Supplemental Material for Multimodal Neurons in Pretrained Text-Only Transformers

S.1. Implementation details

We follow the LiMBeR process for augmenting pretrained GPT-J with vision as described in Merullo *et al.* (2022). Each image is resized to $(224, 224)$ and encoded into a sequence $[i_1, ..., i_k]$ by the image encoder E, where $k = 196$ and each i corresponds to an image patch of size $(16, 16)$. We use self-supervised BEIT as E, trained with no linguistic supervision, which produces $[i_1, ..., i_k]$ of dimensionality 1024. To project image representations i into the transformer-defined embedding space of GPT-J, we use linear layer P from Merullo *et al.* (2022), trained on an image-to-text task (CC3M image captioning). P transforms $[i_1, ..., i_k]$ into soft prompts $[x_1, ..., x_k]$ of dimensionality 4096, which we refer to as the image prompt. Following convention from SimVLM, MAGMA and LiMBeR, we append the text prefix "A picture of" after every every image prompt. Thus for each image, GPT-J receives as input a $(199, 4096)$ prompt and outputs a probability distribution y over next-token continuations of that prompt.

To calculate neuron attribution scores, we generate a caption for each image by sampling from y using temperature $T = 0$, which selects the token with the highest probability at each step. The attribution score $g_{k,c}$ of neuron k is then calculated with respect to token c, where c is the first noun in the generated caption (which directly follows the image prompt and is less influenced by earlier token predictions). In the rare case where this noun is comprised of multiple tokens, we let c be the first of these tokens. This attribution score lets us rank multimodal neurons by how much they contribute to the crossmodal image captioning task.

S.2. Example multimodal neurons

Table S.1 shows additional examples of multimodal neurons detected and decoded for randomly sampled images from the COCO 2017 validation set. The table shows the top 20 neurons across all MLP layers for each image. In analyses where we filter for interpretable neurons that correspond to objects or object features in images, we remove neurons that decode primarily to word fragments or punctuation. Interpretable units (units where at least 7 of the top 10 tokens are words in the SCOWL English dictionary, for en-US or en-GB, with ≥ 3 letters) are highlighted in bold.

S.3. Evaluating agreement with image captions

We use BERTScore (F1) as a metric for evaluating how well a list of tokens corresponds to the semantic content of an image caption. Section 2.2 uses this metric to evaluate multimodal neurons relative to ground-truth human annotations from COCO, and Section 3.1 uses the metric to determine whether projection layer P translates $[i_1, ..., i_k]$ into $[x_1, ..., x_k]$ that already map visual features onto related language before reaching transformer MLPs. Given that $[x_1, ..., x_k]$ do not correspond to discrete tokens, we map each x onto the 5 token vectors with highest cosine similarity in the transformer embedding space for analysis.

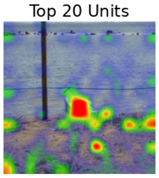
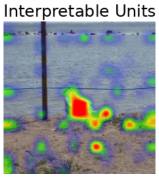
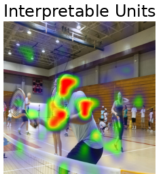
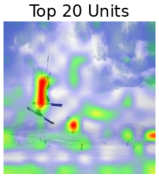
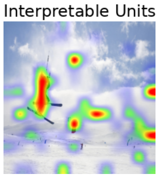
Table S.2 shows example decoded soft prompts for a randomly sampled COCO image. For comparison, we sample random vectors of size 4096 and use the same procedure to map them onto their nearest neighbors in the GPT-J embedding space. BERTScores for the random soft prompts are shown alongside scores for the image soft prompts. The means of these BERTScores, as well as the maximum values, are indistinguishable for real and random soft prompts (see Table S.2 for a single image and Figure 3 in the main paper for the distribution across COCO images). Thus we conclude that P produces image prompts that fit within the GPT-J embedding space, but do not already map image features onto related language: this occurs deeper inside the transformer.

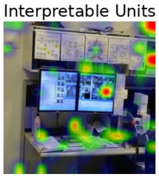
S.4. Selectivity of multimodal neurons

Figure S.1 shows additional examples of activation masks of individual multimodal neurons over COCO validation images, and IoU scores comparing each activation mask with COCO object annotations.

We conduct an additional experiment to test whether multimodal neurons are selectively active for images containing particular concepts. If unit k is selective for the images it describes (and not, for instance, for many images), then we expect greater $A^k_{x_i}$ on images where it relevant to the caption than on images where it is irrelevant. It is conceivable that our method merely extracts a set of high-activating neurons, not a set of neurons that are selectively active on the inputs we claim they are relevant to captioning.

We select 10 diverse ImageNet classes (see Figure S.2) and compute the top 100 scoring units per image on each of 200 randomly sampled images per class in the ImageNet training set, filtered for interpretable units. Then for each class, we select the 20 units that appear in the most images for that class. We measure the mean activation of these units across all patches in the ImageNet validation images for each of the 10 classes. Figure S.2(a) shows the comparison of activations across each of the categories. We find that neurons activate more frequently on images in their own category than for others. This implies that our pipeline does not extract a set of general visually attentive units, but rather units that are specifically tied to image semantics.

