Supplementary for "Generative learning enables brain signal recovery from multimodal prompts"

Ya-Li Li, Xin Liu, Jichuan Zhang, Shengjin Wang*

Department of Electronic Engineering, Tsinghua University

Beijing National Research Center for Information Science and Technology

*Correspondence: wgsgj@tsinghua.edu.cn

Abstract

In this supplementary material, we present more detailed results and visualization of predition differences. Especially, we provide more visualization for all the eight subjects in NSD dataset, for individual diversity investigation and analysis. Besides, we provide the additional investigative study from the deep learning perspective, *e.g.*, the importance of cross-modality alignment for multimodal generative learning.

1 Additional Results

1.1 More analysis on multimodal promptable generation

As depicted in the main paper, we designed the NeoDiffuser to adapt to multimodal prompts. These prompts are used as guidance for fMRI data generation. The common ones are visual prompts, correspond to the stimulus images revealed to participants (subjects). For layout prompts, the NeoDiffuser model produces fMRI signals from sparsely-distributed tagged objects. For (fMRI) signal prompts, the masked or polluted signals are used for data recovery with voxel relevance. In Fig.2 of main paper, we have presented the prediction accuracy of fMRI generation by multimodal prompts, *i.e.*, brain response simulation. The unnormalized prediction accuracy is measured by the unnormalized Pearson correleation coefficients (ρ_c) between the generated BOLD signals and real captured ones(ground truths). We further presented the detailed accuracy in the table below for reference. For the 8 subjects, the generation accuracy can be ranked as three groups with the layout and visual prompts. `S1`, `S2` and `S5` show superior prediction accuracy ($\rho_c \geq 45\%$), while `S3`, `S4`, `S6`, `S7` have moderate accuracy ($40\% > \rho_c \geq 35\%$). The unnormalized generation accuracy of $S8$ is the lowest, due to the signal capturing. With signal prompts, most of the subjects show comparable

Table 1: **Detailed prediction accuracy with different prompts, measured by mean Pearson coefficients (%)**

		S1	S2	S3	S4	S5	S6	S7	S8
layout	LH	26.69	24.55	21.00	19.88	35.41	18.78	20.88	16.96
	RH	27.44	28.83	23.96	22.24	35.35	18.87	24.61	18.87
visual	LH	46.36	48.16	37.54	35.29	50.27	34.34	36.06	27.46
	RH	46.12	47.68	37.42	39.66	49.80	37.12	35.38	27.85
signal	LH	64.36	61.39	66.17	67.39	73.90	68.53	66.78	67.80
	RH	66.97	59.36	68.24	69.48	72.67	69.71	65.91	67.88

performance, and the unnormalized generation accuracy of S5 is higher, indicating the signal quality and predictability.

We visualized the vertex-wise accuracy of layout promptable fMRI generation in Fig. I(a). We adopted the PyCortex to visualize the cortical flatmap of visual areas across 8 subjects (S1-S8). To be specific, we designed the layout prompts in accordance with functional localization. Only the object catogories and locations are used as generation guidance, while the fine visual details are removed. Despite the generation accuracy varies among multiple subjects, the tendancy is similar. It can be observed that the prediction accuracy is higher within the visual regions corresponding to high-level properties, but lower in primary visual cortex. To supplement the multimodal analysis and comparison, we also presented the vertex-wise accuracy of visual promptable generation in Fig. I(b). The prediction with visual prompts (images), by complementing the fine details such as colors, texture, contours, is relatively smooth across the whole visual area. The visualization of mutlimodal generation aligns with the hypothethized visual hierarchy and brain activation preferences.

As the complement measurement, we further visualize the difference bewteen visual or layout prompts. We calculate the vertex-wise R^2 values between the generated signals with visual or layout prompts. The visualization of R^2 values is presented in Fig. II, supports the visualization of different prompts in Fig. I of this supplementary material. The primary visual areas from V1-V3 are with lower R^2 values, indicating that the fMRI generation requires the fine-detailed guidance from visual prompts. On the contrary, the regions associated with high-level properties are robust to prompt changes. The multimodal analysis on fMRI generation provides the circumstantial evidence for supporting the hypothetical visual hierarchy.

