LaVCa: LLM-assisted Visual Cortex Captioning

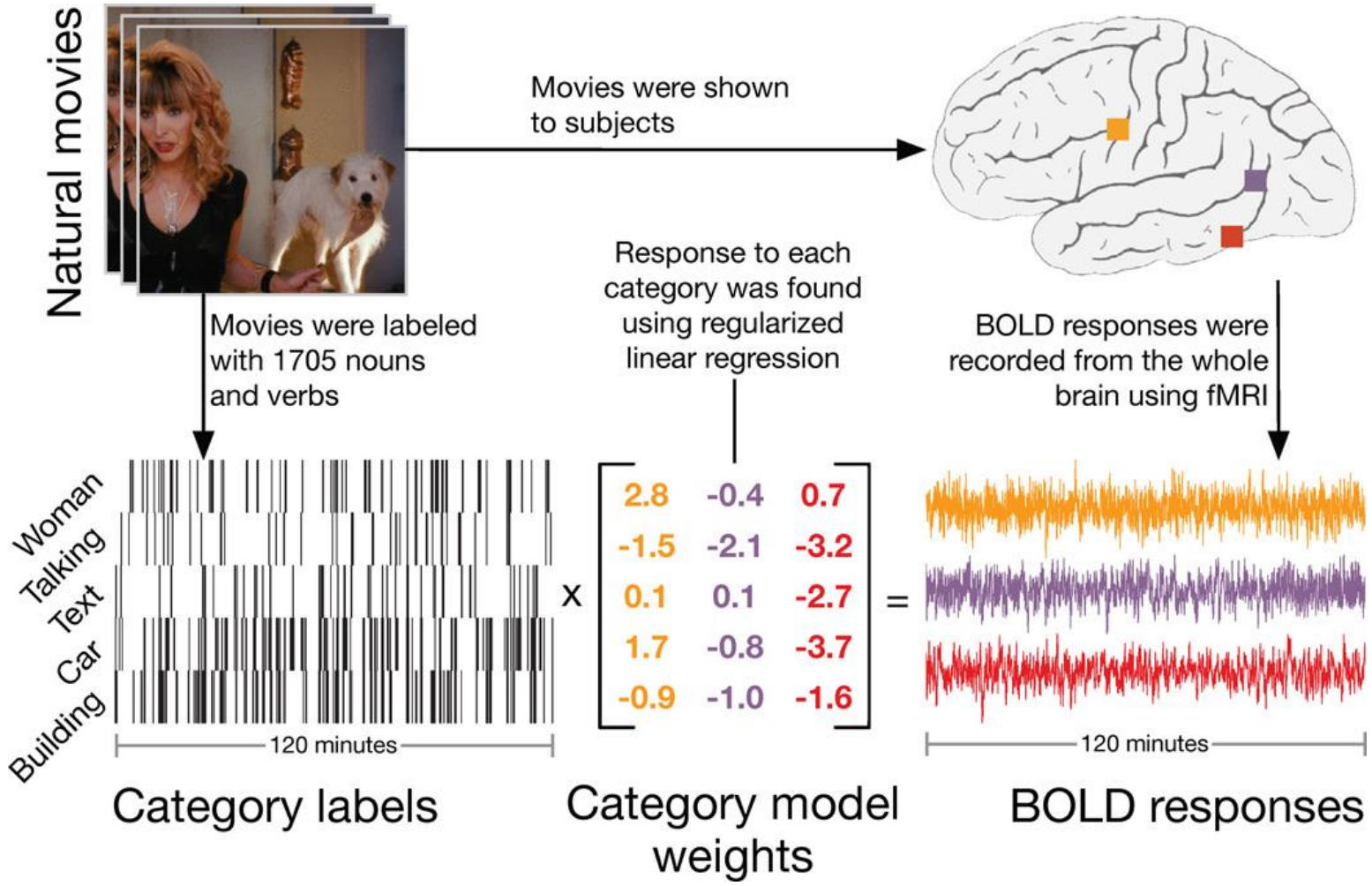
ICLR 2026

Takuya Matsuyama[1,2], **Shinji Nishimoto**[1,2*], **Yu Takagi**[1,2,3*]

[1]University of Osaka, Japan [2]National Institute of Information and Communications Technology, Japan
[3]Nagoya Institute of Technology, Japan

