VLMs underpreform on visual data compared to text.

Why?

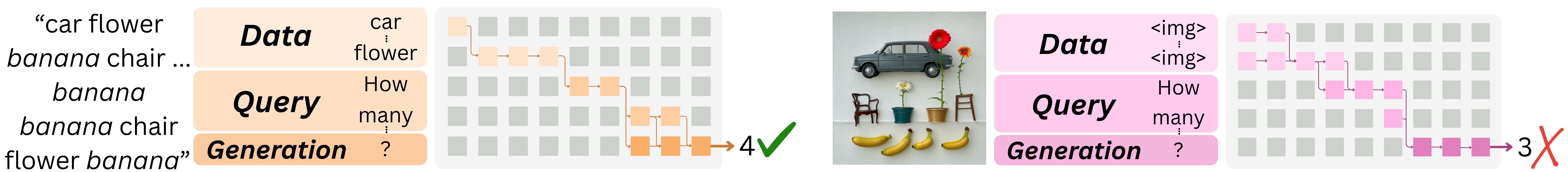
Same Task, Different Circuits: Disentangling Modality-Specific Mechanisms in VLMs

Yaniv Nikankin, Dana Arad, Yossi Gandelsman, Yonatan Belinkov
yaniv.n@campus.technion.ac.il

TECHNION Israel Institute of Technology
BAIR BERKELEY ARTIFICIAL INTELLIGENCE RESEARCH

NEURAL INFORMATION PROCESSING SYSTEMS

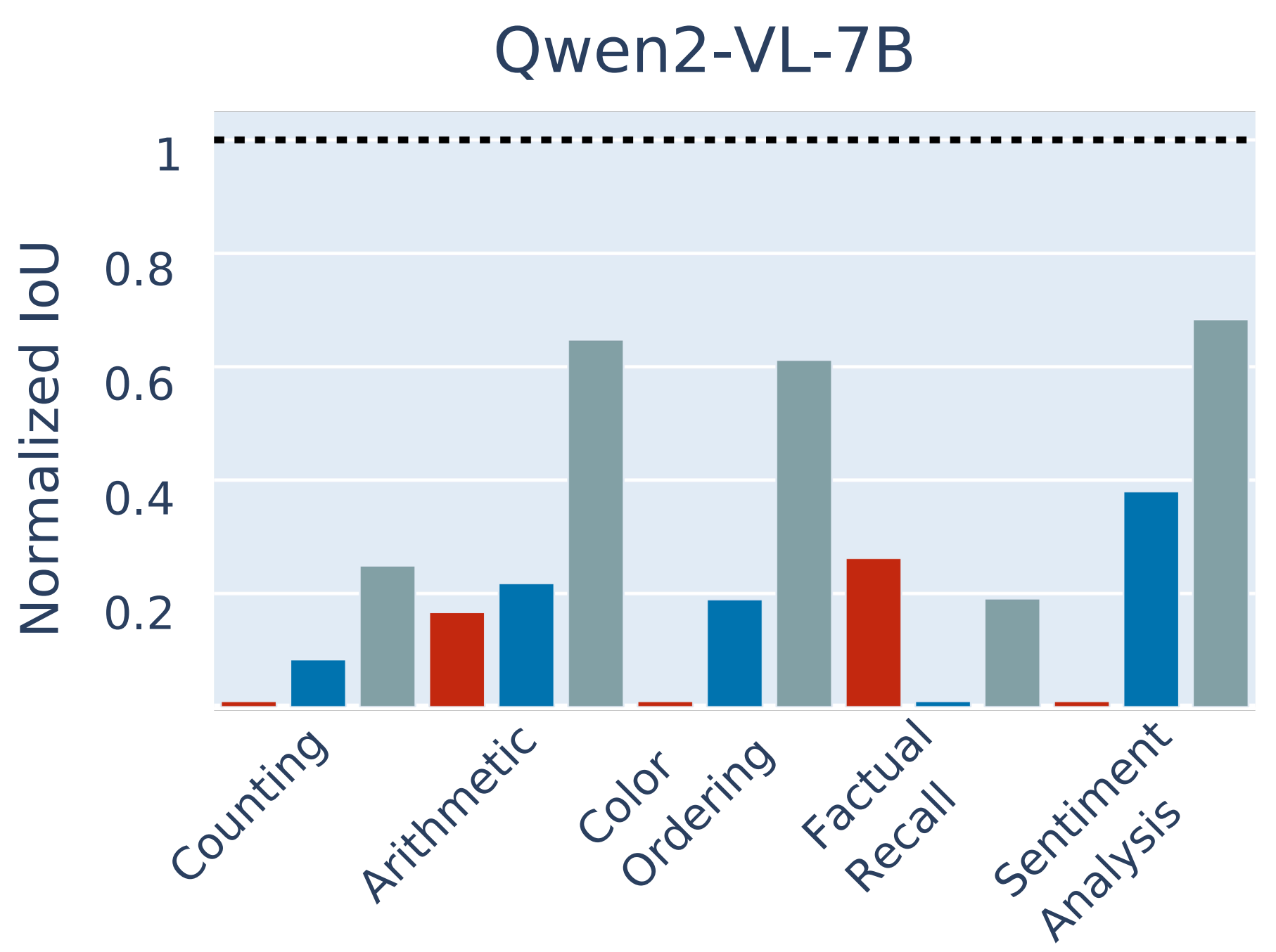
Q: *How many "banana" are in the sequence/image?*