The fMRI signals can be used to explore the stimuli with similar activation patterns. An investigative study on image retrieval is performed with real fMRI signals and generated ones. Fig.2e-f in the main paper provide the examples of image retrival with generated fMRI signals. We add the comparison with fMRI signals as "mediator" to retrieve stimulus images with similar brain responses, as in Fig. III of this supplementary material. The reference images are red-boxed, and the cosine similarity of fMRI data is used as measurement to select top-K related images. With

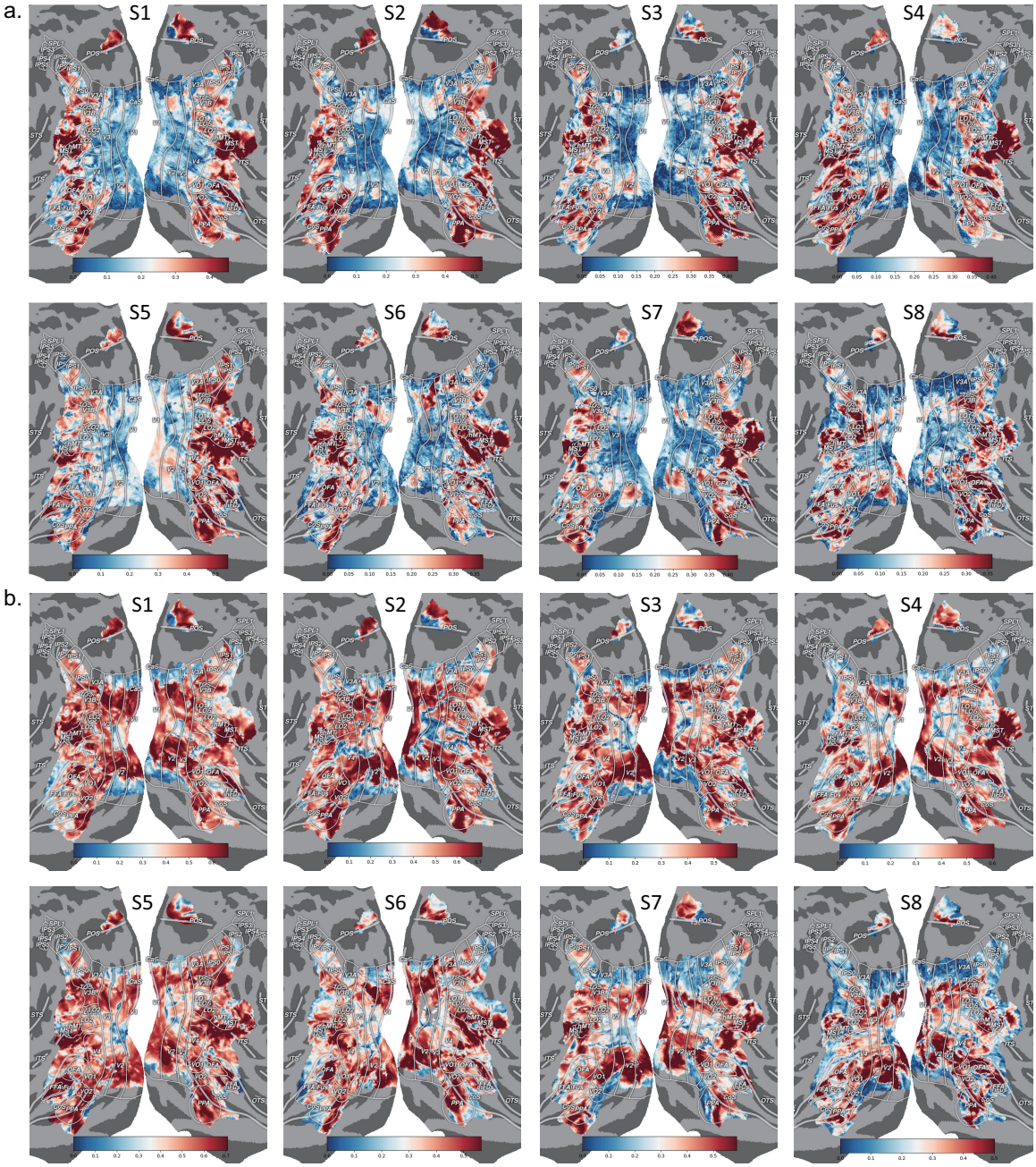

Fig. I: Visualization the accuracy of unified fMRI generation, measured by the Pearson correleation coefficients ρ_c (S1-S8). (a). With layout prompts, the fMRI generation is relatively better with higher-tier processing visual regions. **(b).**With visual prompts, the fMRI generation performance is relatively smooth with across visual processing areas.