Images	Layer.unit	Patch	Decoding (top 5 tokens)	Attr. score
Raw Image 	**L7.u15772**	**119**	**‘ animals’, ‘ embryos’, ‘ kittens’, ‘ mammals’, ‘ eggs’**	**0.0214**
	L5.u4923	**119**	**‘ birds’, ‘ cages’, ‘ species’, ‘ breeding’, ‘ insects’**	**0.0145**
	L7.u12134	**119**	**‘ aircraft’, ‘ flight’, ‘ airplanes’, ‘ Flight’, ‘ Aircraft’**	**0.0113**
	L5.u4888	**119**	**‘ Boat’, ‘ sails’, ‘voy’, ‘ boats’, ‘ ships’**	**0.0085**
	L7.u5875	**119**	**‘ larvae’, ‘ insects’, ‘ mosquitoes’, ‘ flies’, ‘ species’**	**0.0083**
	L8.u2012	**105**	**‘ whales’, ‘ turtles’, ‘ whale’, ‘ birds’, ‘ fishes’**	**0.0081**
	L7.u3030	**119**	**‘ Island’, ‘ island’, ‘ Islands’, ‘ islands’, ‘ shore’**	**0.0078**
Top 20 Units 	L7.u14308	119	‘uses’, ‘ dec’, ‘bill’, ‘oid’, ‘FS’	0.0078
	L9.u12771	**119**	**‘ satellites’, ‘ Flight’, ‘ orbiting’, ‘ spacecraft’, ‘ ship’**	**0.0075**
	L4.u12317	**119**	**‘ embryos’, ‘ chicken’, ‘ meat’, ‘ fruits’, ‘ cows’**	**0.0071**
	L8.u2012	**119**	**‘ whales’, ‘ turtles’, ‘ whale’, ‘ birds’, ‘ fishes’**	**0.0062**
	L5.u4530	**119**	**‘ herds’, ‘ livestock’, ‘ cattle’, ‘ herd’, ‘ manure’**	**0.0056**
	L5.u4923	**105**	**‘ birds’, ‘ cages’, ‘ species’, ‘ breeding’, ‘ insects’**	**0.0055**
	L6.u8956	**119**	**‘ virus’, ‘ strains’, ‘ infect’, ‘ viruses’, ‘ parasites’**	**0.0052**
Interpretable Units 	L7.u2159	105	‘ species’, ‘species’, ‘ bacteria’, ‘ genus’, ‘ Species’	0.0051
	L10.u4819	119	‘çĶ°’, ‘¬¼’, ‘”””’, ‘ Marketable’, ‘å§’	0.0051
	L5.u4923	**118**	**‘ birds’, ‘ cages’, ‘ species’, ‘ breeding’, ‘ insects’**	**0.0050**
	L10.u927	3	‘onds’, ‘rog’, ‘lys’, ‘arrow’, ‘ond’	0.0050
	L11.u7635	**119**	**‘ birds’, ‘birds’, ‘ butterflies’, ‘ kittens’, ‘ bird’**	**0.0049**
	L9.u15445	**119**	**‘ radar’, ‘ standby’, ‘ operational’, ‘ flight’, ‘ readiness’**	**0.0048**
Raw Image 	**L5.u15728**	**119**	**‘ playoff’, ‘ players’, ‘ teammate’, ‘ player’, ‘Players’**	**0.0039**
	L12.u11268	113	‘elson’, ‘ISA’, ‘Me’, ‘PRES’, ‘SO’	0.0039
	L5.u9667	**119**	**‘ workouts’, ‘ workout’, ‘ Training’, ‘ trainer’, ‘ exercises’**	**0.0034**
	L9.u15864	182	‘lihood’, ‘/**’, ‘Advertisements’, ‘.”.’, ‘”””’	0.0034
	L9.u9766	**119**	**‘ soccer’, ‘ football’, ‘ player’, ‘ baseball’, ‘player’**	**0.0033**
	L10.u4819	182	‘çĶ°’, ‘¬¼’, ‘”””’, ‘ Marketable’, ‘å§’	0.0033
	L18.u15557	150	‘imer’, ‘ohan’, ‘ellow’, ‘ims’, ‘gue’	0.0032
Top 20 Units 	L12.u6426	160	‘â¢’, ‘ Â®’, ‘ syndrome’, ‘ Productions’, ‘ Ltd’	0.0032
	L8.u15435	**119**	**‘ tennis’, ‘ tournaments’, ‘ tournament’, ‘ golf’, ‘ racing’**	**0.0032**
	L11.u4236	**75**	**‘ starring’, ‘ played’, ‘ playable’, ‘ Written’, ‘ its’**	**0.0031**
	L8.u6207	**119**	**‘ player’, ‘ players’, ‘ Player’, ‘Ä’, ‘ talent’**	**0.0031**
	L6.u5975	**119**	**‘ football’, ‘ soccer’, ‘ basketball’, ‘ Soccer’, ‘ Football’**	**0.0030**
	L2.u10316	75	‘ï’, ‘/**’, ‘Q’, ‘The’, ‘//’	0.0028
	L12.u8390	89	‘etheless’, ‘viously’, ‘theless’, ‘bsite’, ‘terday’	0.0028
Interpretable Units 	**L5.u7958**	**89**	**‘ rugby’, ‘ football’, ‘ player’, ‘ soccer’, ‘ footballer’**	**0.0028**
	L20.u9909	89	‘ Associates’, ‘ Alt’, ‘ para’, ‘ Lt’, ‘ similarly’	0.0026
	L5.u8219	**75**	**‘ portion’, ‘ regime’, ‘ sector’, ‘ situation’, ‘ component’**	**0.0026**
	L11.u7264	**75**	**‘ portion’, ‘ finale’, ‘ environment’, ‘iest’, ‘ mantle’**	**0.0026**
	L20.u452	103	‘ CLE’, ‘ plain’, ‘ clearly’, ‘ Nil’, ‘ Sullivan’	0.0026
	L7.u16050	89	‘pc’, ‘IER’, ‘ containing’, ‘ formatted’, ‘ supplemented’	0.0026
Raw Image 	L10.u927	73	‘onds’, ‘rog’, ‘lys’, ‘arrow’, ‘ond’	0.0087
	L5.u9667	**101**	**‘ workouts’, ‘ workout’, ‘ Training’, ‘ trainer’, ‘ exercises’**	**0.0081**
	L9.u3561	73	‘ mix’, ‘ CRC’, ‘ critically’, ‘ gulf’, ‘ mechanically’	0.0076
	L9.u5970	**73**	**‘ construct’, ‘ performance’, ‘ global’, ‘ competing’, ‘ transact’**	**0.0054**
	L10.u562	73	‘ prev’, ‘ struct’, ‘ stable’, ‘ marg’, ‘ imp’	0.0054
	L6.u14388	**87**	**‘ march’, ‘ treadmill’, ‘ Championships’, ‘ racing’, ‘ marathon’**	**0.0052**
	L14.u10320	**73**	**‘ print’, ‘ handle’, ‘ thing’, ‘catch’, ‘error’**	**0.0051**
Top 20 Units 	L9.u3053	73	‘essel’, ‘ked’, ‘ ELE’, ‘ument’, ‘ue’	0.0047
	L5.u4932	73	‘eman’, ‘rack’, ‘ago’, ‘anne’, ‘ison’	0.0046
	L9.u7777	101	‘dr’, ‘thur’, ‘tern’, ‘mas’, ‘mass’	0.0042
	L6.u16106	73	‘umble’, ‘archives’, ‘room’, ‘ decentral’, ‘Root’	0.0040
	L5.u14519	**73**	**‘ abstract’, ‘ global’, ‘map’, ‘exec’, ‘kernel’**	**0.0039**
	L11.u10405	73	‘amed’, ‘elect’, ‘1’, ‘vol’, ‘vis’	0.0038
	L9.u325	87	‘ training’, ‘ tournaments’, ‘ango’, ‘ ballet’, ‘ gymn’	0.0038
Interpretable Units 	**L6.u14388**	**101**	**‘ march’, ‘ treadmill’, ‘ Championships’, ‘ racing’, ‘ marathon’**	**0.0038**
	L7.u3844	101	‘DERR’, ‘Charges’, ‘wana’, ‘¬¼’, ‘verages’	0.0036
	L9.u15864	101	‘lihood’, ‘/**’, ‘Advertisements’, ‘.”.’, ‘”””’	0.0036
	L7.u3330	**101**	**‘ Officers’, ‘ officers’, ‘ patrolling’, ‘ patrols’, ‘ troops’**	**0.0036**
	L8.u8807	**73**	**‘ program’, ‘ updates’, ‘ programs’, ‘ document’, ‘ format’**	**0.0034**
	L6.u12536	**87**	**‘ ankles’, ‘ joints’, ‘ biome’, ‘ injuries’, ‘ injury’**	**0.0034**