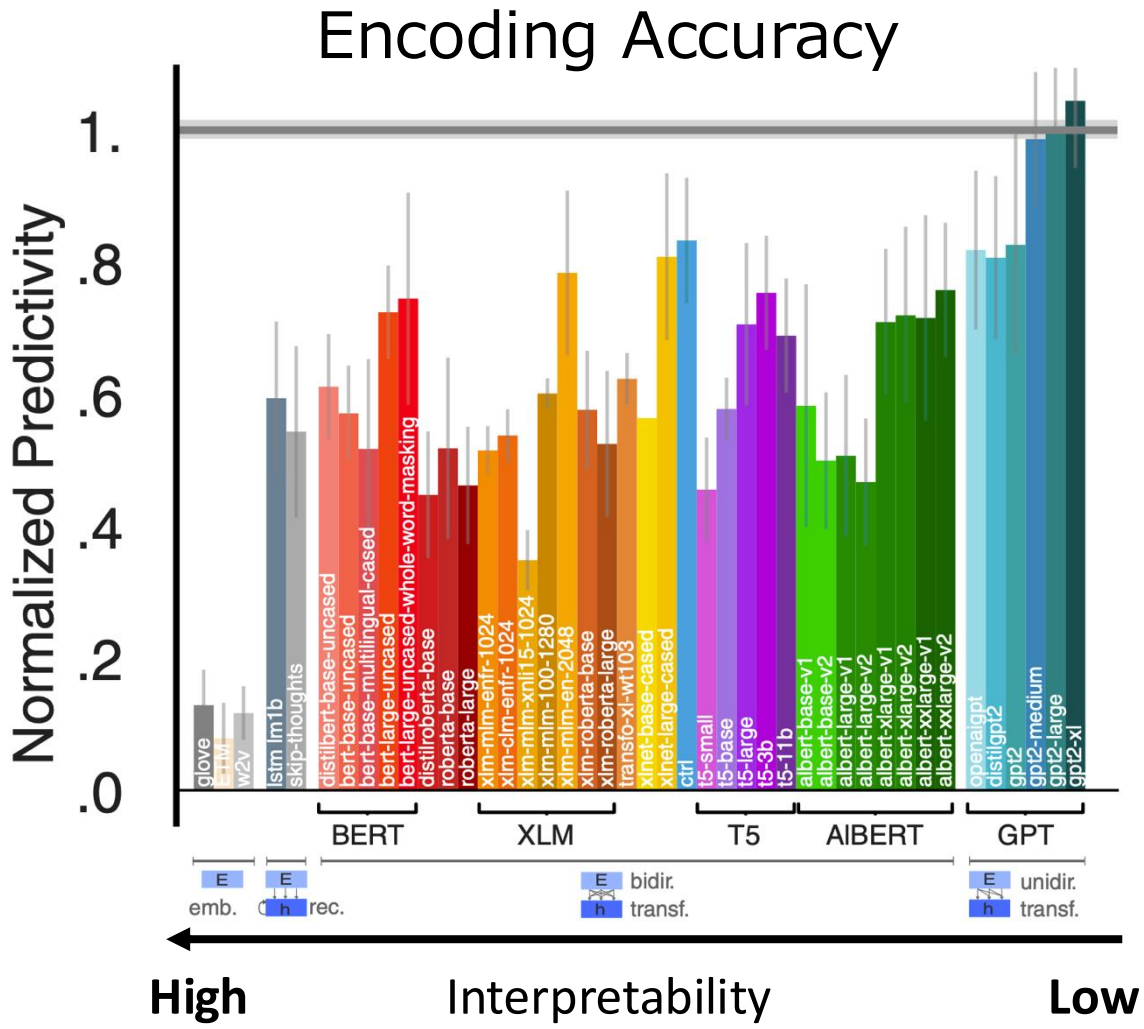
*Equal last author.

Encoding models have been used to elucidate visual representations in the human brain.