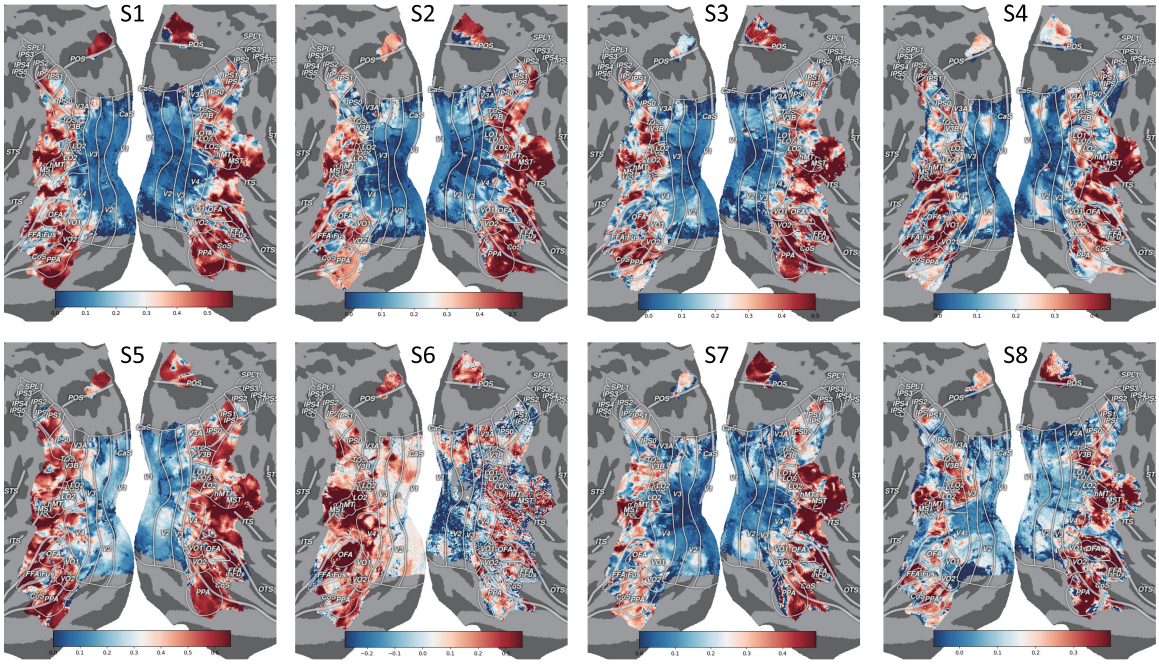

Fig. II: Comparison of prediction difference with visual or layout prompts, measured by R^2 values. Higher R^2 values correspond to red-colored regions, indicate that the BOLD signals of the fMRI voxels are robust to prompt changes.

real-captured fMRI signals ("GT" row), the retrieved images with real-captured fMRI signals mostly share the similar semantics, but also distracted by some other images. For example 2, the rank-3 retrieved image with "a boy with pizza" is "a dog and a cat". For example 4, the rank-1 and rank-2 images are "bell tower", sharing the similar shape with "beer bottle". It can be attributed to the complexity of brain responses and the noise capturing. By visual prompting, the generated brain responses attend to a broad range of imaging **appearance**, yielding similar but not identical retrieved results. For example 1, the generated brain responses of "skating man" image can be retrieved with the similar "sports" images under different scenarios (*e.g.*, streets, ski fields, grass). For example 3, the activation patterns of "sandwich" image can be associated with other "food" images (*e.g.*, pizza, cake). By layout prompting, the generated brain responses tend to attend the main **concept** objects in images, which might neglect other objects in complicated scenes. For example 2, the layout promptable generation identifies the "person", but ignores the "pizza" and "plate", yielding retrieved images are mostly "people"-related. For example 4, the generated signals focus on the "half-pizza" caused by trucation, therefore the retrieved images are mostly "food"-related. Only the rank-2 image is "a refrigerator packed with bottles" . As a potential approach to explore brain activation patterns, the NeoDiffuser can be used to associate **appearance**-related and **concept**-related stimulus images.