Images	Layer.unit	Patch	Decoding (top 5 tokens)	Attr. score
Raw Image	**L8.u14504**	**13**	**‘ upstairs’, ‘ homeowners’, ‘ apartments’, ‘ houses’, ‘ apartment’**	**0.0071**
	**L13.u15107**	**93**	**‘ meals’, ‘ meal’, ‘ dinner’, ‘ dishes’, ‘ cuisine’**	**0.0068**
	L8.u14504	**93**	**‘ upstairs’, ‘ homeowners’, ‘ apartments’, ‘ houses’, ‘ apartment’**	**0.0052**
	L8.u14504	**150**	**‘ upstairs’, ‘ homeowners’, ‘ apartments’, ‘ houses’, ‘ apartment’**	**0.0048**
	L9.u4691	**13**	**‘ houses’, ‘ buildings’, ‘ dwellings’, ‘ apartments’, ‘ homes’**	**0.0043**
	L8.u13681	**93**	**‘ sandwiches’, ‘ foods’, ‘ salad’, ‘ sauce’, ‘ pizza’**	**0.0041**
	L12.u4638	**93**	**‘ wash’, ‘ Darkness’, ‘ Caps’, ‘ blush’, ‘ Highest’**	**0.0040**
Top 20 Units	L9.u3561	93	‘ mix’, ‘ CRC’, ‘ critically’, ‘ gulf’, ‘ mechanically’	0.0040
	**L7.u5533**	**93**	**‘bags’, ‘Items’, ‘ comprehens’, ‘ decor’, ‘bag’**	**0.0039**
	L9.u8687	**93**	**‘ eaten’, ‘ foods’, ‘ food’, ‘ diet’, ‘ eating’**	**0.0037**
	L12.u4109	**93**	**‘ Lakes’, ‘ Hof’, ‘ Kass’, ‘ Cotton’, ‘Council’**	**0.0036**
	L8.u943	**93**	**‘ Foods’, ‘Food’, ‘let’, ‘ lunch’, ‘commercial’**	**0.0036**
	L5.u16106	**93**	**‘ware’, ‘ halls’, ‘ salt’, ‘WARE’, ‘ mat’**	**0.0032**
	L8.u14504	**143**	**‘ upstairs’, ‘ homeowners’, ‘ apartments’, ‘ houses’, ‘ apartment’**	**0.0032**
Interpretable Units	**L9.u11735**	**93**	**‘ hysterical’, ‘ Gould’, ‘ Louie’, ‘ Gamble’, ‘ Brown’**	**0.0031**
	**L8.u14504**	**149**	**‘ upstairs’, ‘ homeowners’, ‘ apartments’, ‘ houses’, ‘ apartment’**	**0.0031**
	L5.u2771	**93**	**‘ occupations’, ‘ industries’, ‘ operations’, ‘ occupational’, ‘ agriculture’**	**0.0029**
	L9.u15864	55	‘lihood’, ‘/**’, ‘Advertisements’, ‘.”.’, ‘””””’	0.0028
	L9.u4691	**149**	**‘ houses’, ‘ buildings’, ‘ dwellings’, ‘ apartments’, ‘ homes’**	**0.0028**
	L7.u10853	**13**	**‘ boutique’, ‘ firm’, ‘ Associates’, ‘ restaurant’, ‘ Gifts’**	**0.0028**
Raw Image	**L8.u15435**	**160**	**‘ tennis’, ‘ tournaments’, ‘ tournament’, ‘ golf’, ‘ racing’**	**0.0038**
	L1.u15996	132	‘276’, ‘PS’, ‘ley’, ‘room’, ‘ Will’	0.0038
	L5.u6439	160	‘ ge’, ‘ fibers’, ‘ hair’, ‘ geometric’, ‘ ori’	0.0037
	L9.u15864	160	‘lihood’, ‘/**’, ‘Advertisements’, ‘.”.’, ‘””””’	0.0034
	L12.u2955	160	‘Untitled’, ‘Welcome’, ‘========’, ‘Newsletter’, ‘====’	0.0033
	L12.u2955	146	‘Untitled’, ‘Welcome’, ‘========’, ‘Newsletter’, ‘====’	0.0032
	L7.u2688	160	‘rection’, ‘itud’, ‘ Ratio’, ‘lat’, ‘ ratio’	0.0031
Top 20 Units	**L8.u4372**	**160**	**‘ footage’, ‘ filmed’, ‘ filming’, ‘ videos’, ‘ clips’**	**0.0029**
	L10.u4819	146	‘çĶ°’, ‘¬¼’, ‘””””’, ‘ Marketable’, ‘å§’	0.0029
	L8.u15435	**93**	**‘ tennis’, ‘ tournaments’, ‘ tournament’, ‘ golf’, ‘ racing’**	**0.0029**
	L8.u15435	**146**	**‘ tennis’, ‘ tournaments’, ‘ tournament’, ‘ golf’, ‘ racing’**	**0.0029**
	L10.u927	132	‘onds’, ‘rog’, ‘lys’, ‘arrow’, ‘ond’	0.0027
	L9.u15864	146	‘lihood’, ‘/**’, ‘Advertisements’, ‘.”.’, ‘””””’	0.0026
	L1.u8731	132	‘ âĢ¦’, ‘ [âĢ¦]’, ‘âĢ¦’, ‘ ...’, ‘ Will’	0.0025
Interpretable Units	**L8.u16330**	**160**	**‘ bouncing’, ‘ hitting’, ‘ bounce’, ‘ moving’, ‘ bounced’**	**0.0025**
	L9.u1908	146	‘ members’, ‘ country’, ‘ VIII’, ‘ Spanish’, ‘ 330’	0.0024
	L10.u4819	160	‘çĶ°’, ‘¬¼’, ‘””””’, ‘ Marketable’, ‘å§’	0.0024
	L11.u14710	**160**	**‘Search’, ‘Follow’, ‘Early’, ‘Compar’, ‘Category’**	**0.0024**
	L6.u132	160	‘ manually’, ‘ replace’, ‘ concurrently’, ‘otropic’, ‘ foregoing’	0.0024
	L7.u5002	**160**	**‘ painting’, ‘ paintings’, ‘ sculpture’, ‘ sculptures’, ‘ painted’**	**0.0024**