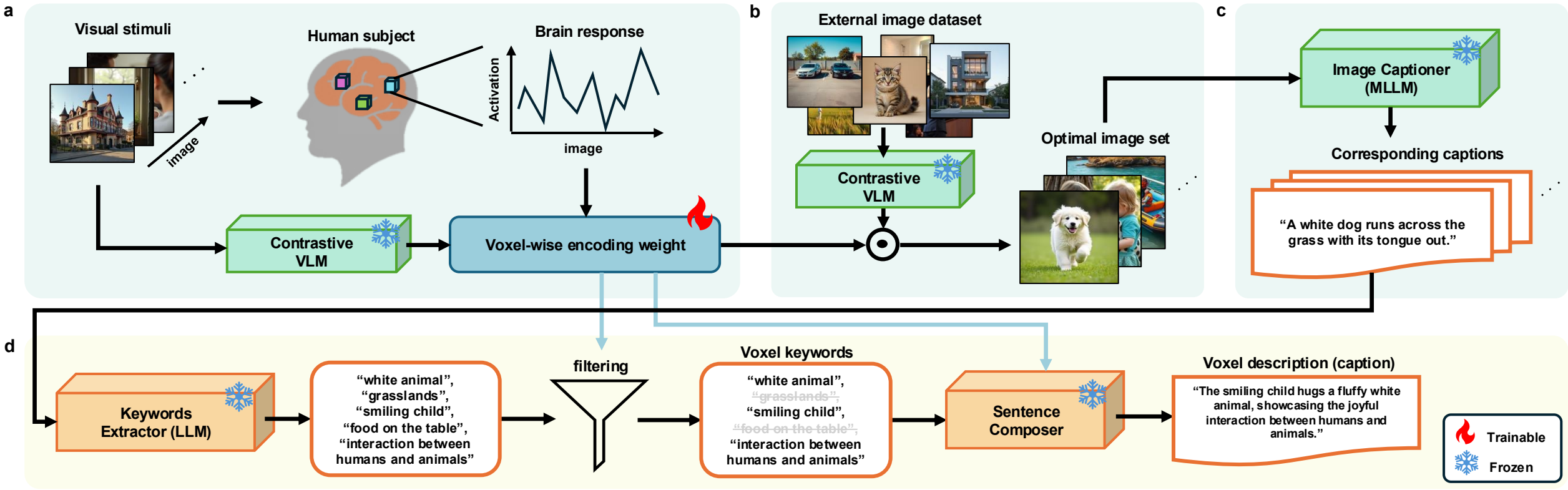
Huth et al. (2011)

DNN-based encoding models achieve high prediction accuracy, but their black-box nature makes them difficult to interpret.



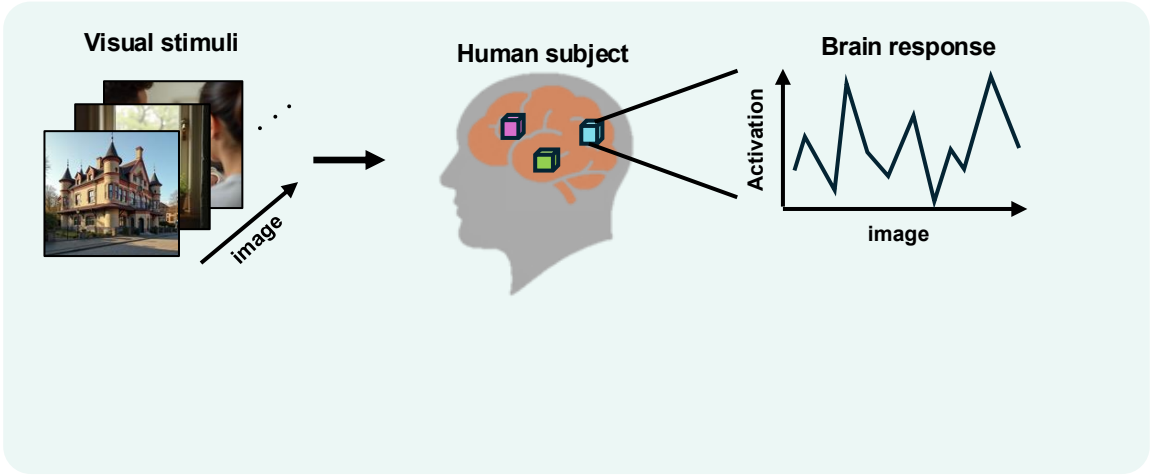
Adapted from Schrimpf et al. (2021)

In this study, we propose a method to interpret DNN-based encoding models using large language models (LLMs).