We investigated the generation differences of visual cortex regions. To remove

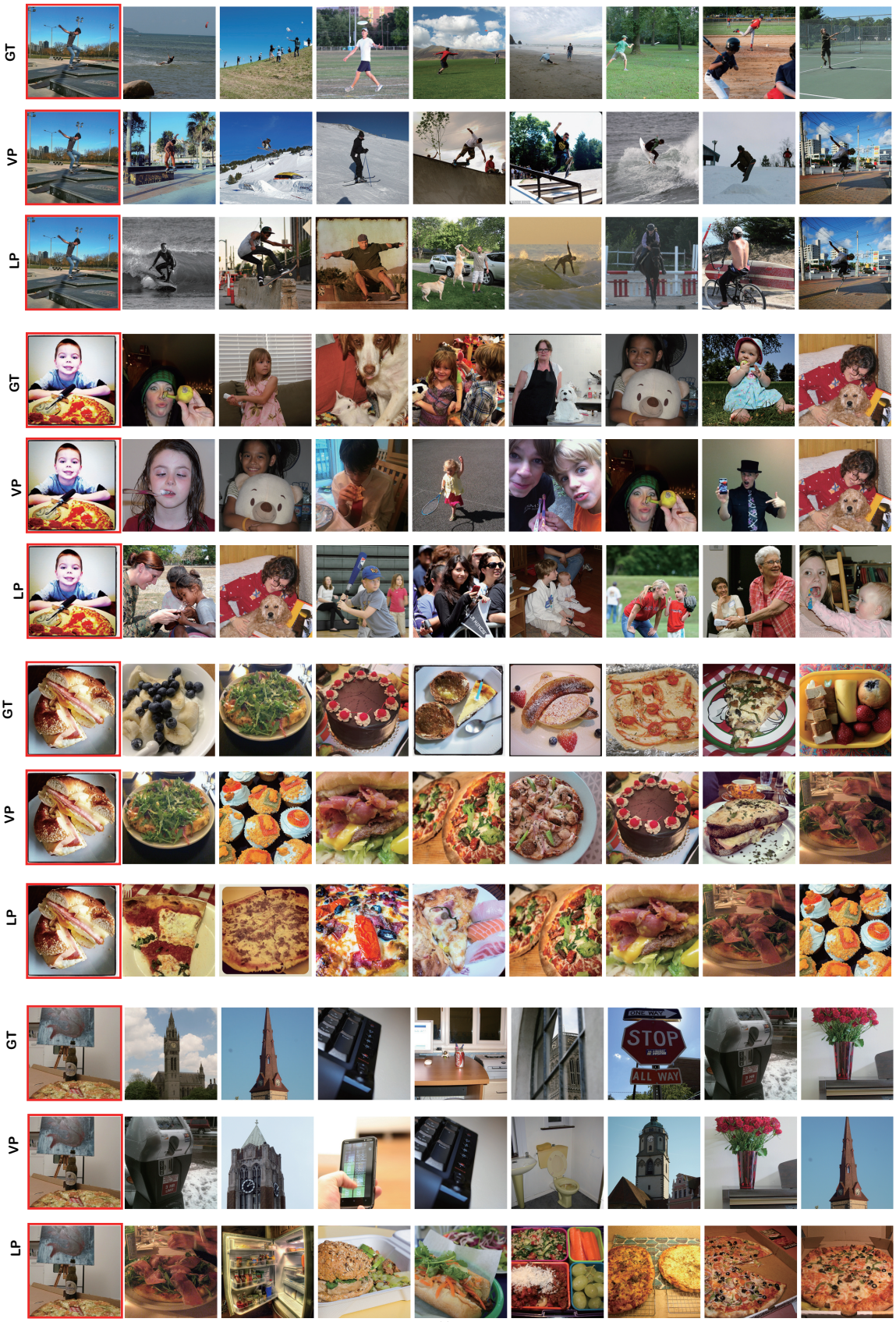

Fig. III: Image retrival with real-captured fMRI signals and generated ones. `GT` indicates `Ground Truth` signals, which are real captured data from the NSD dataset. `VP` and `LP` are abbrivations of `Visual Prompt` and `Layout Prompt`, respectively.