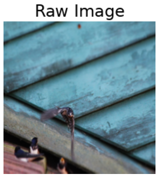
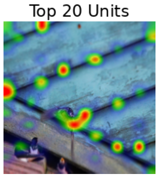
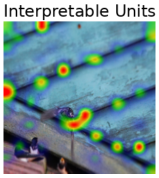
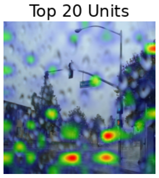
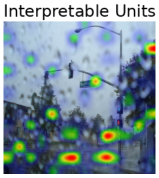
Images	Layer.unit	Patch	Decoding (top 5 tokens)	Attr. score
Raw Image 	**L5.u13680**	**132**	**‘ driver’, ‘ drivers’, ‘ cars’, ‘heading’, ‘cars’**	**0.0091**
	L11.u9566	**132**	**‘ traffic’, ‘ network’, ‘ networks’, ‘ Traffic’, ‘network’**	**0.0090**
	L12.u11606	**132**	**‘ chassis’, ‘ automotive’, ‘ design’, ‘ electronics’, ‘ specs’**	**0.0078**
	L7.u6109	**132**	**‘ automobile’, ‘ automobiles’, ‘ engine’, ‘ Engine’, ‘ cars’**	**0.0078**
	L6.u11916	**132**	**‘ herd’, ‘loads’, ‘ racing’, ‘ herds’, ‘ horses’**	**0.0071**
	L8.u562	**132**	**‘ vehicles’, ‘ vehicle’, ‘ cars’, ‘veh’, ‘ Vehicles’**	**0.0063**
	L7.u3273	**132**	**‘ride’, ‘ riders’, ‘ rides’, ‘ ridden’, ‘ rider’**	**0.0062**
Top 20 Units 	**L13.u5734**	**132**	**‘ Chevrolet’, ‘ Motorsport’, ‘ cars’, ‘ automotive’, ‘ vehicle’**	**0.0062**
	L8.u2952	**132**	**‘ rigging’, ‘ valves’, ‘ nozzle’, ‘ pipes’, ‘ tubing’**	**0.0059**
	L13.u8962	**132**	**‘ cruising’, ‘ flying’, ‘ flight’, ‘ refuel’, ‘ Flying’**	**0.0052**
	L9.u3561	116	‘ mix’, ‘ CRC’, ‘ critically’, ‘ gulf’, ‘ mechanically’	0.0051
	L13.u107	**132**	**‘ trucks’, ‘ truck’, ‘ trailer’, ‘ parked’, ‘ driver’**	**0.0050**
	L14.u10852	**132**	**‘Veh’, ‘ driver’, ‘ automotive’, ‘ automakers’, ‘Driver’**	**0.0049**
	L6.u1989	132	‘text’, ‘light’, ‘TL’, ‘X’, ‘background’	0.0049
Interpretable Units 	L2.u14243	132	‘ousel’, ‘ Warriors’, ‘riages’, ‘illion’, ‘Ord’	0.0048
	L5.u6589	**132**	**‘ vehicles’, ‘ motorcycles’, ‘ aircraft’, ‘ tyres’, ‘ cars’**	**0.0046**
	L7.u4574	**132**	**‘ plants’, ‘ plant’, ‘ roof’, ‘ compost’, ‘ wastewater’**	**0.0045**
	L7.u6543	**132**	**‘ distance’, ‘ downhill’, ‘ biking’, ‘ riders’, ‘ journeys’**	**0.0045**
	L16.u9154	**132**	**‘ driver’, ‘ drivers’, ‘ vehicle’, ‘ vehicles’, ‘driver’**	**0.0045**
	L12.u7344	**132**	**‘ commemor’, ‘ streets’, ‘ celebrations’, ‘ Streets’, ‘ highways’**	**0.0044**
Raw Image 	**L12.u9058**	**174**	**‘ swimming’, ‘ Swim’, ‘ swim’, ‘ fishes’, ‘ water’**	**0.0062**
	L17.u10507	**174**	**‘ rivers’, ‘ river’, ‘ lake’, ‘ lakes’, ‘ River’**	**0.0049**
	L7.u3138	**174**	**‘ basin’, ‘ ocean’, ‘ islands’, ‘ valleys’, ‘ mountains’**	**0.0046**
	L5.u6930	**149**	**‘ rivers’, ‘ river’, ‘ River’, ‘ waters’, ‘ waterways’**	**0.0042**
	L7.u14218	**174**	**‘ docks’, ‘ Coast’, ‘ swimming’, ‘ swim’, ‘melon’**	**0.0040**
	L9.u4379	**149**	**‘ river’, ‘ stream’, ‘ River’, ‘ Valley’, ‘ flow’**	**0.0038**
	L6.u5868	**149**	**‘water’, ‘ water’, ‘ waters’, ‘ river’, ‘ River’**	**0.0036**
Top 20 Units 	**L9.u4379**	**174**	**‘ river’, ‘ stream’, ‘ River’, ‘ Valley’, ‘ flow’**	**0.0036**
	L5.u6930	**174**	**‘ rivers’, ‘ river’, ‘ River’, ‘ waters’, ‘ waterways’**	**0.0032**
	L7.u3138	**149**	**‘ basin’, ‘ ocean’, ‘ islands’, ‘ valleys’, ‘ mountains’**	**0.0029**
	L6.u5868	**174**	**‘water’, ‘ water’, ‘ waters’, ‘ river’, ‘ River’**	**0.0028**
	L7.u416	136	‘ praise’, ‘ glimpse’, ‘ glimps’, ‘ palate’, ‘ flavours’	0.0027
	L10.u15235	**149**	**‘ water’, ‘ waters’, ‘water’, ‘ lake’, ‘ lakes’**	**0.0026**
	L4.u2665	**136**	**‘ levels’, ‘ absorbed’, ‘ density’, ‘ absorption’, ‘ equilibrium’**	**0.0026**
Interpretable Units 	**L10.u14355**	**149**	**‘ roads’, ‘ paths’, ‘ flows’, ‘ routes’, ‘ streams’**	**0.0026**
	L17.u10507	**149**	**‘ rivers’, ‘ river’, ‘ lake’, ‘ lakes’, ‘ River’**	**0.0024**
	L7.u7669	**174**	**‘ weather’, ‘ season’, ‘ forecast’, ‘ rains’, ‘ winters’**	**0.0024**
	L8.u9322	**136**	**‘ combustion’, ‘ turbulence’, ‘ recoil’, ‘ vibration’, ‘ hydrogen’**	**0.0024**
	L9.u15864	182	‘lihood’, ‘/**’, ‘Advertisements’, ‘.”.’, ‘”””’	0.0022
	L7.u3138	**78**	**‘ basin’, ‘ ocean’, ‘ islands’, ‘ valleys’, ‘ mountains’**	**0.0021**