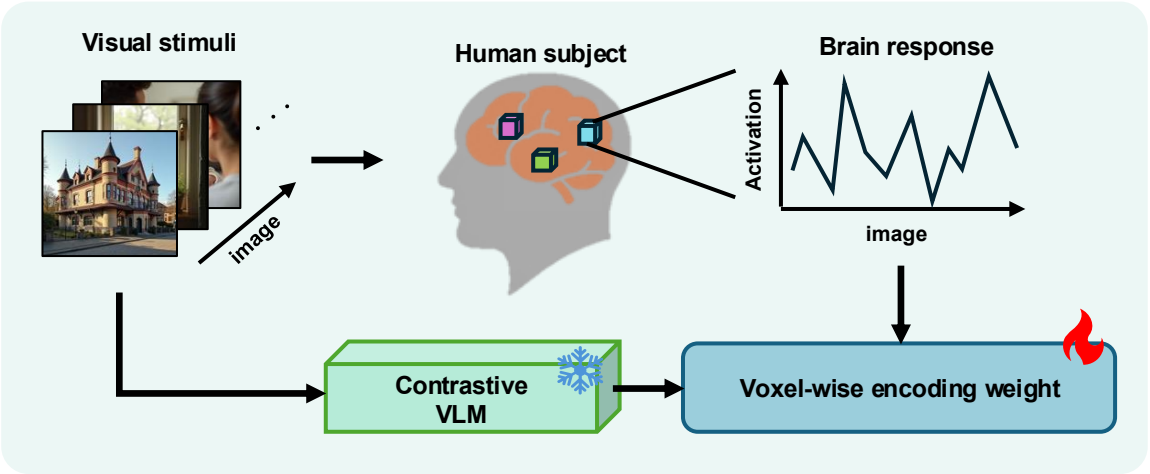
a. Construct voxel-wise encoding models

a

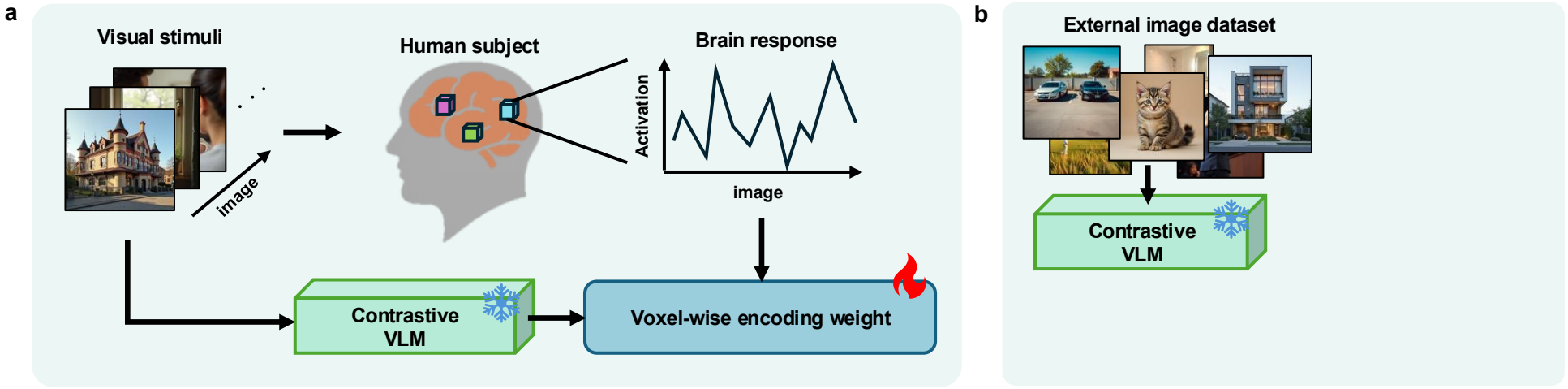


a. Construct voxel-wise encoding models

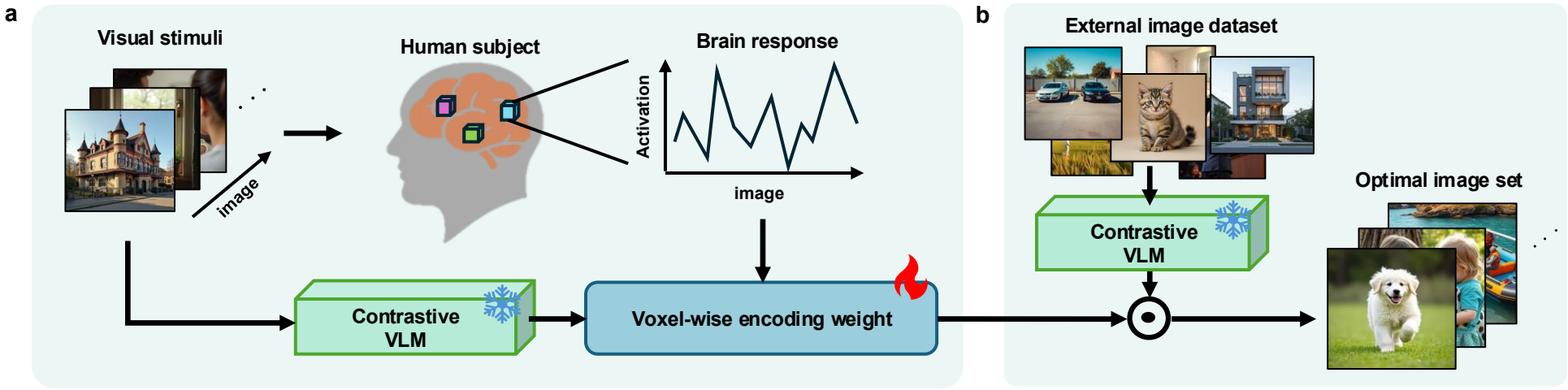
a



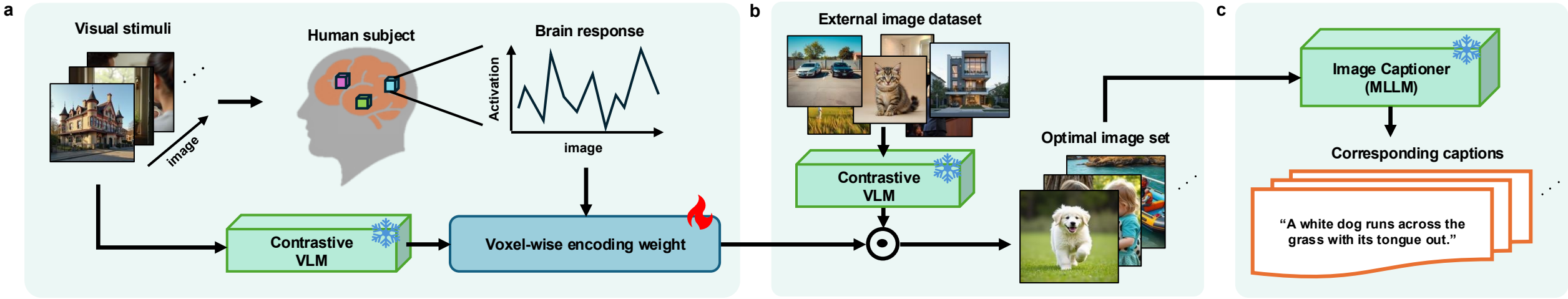