the subject-wise prediction difference, the Pearson correlation coefficients of divided visual ROIs are subtracted by the mean Pearson coefficients of the whole visual cortex, denoted as the relative predictive accuracy $\Delta\rho_c$. We illustrated the relative prediction differences for the `prf-visualrois` and `streams` ROIs in Fig.IV(a)-(b), respectively. The `prf-visualrois` group focuses on primary visual cortex, in which the fMRI vertices are allocated into 8 ROIs, marked as `V1v, V1d, V2v, V2d, V3v, V3d, hV4` and the remaining `Unk(prf-vis)`. Across 8 subjects, most of the primary visual ROIs show higher $\Delta\rho_c$ with visual prompts, yet lower ones with layout prompts. It substantiates that fMRI vertices falling outside of primary ROIs shows better stability when only high-level semantics are reserved. Moreover, the performance differences between the visual and layout prompts of the last `prf-visualrois` ROIs (`V3v, V3d, hV4`) decreases less than the primary `V1v, V1d, V2v, V2d` ROIs, validates their higher relevance with conceptual layouts. The `streams` group is broadly partitioned by the processing streams in cerebral cortex. With layout prompts, progressive increase can be roughly observed along the stream `early, midventral, midlateral, midparietal, ventral, lateral`, in accordance of the levels of visual hierarchy [1]. On right hemisphere (RH), the prediction for fMRI vertices within `lateral` ROIs is statistically better, indicating the right superiority of visuospatial processing [2]. On left hemisphere (LH), the `midventral` ROIs perform better generation with visual prompts, consistent with left lateralization of the functional regions such as `OWFA` [3]. The left and right hemispheres show specialization for visual processing, but the lagging of generative capability on `parietal` ROIs suggests relatively low relevance with visual clues.

We computed R^2 values between generated fMRI signals and the real ones, to measure the impacts of prompts. In Fig.3b of main paper, we have presented the scatters of R^2 values inside visual ROIs along the processing `streams`. In this supplementary material, Fig. V additionally presents the R^2 distributions of vertices inside different functional ROIs. X-axes denote the R^2 values of layout promptable generation and Y-axes denote those of visual promptable generation. In this first row, the R^2 values of vertices inside `floc-bodies` are provided. In the second row, we present the scatters of R^2 values for vertices inside `floc-faces`. Similarly, the R^2 scatters for `floc-places` and `floc-words` are provided in the third and fourth rows. As the information processing proceeds, the effects of visual and layout prompts become close, near the $y = x$ line. The 2D vertex-wise scatters of distribute in clusters, aligning with the hierarchical processing of visual system. For `floc-bodies` ROIs, the scatters of `FBA-2` is appearantly near the $y = x$ line, indicatingf that the elimination of details marginally impact the fMRI generation. For `mTL-bodies` (medial Temporal Lobe-associated bodies), only the divisions of `S6` and `S8` were provided in the dataset. The R^2 values are consistently inferior, since the fMRI signals were triggered by visual stimuli while `mTL-bodies` facilitates multimodal information processing. For `floc-faces` ROIs, the scatters of `S1` are broadly separable. The vertices within `OFA` have low R^2 with layout prompts. The `FFA-2` is less impacted than `FFA-1`. For