Table S.1. Results of attribution analysis for randomly sampled images from the COCO validation set. Includes decoded tokens for the top 20 units by attribution score. The first column shows the COCO image and superimposed heatmaps of the mean activations from the top 20 units and the top interpretable units (shown in **bold**). Units can repeat if they attain a high attribution score on multiple image patches.

Image	COCO Human Captions	GPT Caption
	A man riding a snowboard down the side of a snow covered slope. A man snowboarding down the side of a snowy mountain. Person snowboarding down a steep snow covered slope. A person snowboards on top of a snowy path. The person holds both hands in the air while snowboarding.	A person jumping on the ice.

Patch	Image soft prompt (nearest neighbor tokens)	BSc.	Random soft prompt (nearest neighbor tokens)	BSc.
144	['nav', 'GY', '+++', 'done', 'Sets']	.29	['Movement', 'Ord', 'CLUD', 'levy', 'LI']	.31
80	['heels', 'merits', 'flames', 'platform', 'fledged']	.36	['adic', 'Stub', 'imb', 'VER', 'stroke']	.34
169	['ear', 'Nelson', 'Garden', 'Phill', 'Gun']	.32	['Thank', 'zilla', 'Develop', 'Invest', 'Fair']	.31
81	['vanilla', 'Poc', 'Heritage', 'Tarant', 'bridge']	.33	['Greek', 'eph', 'jobs', 'phylogen', 'TM']	.30
89	['oily', 'stant', 'cement', 'Caribbean', 'Nad']	.37	['Forestry', 'Mage', 'Hatch', 'Buddh', 'Beaut']	.34
124	['ension', 'ideas', 'GY', 'uler', 'Nelson']	.32	['itone', 'gest', 'Af', 'iple', 'Dial']	.30
5	['proves', 'Feed', 'meaning', 'zzle', 'stripe']	.31	['multitude', 'psychologically', 'Taliban', 'Elf', 'Pakistan']	.36
175	['util', 'elson', 'asser', 'seek', '////////////////////']	.26	['ags', 'Git', 'mm', 'Morning', 'Cit']	.33
55	['Judicial', 'wasting', 'oen', 'oplan', 'trade']	.34	['odd', 'alo', 'rophic', 'perv', 'pei']	.34
61	['+++', 'DEP', 'enum', 'vernight', 'posted']	.33	['Newspaper', 'iii', 'INK', 'Graph', 'UT']	.35
103	['Doc', 'Barth', 'details', 'DEF', 'buckets']	.34	['pleas', 'Eclipse', 'plots', 'cb', 'Menu']	.36
99	['+++', 'Condition', 'Daytona', 'oir', 'research']	.35	['Salary', 'card', 'mobile', 'Cour', 'Hawth']	.35
155	['Named', '910', 'collar', 'Lars', 'Cats']	.33	['Champ', 'falsely', 'atism', 'styles', 'Champ']	.30
145	['cer', 'args', 'olis', 'te', 'atin']	.30	['Chuck', 'goose', 'anthem', 'wise', 'fare']	.33
189	['MOD', 'Pres', 'News', 'Early', 'Herz']	.33	['Organ', 'CES', 'POL', '201', 'Stan']	.31
49	['Pir', 'Pir', 'uum', 'akable', 'Prairie']	.30	['flame', 'roc', 'module', 'swaps', 'Faction']	.33
20	['ear', 'feed', 'attire', 'demise', 'peg']	.33	['Chart', 'iw', 'Kirst', 'PATH', 'rhy']	.36
110	['+++', 'Bee', 'limits', 'Fore', 'seeking']	.31	['imped', 'iola', 'Prince', 'inel', 'law']	.33
6	['SIGN', 'Kob', 'Ship', 'Near', 'buzz']	.36	['Tower', '767', 'Kok', 'Tele', 'Arbit']	.33
46	['childhood', 'death', 'ma', 'vision', 'Dire']	.36	['Fram', 'exper', 'Pain', 'ader', 'unprotected']	.33
113	['Decl', 'Hide', 'Global', 'orig', 'meas']	.32	['usercontent', 'OTUS', 'Georgia', 'ech', 'GRE']	.32
32	['ideas', 'GY', '+++', 'Bake', 'Seed']	.32	['GGGGGGGG', 'dictators', 'david', 'ugh', 'BY']	.31
98	['Near', 'Near', 'LIN', 'Bee', 'threat']	.30	['Lavrov', 'Debor', 'Hegel', 'Advertisement', 'iak']	.34
185	['ceans', 'Stage', 'Dot', 'Price', 'Grid']	.33	['wholesale', 'Cellular', 'Magn', 'Ingredients', 'Magn']	.32
166	['bys', '767', '+++', 'bottles', 'gif']	.32	['Bras', 'discipl', 'gp', 'AR', 'Toys']	.33
52	['Kob', 'Site', 'reed', 'Wiley', 'âĻ']	.29	['THER', 'FAQ', 'ibility', 'ilities', 'twitter']	.34
90	['cytok', 'attack', 'Plug', 'strategies', 'uddle']	.32	['Boots', 'Truman', 'CFR', 'ãĤ£', 'Shin']	.33
13	['nard', 'Planetary', 'lawful', 'Court', 'eman']	.33	['Nebraska', 'tails', 'ÅŁ', 'DEC', 'Despair']	.33
47	['pport', 'overnight', 'Doc', 'ierra', 'Unknown']	.34	['boiling', 'A', 'Ada', 'itude', 'flawed']	.31
19	['mocking', 'chicks', 'GY', 'ear', 'done']	.35	['illet', 'severely', 'nton', 'arrest', 'Volunteers']	.33
112	['avenue', 'gio', 'Parking', 'riages', 'Herald']	.35	['griev', 'Swanson', 'Guilty', 'Sent', 'Pac']	.32
133	['ãĤĬ', 'itto', 'iation', 'asley', 'Included']	.32	['Purs', 'reproductive', 'sniper', 'instruct', 'Population']	.33
102	['drawn', 'Super', 'gency', 'Type', 'blames']	.33	['metric', 'Young', 'princip', 'scal', 'Young']	.31
79	['Vand', 'inement', 'straw', 'ridiculous', 'Chick']	.34	['Rez', 'song', 'LEGO', 'Login', 'pot']	.37
105	['link', 'ede', 'Dunk', 'Pegasus', 'Mao']	.32	['visas', 'Mental', 'verbal', 'WOM', 'nda']	.30
	Average	.33		.33