b. Identify the optimal image set by finding the top-N images that most strongly activate each voxel



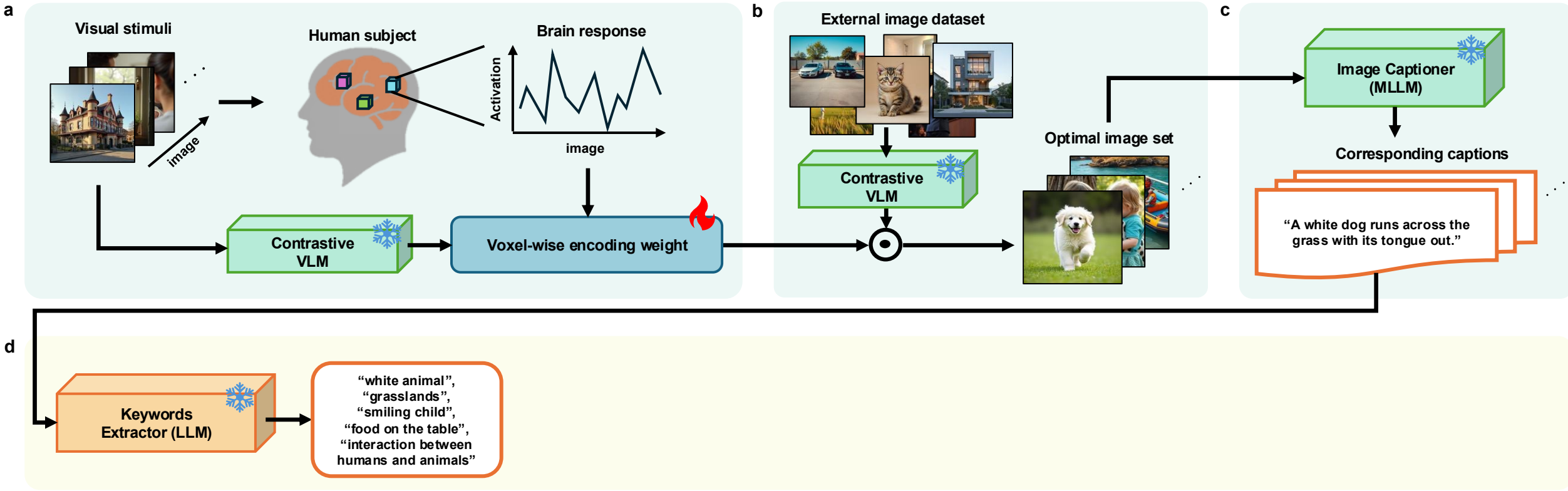
b. Identify the optimal image set by finding the top-N images that most strongly activate each voxel