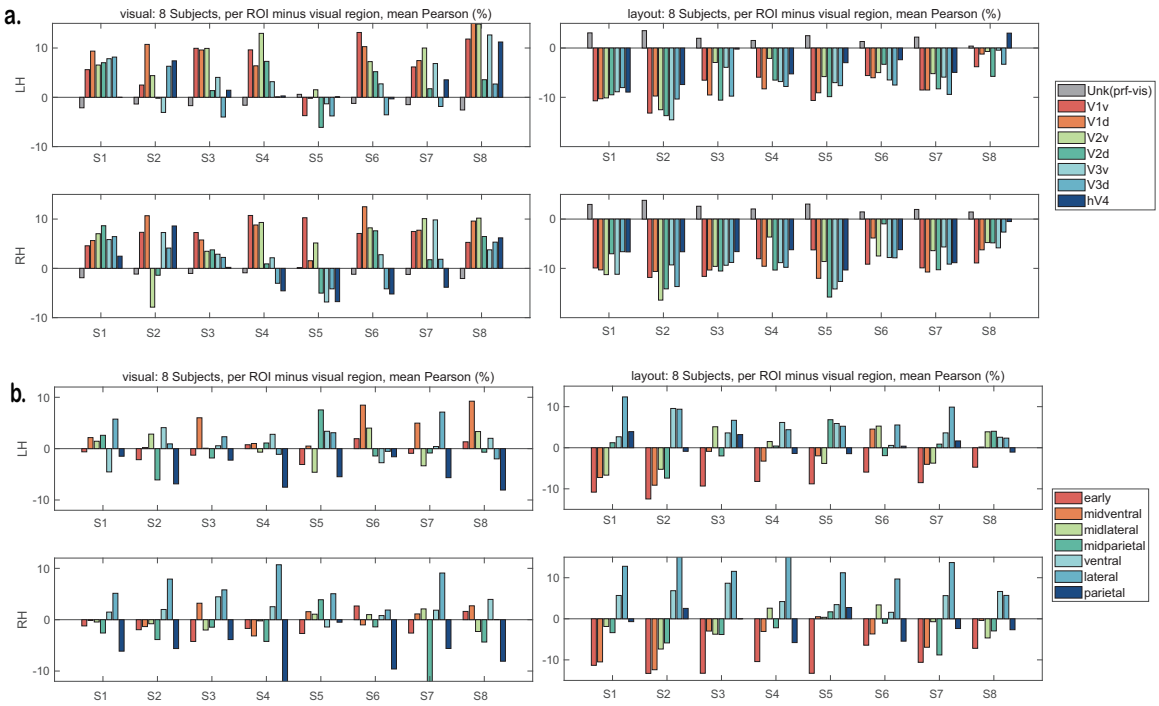

Fig. IV: Distributed statistics of NeoDiffuser based multimodal promptable fMRI generation in divided visual ROIs. (a). The prf-visualrois (primary functional visual ROIs) regions show higher prediction accuracy with visual prompts, but lower one with layout prompts. **(b).** The visual ROIs of processing streams show diverse prediction performance with multimodal prompts.

`floc-places` ROIs, the clusters of `S1` and `S5` are similarly distributed, with the scatters of `RSC` closer to $y = x$ line, indicating stablity to prompt changes. The R^2 values of `S7` and `S8` are relatively lower. For `floc-words` ROIs, the `OWFA`, `VWFA-1` and `VWFA-2` are generally seperable for `S1`, `S5`. But the fMRI vertices from `mfs-words` are different for subjects. For `S5`, the cluster of `mfs-words` is broadly distributed, but it is densely clustered for other subjects.

1.2 More visualization on impacts of positional representations

In Fig.4 of main paper, we present the visualization of sensitivity to positional features of `S1`, `S2`, `S4`, and `S5`. The sensivitiy is measured by the R^2 values between the generated fMRI signals with prompt codes from different networks. For the basic experiment, we designed the first kind of network, which adopts *FC (fully-connected) layer* to integrate the visual features from spatially-distributed positions, denoted as "w positional features". For the controlled experiment, the network leverages the *MAXPOOL (max pooling) layer* to extract the salient visual features yet ignore where these features locates, denoted as "w/o positional features". By these designs, the