We compare **circuits** — task-specific computational sub-graphs — used to solve textual and visual variants of the same task.

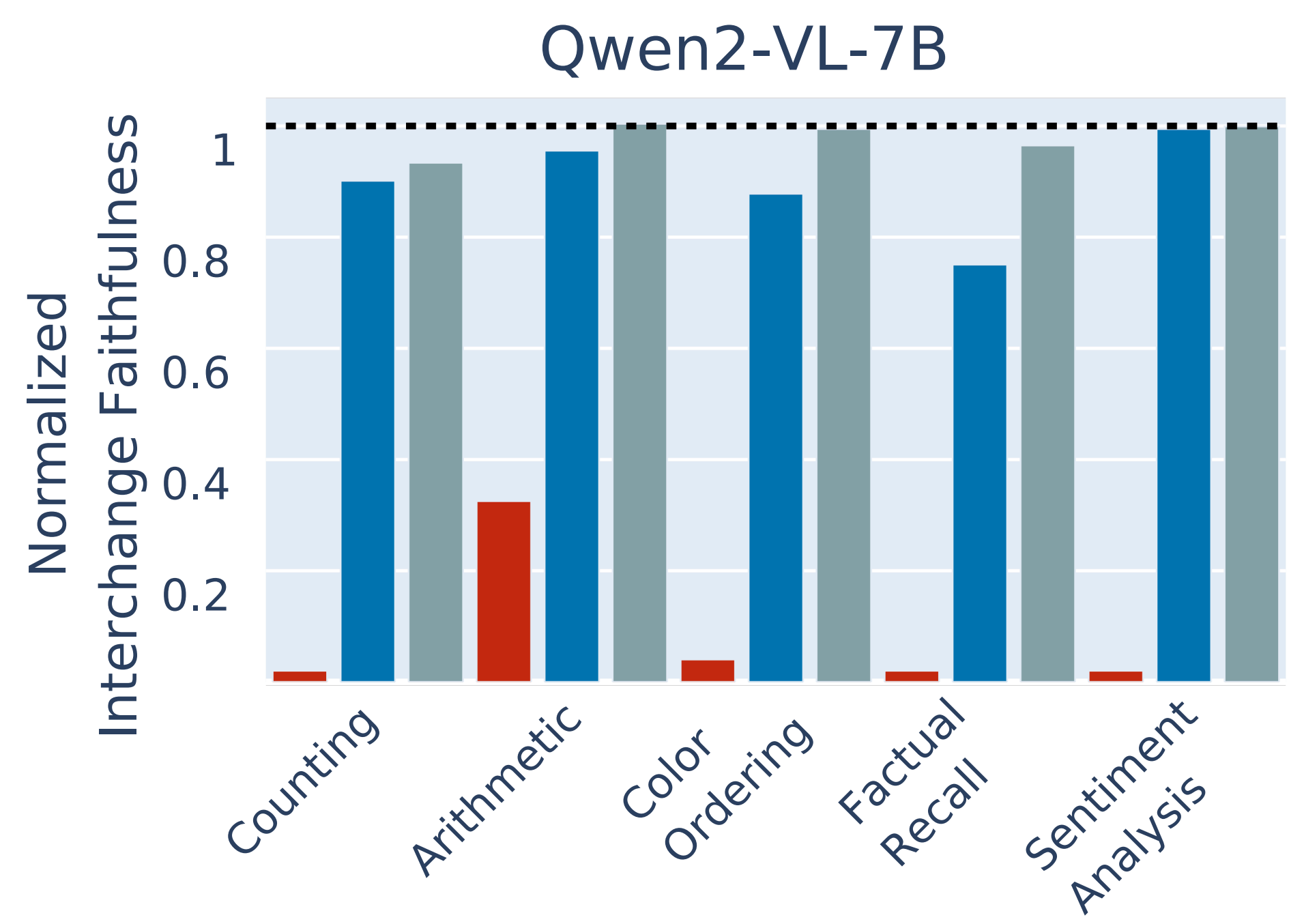
(1) VLMs use **different** computation paths for vision and language variants of the same task.