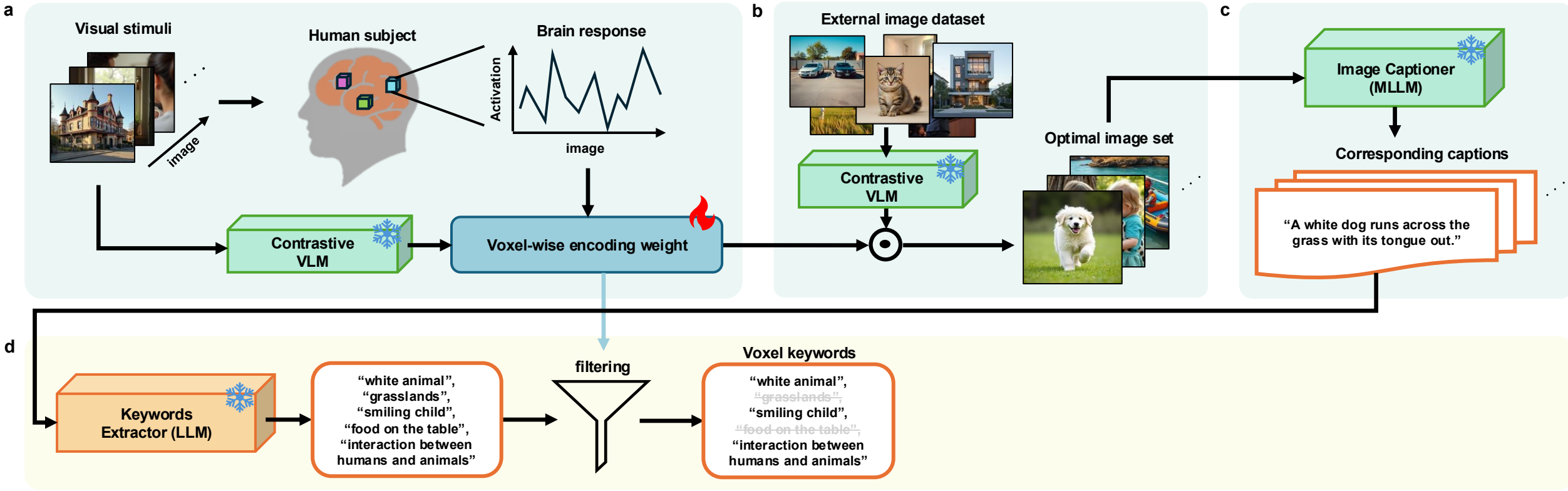
c. Generate captions for these optimal images using a multimodal large language model (MLLM) for summarization



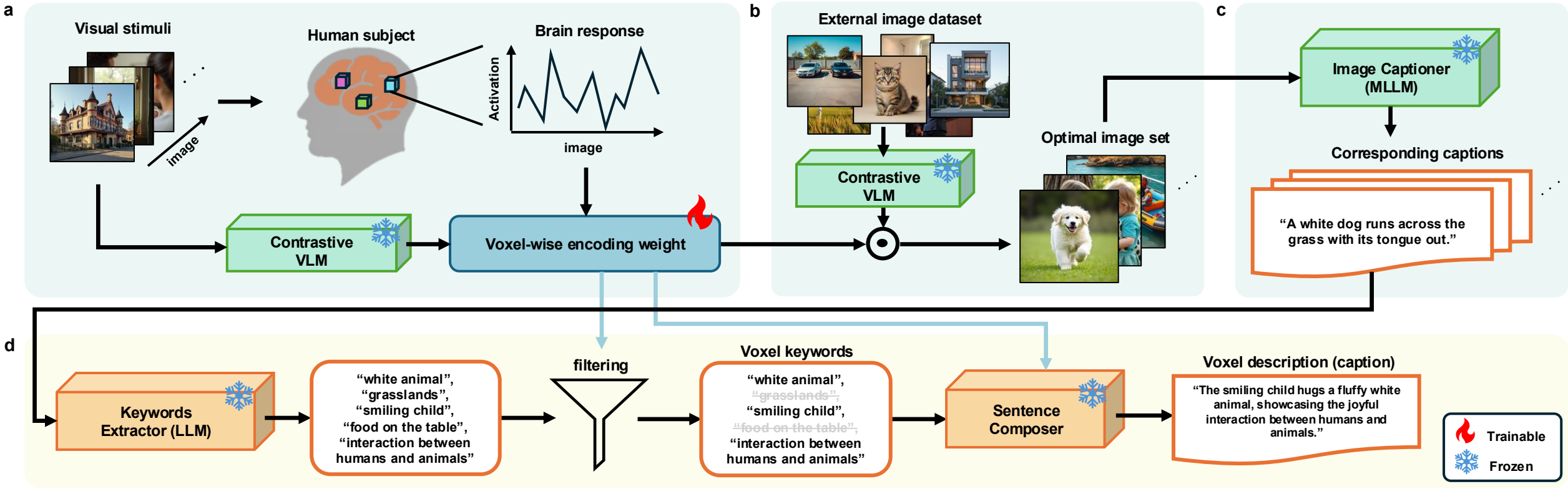
d. Derive concise voxel captions by extracting and filtering keywords from the image captions, then feeding these keywords into a “Sentence Composer.”