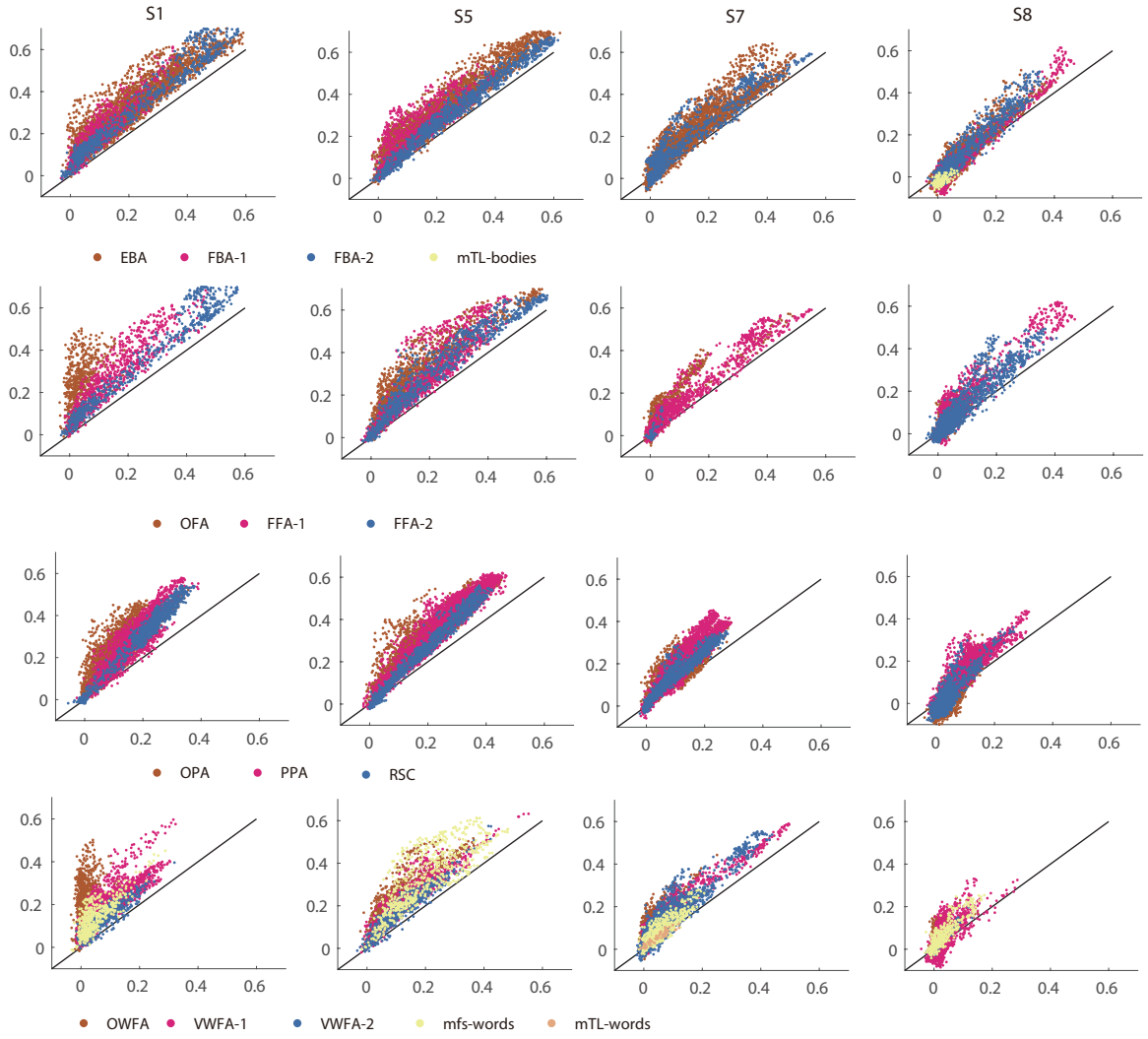

Fig. V: Scatters of R^2 distribution with multimodal promptable fMRI generation in functional divided visual ROIs. The R^2 values are computed between the generated fMRI signals and real-captured ones (ground truth). X-axes are the R^2 values of layout promptable generation and Y-axes are those of visual promptable generation. Each point in the scatter corresponds to one single fMRI vertex. In the first row, the scatters of `floc-bodies` ROIs (`EBA`, `FBA`, etc) are presented. In the second row, the scattes of `floc-faces` ROIs (`OFA`, `FFA`, etc) are shown. The scatters of `floc-places` and `floc-words` are presented in the third and fourth rows. Generally, the R^2 values within functional regions distribute in clusters. The clusters near $y = x$ lines indicate comparable prediction accuracy with layout, visual prompts.