Table S.2. Image soft prompts are indistinguishable from random soft prompts via BERTScore. Each image is encoded as a sequence of 196 soft prompts, corresponding to image patches, that serve as input to GPT-J. Here we randomly sample 35 patches for a single COCO image and map them onto nearest-neighbor tokens in transformer embedding space. BERTScore is measured relative to COCO human annotations of the same image (we report the mean score over the 5 human captions). For comparison we sample random vectors in the transformer embedding space and compute BERTScores using the same procedure.

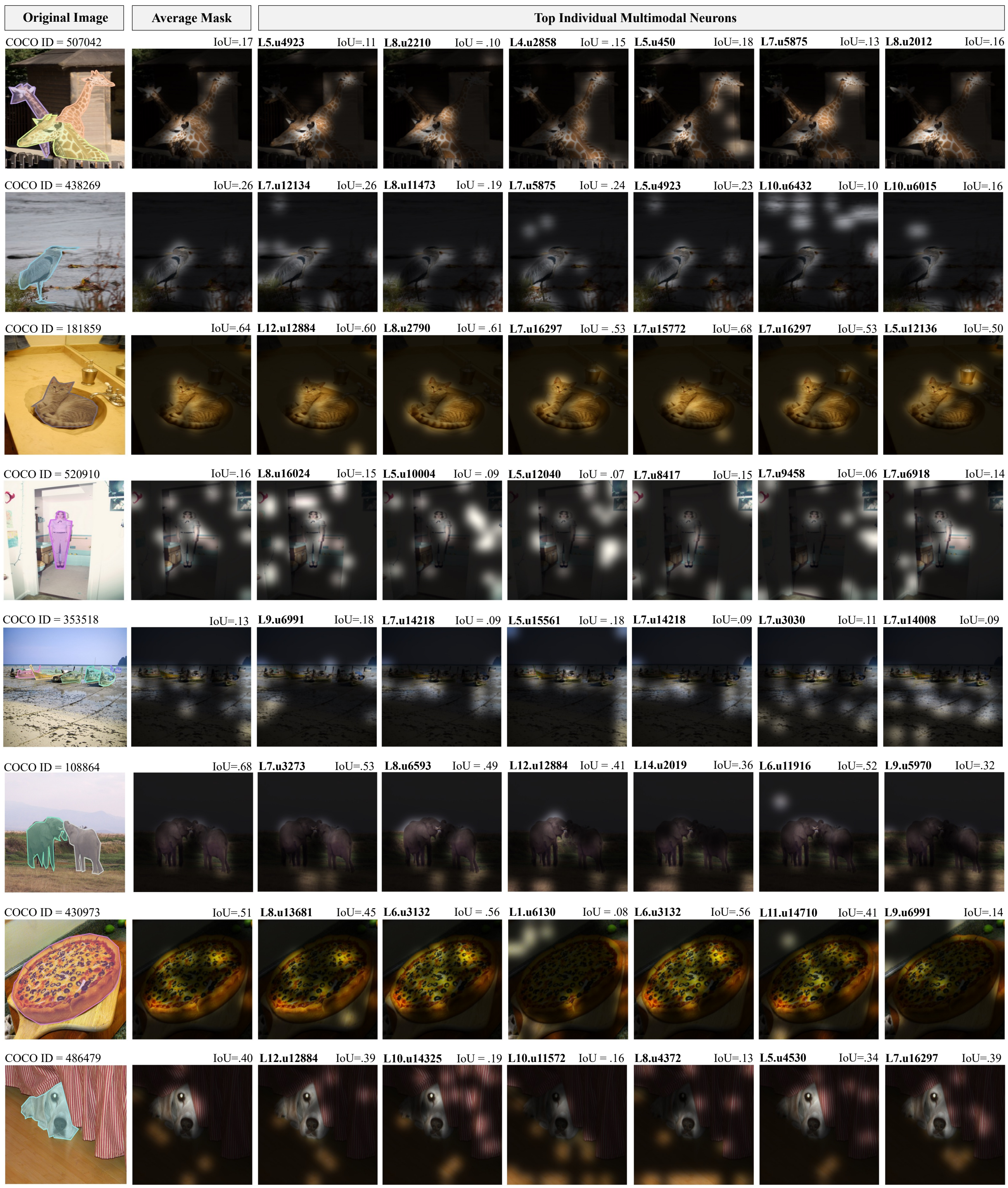


Figure S.1. Multimodal neurons are selective for objects in images. For 8 example images sampled from the COCO categories described in Section 3.2 of the main paper, we show activation masks of individual multimodal neurons over the image, as well as mean activation masks over all top multimodal neurons. We use IoU to compare these activation masks to COCO object annotations. IoU is calculated by upsampling each activation mask to the size of the original image (224) using bilinear interpolation, and thresholding activations in the 0.95 percentile to produce a binary segmentation mask.

