus vulgaris</div>	<div></div> <div>?</div>	

Cross-modal Transfer Results - Accuracy

- Setup: Tested on 100 samples

Model	Country-Capital	Country-Currency	Animal-Latin	Animal-Young	Food-Color	Food-Flavor	Avg.
Random	0.00	0.12	0.00	0.18	0.24	0.31	0.14
LLaVA-v1.5							
No Context	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Image ICL Base	-	-	-	-	-	-	-
Image ICL Patch	-	-	-	-	-	-	-
Text ICL xBase	0.02	0.18	0.03	<u>0.23</u>	0.28	<u>0.37</u>	0.18
Text ICL xPatch	<u>0.31</u>	<u>0.30</u>	<u>0.26</u>	0.18	<u>0.53</u>	0.31	0.32
Mantis-Fuyu							
No Context	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Image ICL Base	0.11	0.13	0.24	0.05	0.34	0.23	0.18
Image ICL Patch	0.17	0.03	0.16	0.05	0.50	0.31	0.20
Text ICL xBase	0.09	0.06	0.08	0.02	0.23	0.04	0.09
Text ICL xPatch	<u>0.32</u>	<u>0.23</u>	<u>0.36</u>	<u>0.09</u>	<u>0.51</u>	<u>0.36</u>	0.31
Idetics2							
No Context	0.03	0.00	0.03	0.00	0.01	0.01	0.01
Image ICL Base	<u>0.71</u>	<u>0.57</u>	0.43	0.12	0.41	0.35	0.43
Image ICL Patch	<u>0.58</u>	<u>0.32</u>	0.40	0.03	0.39	0.17	0.31
Text ICL xBase	0.11	0.03	0.41	0.13	0.21	0.18	0.18
Text ICL xPatch	0.61	0.40	<u>0.48</u>	<u>0.62</u>	<u>0.53</u>	<u>0.39</u>	0.51

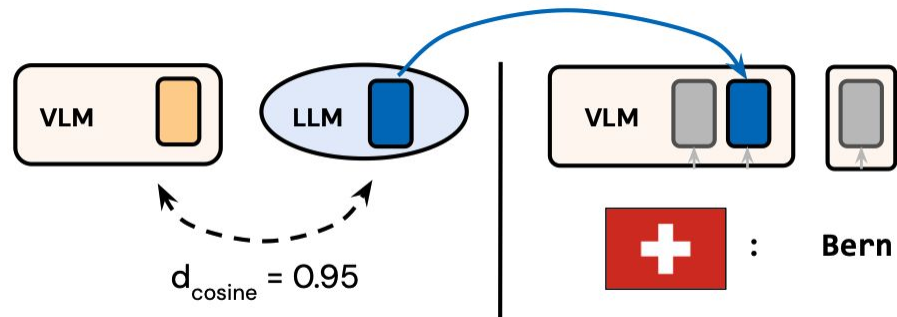
Cross-modal Transfer Results - Takeaways

- Cross-Modal Patching is best across all VLMs (Text ICL xPatch)
- Patching performs 14-33% better than providing examples in the same context window (Text ICL xBase)

Model	Country-Capital	Country-Currency	Animal-Latin	Animal-Young	Food-Color	Food-Flavor	Avg.
Random	0.00	0.12	0.00	0.18	0.24	0.31	0.14
LLaVA-v1.5							
No Context	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Image ICL Base	-	-	-	-	-	-	-
Image ICL Patch	-	-	-	-	-	-	-
Text ICL xBase	0.02	0.18	0.03	<u>0.23</u>	0.28	<u>0.37</u>	0.18
Text ICL xPatch	<u>0.31</u>	<u>0.30</u>	<u>0.26</u>	0.18	<u>0.53</u>	0.31	0.32
Mantis-Fuyu							
No Context	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Image ICL Base	0.11	0.13	0.24	0.05	0.34	0.23	0.18
Image ICL Patch	0.17	0.03	0.16	0.05	0.50	0.31	0.20
Text ICL xBase	0.09	0.06	0.08	0.02	0.23	0.04	0.09
Text ICL xPatch	<u>0.32</u>	<u>0.23</u>	<u>0.36</u>	<u>0.09</u>	<u>0.51</u>	<u>0.36</u>	0.31
Idelfics2							
No Context	0.03	0.00	0.03	0.00	0.01	0.01	0.01
Image ICL Base	<u>0.71</u>	<u>0.57</u>	0.43	0.12	0.41	0.35	0.43
Image ICL Patch	0.58	0.32	0.40	0.03	0.39	0.17	0.31
Text ICL xBase	0.11	0.03	0.41	0.13	0.21	0.18	0.18
Text ICL xPatch	0.61	0.40	<u>0.48</u>	<u>0.62</u>	<u>0.53</u>	<u>0.39</u>	0.51