We find low overlap in **data** and **query** positions, and some overlap in the **generation** position, calculated by IoU of circuit components.

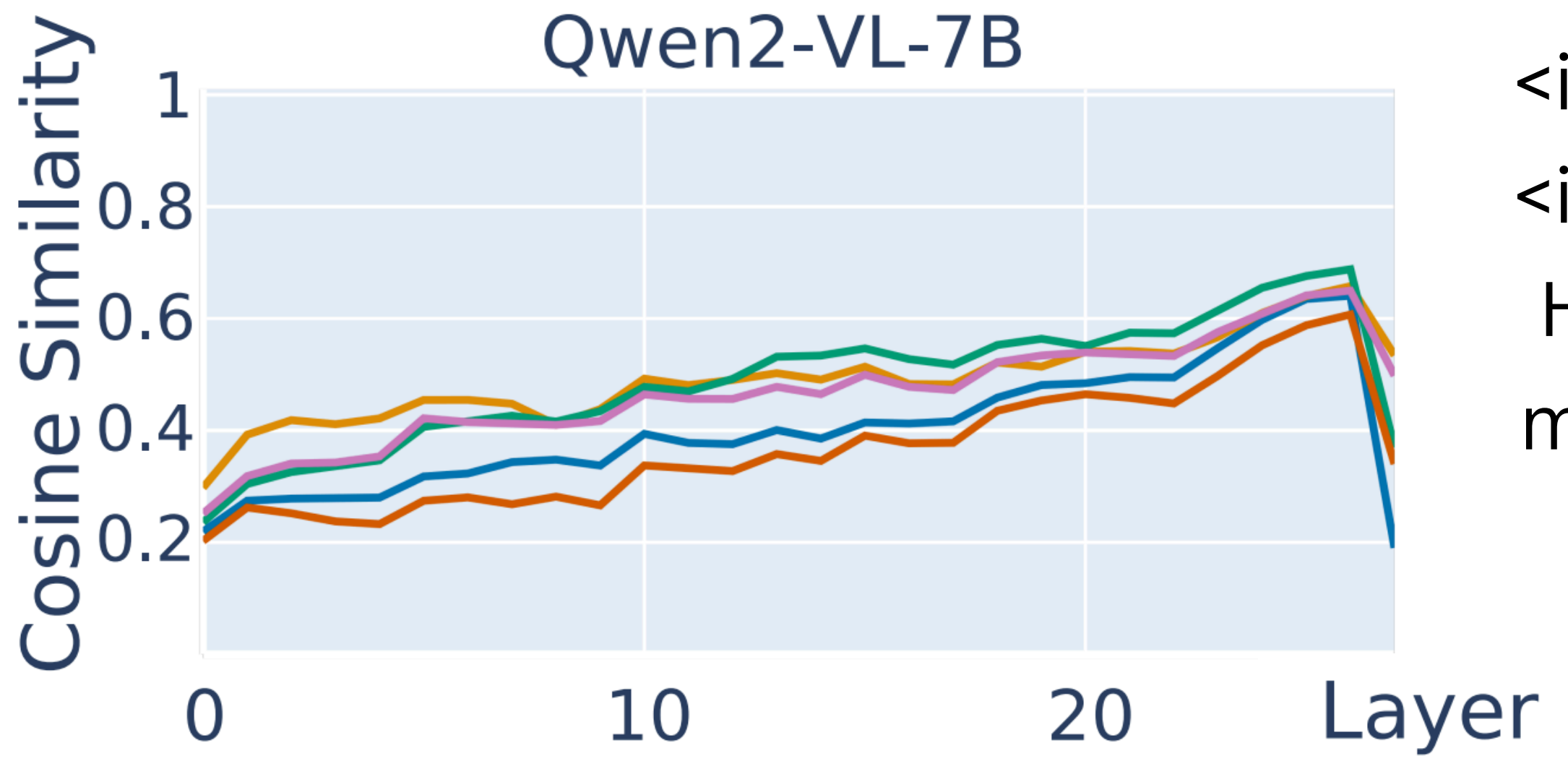


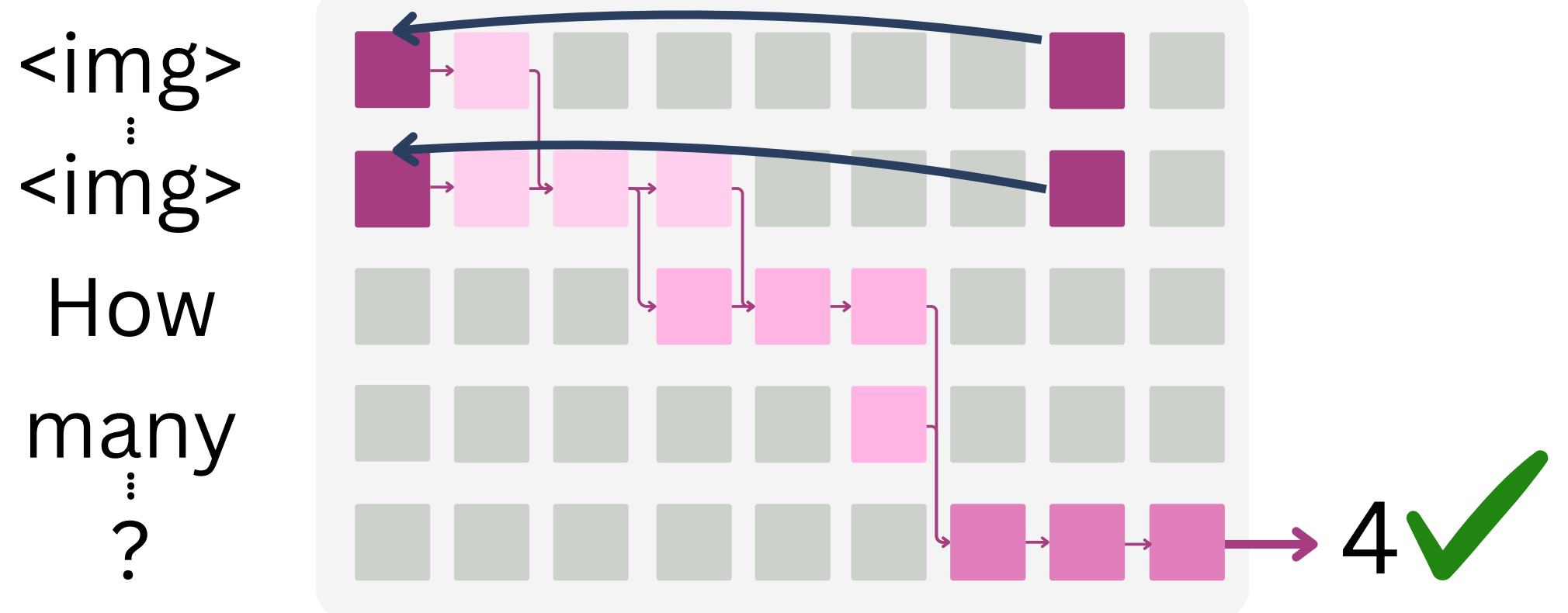
(2) **Query** and **Generation** components implement **similar** functionality across modalities; **Data** components are modality-specific.

We swap components at matched positions and re-evaluate faithfulness, quantifying how much behavior is preserved.



(3) **Visual** data tokens align with **text representations** only in late layers, too late to influence subsequent positions. Back-patching closes a third of the performance gap!





Model	Counting	Arithmetic	Sentiment Analysis
Qwen2-7B-VL	+4.2%	+9.4%	+2.1%
Pixtral-12B	+1.5%	+7.6%	+17.8%
Gemma-3-12B	+2.9%	+6.7%	+5.5%

Funded by the European Union (ERC, Control-LM,101165402). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.