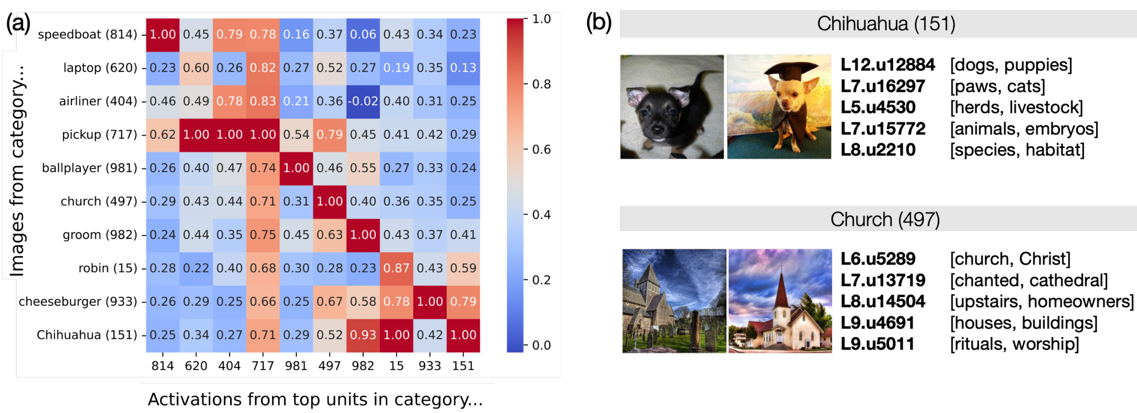


Figure S.2. Multimodal neurons are selective for image categories. (a) For 10 ImageNet classes we construct the set of interpretable multimodal neurons with the highest attribution scores on training images in that class, and calculate their activations on validation images. For each class, we report the average activation value of top-scoring multimodal units relative to the maximum value of their average activations on any class. Multimodal neurons are maximally active on classes where their attribution scores are highest. (b) Sample images and top-scoring units from two classes.

S.5. Ablating Multimodal Neurons

In Section 3.3 of the main paper, we show that ablating multimodal neurons causally effects the probability of outputting the original token. To investigate the effect of removing multimodal neurons on model output, we ablate the top k units by attribution score for an image, where $k \in \{0, 50, 100, 200, 400, 800, 1600, 3200, 6400\}$, and compute the BERTScore between the model's original caption and the newly-generated zero-temperature caption. Whether we remove the top k units by attribution score, or only those that are interpretable, we observe a strong decrease in caption similarity. Table S.3 shows examples of the effect of ablating top neurons on randomly sampled COCO validation images, compared to the effect of ablating random neurons. Figure S.3 shows the average BERTScore after ablating k units across all COCO validation images.

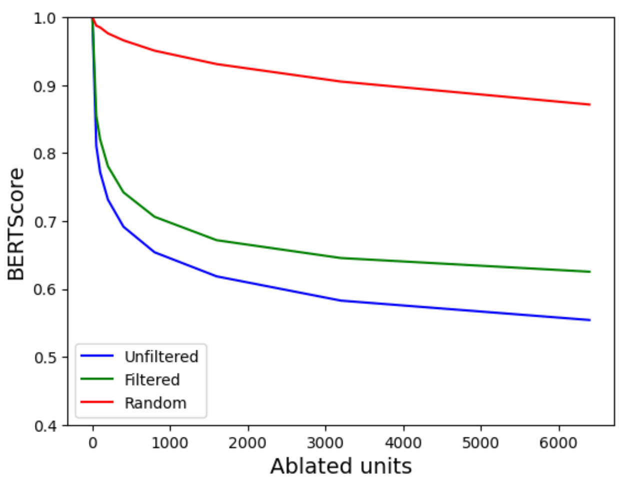


Figure S.3. BERTScores of generated captions decrease when multimodal neurons are ablated, compared to the ablation of random neurons from the same layers.

S.6. Distribution of Multimodal Neurons

We perform a simple analysis of the distribution of multimodal neurons by layer. Specifically, we extract the top 100 scoring neurons for all COCO validation images. Most of these neurons are found between layers 5 and 10 of GPT-J ($L = 28$), suggesting translation of semantic content between modalities occurs in earlier transformer layers.

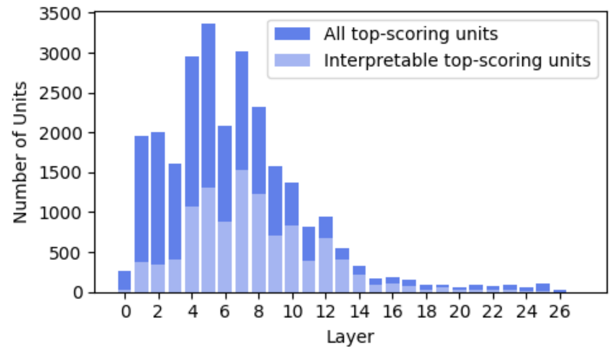


Figure S.4. Unique multimodal neurons per layer chosen using the top 100 attribution scores for each COCO validation image. Interpretable units are those for which at least 7 of the top 10 logits are words in the English dictionary containing ≥ 3 letters.