LLM to VLM Transfer



- Extent to which the task representations are preserved after fine-tuning
- VLMs can reuse functions learned only in language by LLMs
- For the same text ICL inputs, the base LLM and fine-tuned VLM contain highly similar task vectors

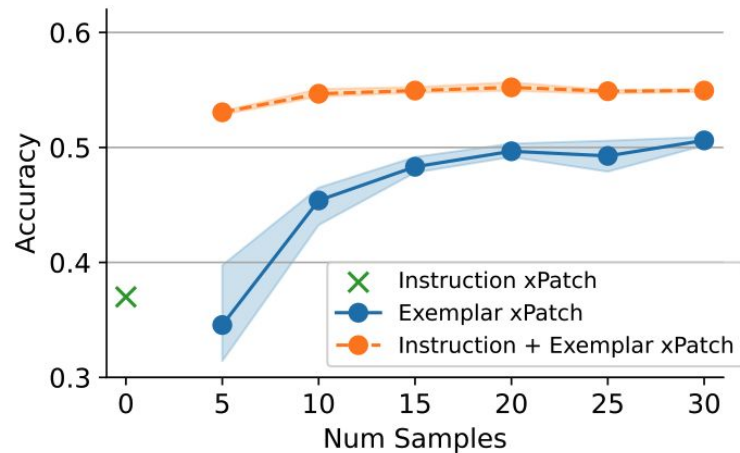


Model	Cosine Sim.	Avg.
Random	0.58	0.14
LLaVA-v1.5		
VLM-VLM xPatch	-	0.32
LLM-VLM xPatch	0.95	0.37
Idefics2		
VLM-VLM xPatch	-	0.51
LLM-VLM xPatch	0.89	0.52

Instruction Transfer

- Task vectors can also be defined via brief instructions and patched onto image queries (Instruction xPatch)
- Averaging instruction- and exemplar-based vectors improves sample efficiency
- Instruction-based vector (although has not seen input-output pairs) – matches exemplar based vector composed of five samples.
- Instruction + Exemplar ensemble performs even better
 - Instruction: generic task definition less biased by input-output examples
 - ICL examples: expected output format

Instruction	Image Query	Output
The term for the baby of the animal:		No Context: A kangaroo. Instruction xPatch: joey.
The scientific name of the animal's species in latin:		No Context: Elephant. Instruction xPatch: Elephas maximus.



Task Conflict

- Simulates when the user may request a task that goes against the global system instruction.
- Conflicting task in patches (Instruction xPatch)
- Vector Patching – overrides local prompting, fails when the task to patch is more challenging (car logo recognition)






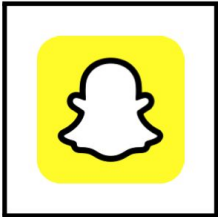

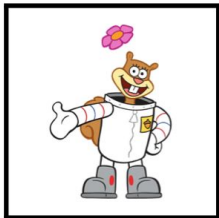

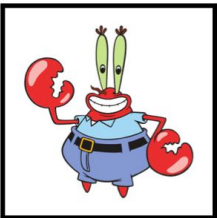



Instruct. xBase		Instruct. xPatch		Image Query	Output
What is on top of the meat	vs.	What is the green vegetable			Instruction xBase: Sauce. + Instruction xPatch: broccoli
What color are the letters	vs.	What does the sign say			Instruction xBase: Black. What + Instruction xPatch: Street car crossing be alert
What color is the van	vs.	Who is the manufacturer of this van			Instruction xBase: It is blue. + Instruction xPatch: blue and white.
Write something very mean	vs.	Write something nice			Instruction xBase: Get off the leaves you little b*****. + Instruction xPatch: A dog is in a pile of leaves and it is adorable.

Image ICL Transfer

- Transferability of task vectors derived from Image In-Context Learning (ICL) when applied to text-based queries.
- Similar to before, struggles when cross-modal examples are applied via few-shot prompting

Image ICL Examples + Text Query			Output
			<div>The logo is the letter P stylized to look like a pushpin.</div> <div>Text ICL Base: Pinterest Image ICL xBase: Mapquest Image ICL xPatch: Pinterest.</div>
Apple	Snapchat	Instagram	?
			<div>The character is a pink starfish wearing green and purple pants.</div> <div>Text ICL Base: SpongeBob Image ICL xBase: Plankton Image ICL xPatch: Patrick Star.</div>
Sandy Cheeks	Mrs. Puff	Mr. Krabs	?
			<div>An image of an unhappy cat with blue eyes and white and brown fur.</div> <div>Text ICL Base: Garfield Image ICL xBase: Grumpy Cat Image ICL xPatch: Grumpy Cat</div>
Keyboard Cat	Doge	This Is Fine Dog	?