d. Derive concise voxel captions by extracting and filtering keywords from the image captions, then feeding these keywords into a “Sentence Composer.”

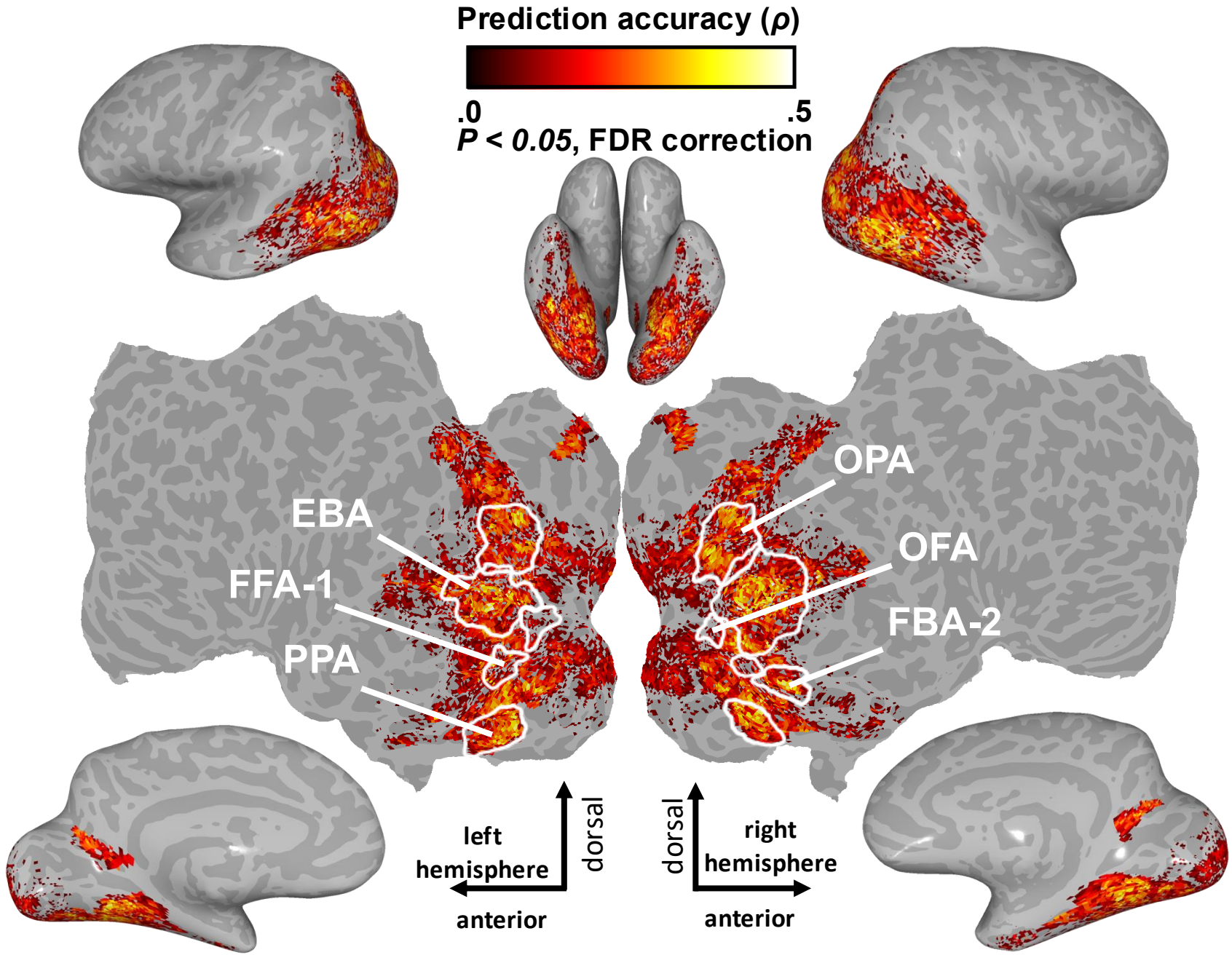


d. Derive concise voxel captions by extracting and filtering keywords from the image captions, then feeding these keywords into a “Sentence Composer.”



Results

Voxel captions enabled prediction across the entire visual cortex.



Our method outperformed existing methods

Model	# keywords	Sentence Composer	Sentence level subj01	subj02	subj05	subj07
Shuffled	–	–	0.007 ± 0.199	0.058 ± 0.223	0.068 ± 0.243	0.009 ± 0.175
BrainSCUBA	–	–	0.207 ± 0.062	0.251 ± 0.071	0.264 ± 0.084	0.182 ± 0.065
LaVCa (Ours)	1	✗	0.205±0.068	0.250±0.075	0.272±0.086	0.186±0.072
LaVCa (Ours)	5	✓	**0.246 ± 0.066**	**0.287 ± 0.075**	**0.306 ± 0.084**	**0.218 ± 0.073**

Model	# keywords	Sentence Composer	Image level subj01	subj02	subj05	subj07
Shuffled	–	–	0.017 ± 0.163	0.052 ± 0.185	0.066 ± 0.204	0.009 ± 0.149
BrainSCUBA	–	–	0.188 ± 0.067	0.226 ± 0.070	0.250 ± 0.078	0.169 ± 0.069
LaVCa (Ours)	1	✗	0.182 ± 0.063	0.221 ± 0.066	0.252 ± 0.077	0.158 ± 0.064
LaVCa (Ours)	5	✓	**0.213 ± 0.072**	**0.250 ± 0.070**	**0.273 ± 0.079**	**0.187 ± 0.073**

The model generated diverse captions for the Occipital Face Area (OFA).

"This is undoubtedly looking particularly unique coloured but exceptional golden ring, incredible blue eye pupils and a tongue sticking out underneath **the baby bear**."

"**A vintage car** with six clock stickers and **puppy dog** eyes."

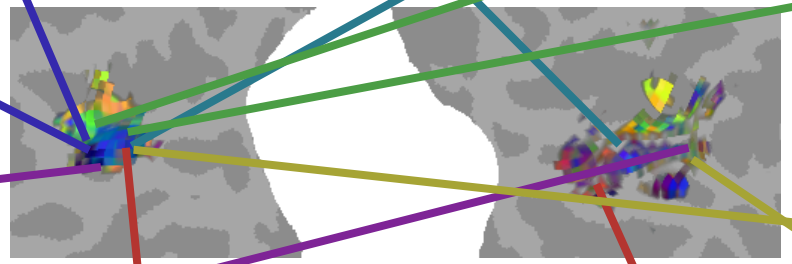