(a) 219578

(b) 131431

(c) 180878

(d) 128675

(e) 289960

(f) 281179

(g) 559842

(h) 47819

		Captions after ablation					
Img. ID	**# Abl.**	**All multimodal**	BSc.	**Interpretable multimodal**	BSc.	**Random neurons**	BSc.
219578	0	a dog with a cat	1.0	a dog with a cat	1.0	a dog with a cat	1.0
	50	a dog and a cat	.83	a dog and a cat	.83	a dog with a cat	1.0
	100	a lion and a zebra	.71	a dog and cat	.80	a dog with a cat	1.0
	200	a dog and a cat	.83	a dog and a cat	.83	a dog with a cat	1.0
	400	a lion and a lioness	.64	a dog and a cat	.83	a dog with a cat	1.0
	800	a tiger and a tiger	.63	a lion and a zebra	.71	a dog with a cat	1.0
	1600	a tiger and a tiger	.63	a lion and a zebra	.71	a dog with a cat	1.0
	3200	a tiger	.67	a tiger and a tiger	.63	a dog with a cat	1.0
	6400	a tiger	.67	a tiger in the jungle	.60	a dog with a cat	1.0
131431	0	the facade of the cathedral	1.0	the facade of the cathedral	1.0	the facade of the cathedral	1.0
	50	the facade of the church	.93	the facade of the cathedral	1.0	the facade of the cathedral	1.0
	100	the facade of the church	.93	the facade of the cathedral	1.0	the facade of the cathedral	1.0
	200	the facade	.75	the facade	.75	the facade of the cathedral	1.0
	400	the exterior of the church	.80	the facade	.75	the facade of the cathedral	1.0
	800	the exterior of the church	.80	the dome	.65	the facade of the cathedral	1.0
	1600	the dome	.65	the dome	.65	the facade of the cathedral	1.0
	3200	the dome	.65	the dome	.65	the facade of the cathedral	1.0
	6400	the exterior	.61	the dome	.65	the facade	.75
180878	0	a cake with a message written on it.	1.0	a cake with a message written on it.	1.0	a cake with a message written on it.	1.0
	50	a cake with a message written on it.	1.0	a cake with a message written on it.	1.0	a cake with a message written on it.	1.0
	100	a cake with a message written on it.	1.0	a cake for a friend's birthday.	.59	a cake with a message written on it.	1.0
	200	a cake with a message written on it.	1.0	a cake for a friend's birthday.	.59	a cake with a message written on it.	1.0
	400	a cake with a message written on it.	1.0	a cake for a friend's birthday.	.59	a cake with a message written on it.	1.0
	800	a cake	.59	a cake for a birthday party	.56	a cake with a message written on it.	1.0
	1600	a cake	.59	a poster for the film.	.49	a cake with a message written on it.	1.0
	3200	a man who is a fan of football	.42	a typewriter	.44	a cake with a message written on it.	1.0
	6400	the day	.34	a typewriter	.44	a cake with a message written on it.	1.0
128675	0	a man surfing on a wave	1.0	a man surfing on a wave	1.0	a man surfing on a wave	1.0
	50	a man in a kayak on a lake	.74	a man surfing on a wave	1.0	a man surfing on a wave	1.0
	100	a man in a kayak on a lake	.74	a man surfing on a wave	1.0	a man surfing on a wave	1.0
	200	a man in a kayak on a lake	.74	a man surfing a wave	.94	a man surfing on a wave	1.0
	400	a man in a kayak on a lake	.74	a man surfing a wave	.94	a man surfing on a wave	1.0
	800	a man in a kayak	.64	a surfer riding a wave	.84	a man surfing on a wave	1.0
	1600	a girl in a red dress walking on the beach	.66	a surfer riding a wave	.84	a man surfing on a wave	1.0
	3200	a girl in a red dress	.53	a girl in a red dress	.53	a man surfing on a wave	1.0
	6400	a girl in the water	.62	a girl in a dress	.59	a man surfing on a wave	1.0

Img. ID	# Abl.	All multimodal	BSc.	Interpretable multimodal	BSc.	Random neurons	BSc.
289960	0	a man standing on a rock in the sea	1.0	a man standing on a rock in the sea	1.0	a man standing on a rock in the sea	1.0
	50	a man standing on a rock in the sea	1.0	a man standing on a rock in the sea	1.0	a man standing on a rock in the sea	1.0
	100	a man standing on a rock in the sea	1.0	a man standing on a rock in the sea.	.94	a man standing on a rock in the sea	1.0
	200	a kite soaring above the waves	.62	a man standing on a rock in the sea	1.0	a man standing on a rock in the sea	1.0
	400	a kite soaring above the waves	.62	a kite surfer on the beach.	.62	a man standing on a rock in the sea	1.0
	800	a kite soaring above the waves	.62	a bird on a wire	.63	a man standing on a rock in the sea	1.0
	1600	a kite soaring above the clouds	.65	a kite surfer on the beach	.63	a man standing on a rock in the sea	1.0
	3200	a kite soaring above the sea	.69	a bird on a wire	.63	a man standing on a rock in the sea	1.0
	6400	a helicopter flying over the sea	.69	a bird on a wire	.63	a man standing on a rock in the sea	1.0
131431	0	the bridge at night	1.0	the bridge at night	1.0	the bridge at night	1.0
	50	the bridge	.70	the street at night	.82	the bridge at night	1.0
	100	the bridge	.70	the street at night	.82	the bridge at night	1.0
	200	the bridge	.70	the street at night	.82	the bridge at night	1.0
	400	the bridge	.70	the street	.55	the bridge at night	1.0
	800	the bridge	.70	the street	.55	the bridge at night	1.0
	1600	the bridge	.70	the street	.55	the bridge at night	1.0
	3200	the night	.61	the street	.55	the bridge at night	1.0
	6400	the night	.61	the street	.55	the bridge at night	1.0
559842	0	the team during the match.	1.0	the team during the match.	1.0	the team during the match.	1.0
	50	the team.	.70	the team.	.70	the team during the match.	1.0
	100	the team.	.70	the team.	.70	the team during the match.	1.0
	200	the team.	.70	the team.	.70	the team during the match.	1.0
	400	the group of people	.52	the team.	.70	the team during the match.	1.0
	800	the group	.54	the team.	.70	the team during the match.	1.0
	1600	the group	.54	the team.	.70	the team during the match.	1.0
	3200	the group	.54	the team.	.70	the team during the match	1.0
	6400	the kids	.46	the team.	.70	the team during the match.	1.0
47819	0	a man and his horse.	1.0	a man and his horse.	1.0	a man and his horse.	1.0
	50	a man and his horse.	1.0	a man and his horse.	1.0	a man and his horse.	1.0
	100	the soldiers on the road	.47	a man and his horse.	1.0	a man and his horse.	1.0
	200	the soldiers on the road	.47	the soldiers on the road	.47	a man and his horse.	1.0
	400	the soldiers	.46	the soldiers	.46	a man and his horse.	1.0
	800	the soldiers	.46	the soldiers	.46	a man and his horse.	1.0
	1600	the soldiers	.46	the soldiers	.46	a man and his horse.	1.0
	3200	the soldiers	.46	the soldiers	.46	a man and his horse.	1.0
	6400	the soldiers	.46	the soldiers	.46	a man and his horse.	1.0

Table S.3. Captions and BERTScores (relative to original GPT caption) after incremental ablation of multimodal MLP neurons. All multimodal neurons are detected, decoded, and filtered to produce a list of "interpretable" multimodal neurons using the procedure described in Section 2 of the main paper. Random neurons are sampled from the same layers as multimodal neurons for ablation. Images are randomly sampled from the COCO validation set. Captions are generated with temperature $= 0$.