"**A food** packaging features **a smiling person** and a cartoon character**.**"

"These decorated every **food item** or product registration display **the brand name.**"

"Woodnut coloured blue eyes and a funny face painted by one person showing **a pig** sticking out his tongue."

"How humorous detail after detail of **brand name candy** and food."

"The animal photos show how cute it is to comfort **a red cardinal** by touching zoo noses with bread."



"A themed dessert tray decorated like **pound cake**, decorated with cupcake stamps."

"**The person** holding the object poses while breathing while concentrating while playing or a person wearing athletic clothing during an event connected to **an outdoor sporting activity**."

"**People** walking through the snowy landscape, people standing on a body of water."

"**A vegetable** plant covered with colorful gold leaf symbolically faces display such images such as **a group of people** dressed in camouflage representing the market setting."

"Pink frosting resembles a dessert tray and **cupcake**."

Conclusions & Discussion

Conclusions

- LaVCa was able to **describe the selectivity of visual cortex voxels with higher accuracy than existing methods.**
- Furthermore, we revealed that **diverse selectivity exists beyond the known selectivity of ROIs.**

Discussion

- These results suggest **the potential of using LLMs to better understand human brain representations**
- This approach may be **applicable to multimodal** settings, potentially enabling integrated interpretation of brain representations across vision, audition, language, and higher-level cognition.