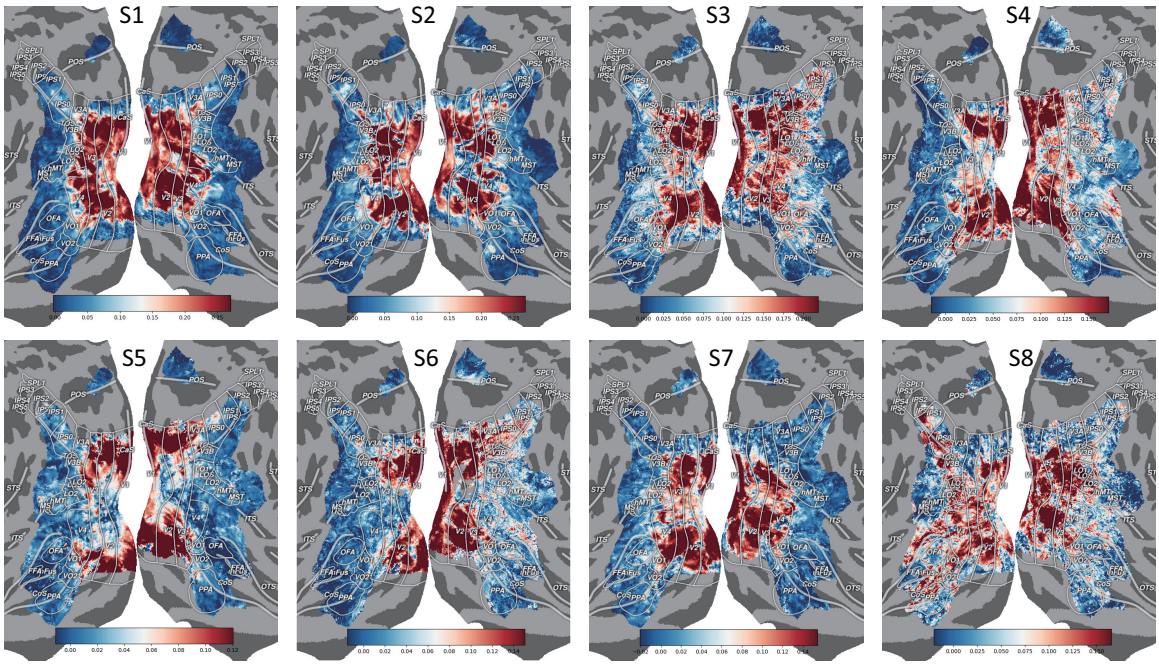

Fig. VI: Sensitivity to positional representations, measured by R^2 values between generated fMRI signals, based on visual promptable codes with different structures.

positional representations are reserved or eliminated, as the controlling factors in computational models, providing a parallel way to investigate the effects of positional features in brain responses. We further provide such visualization of all the 8 subjects in Fig. VI of this supplementary material, to illustrate the individual diversity. Generally, `S1`, `S2`, `S7` manifest the similar sensitivity maps, while `S3`, `S5`, `S6` are analogous. The difference lies in the middle regions of ventral and dorsal areas. Despite of the individual differences, it is the consistent characteristic that the fMRI vertices corresponding to early visual processing benefit more from geometrical representations, *i.e.*, spatial features. The retrosplenial cortex (`RSC`) consistently show low dependence on geometrical representations during the fMRI generation process.

In Fig.4e of the main paper, we provided the bubble chart of positional sensitivity (measured by R^2 values) within visual ROIs, with `S1` for example. In this supplementary material, we further presented the positional sensitivityn of visual ROIs for `S1`, `S2`, `S6` for example, as in Fig.VII. For `S1`, `S2`, the primary visual ROIs (from `V1d, V1v` to `hV4`) and `OFA`, `OWFA` are distinguished from other functional ROIs. The fMRI vertices inside `RSC` of left hemisphere are consistenly stable when inherent positional features are eliminated from prompt codes. We added the bubble chart of `S6` for example, with the division of `mTL-words` and `mTL-bodies`. The `mTL-words` (medial Temporal Lobe-Word Object Responsive Domain), which is functionally specialized for preferentially responds to written words stimuli, exhibits sensitivity to positional

features. But the fMRI vertices within mTL-bodies are stable when positional features are ignored.

1.2.1 Case study: floc-places ROIs

There are several functional areas in visual cortex, which can be evoked with place-related stimuli, denoted as floc-places. Typical ROIs include Occipital Place Area (OPA), Fusiform Place Area (FPA) and Retrosplenial Cortex (RSC). The Occipital Place Area (OPA) is a functionally specialized cortical area located in the dorsal occipitoparietal cortex, playing a central role in visuospatial navigation and scene perception, selectively responding to visual stimuli depicting real-world environments (*e.g.*, landscapes, rooms, and corridors) rather than isolated objects or faces. The Parahippocampal Place Area (PPA) is a functionally specialized region located in the posterior parahippocampal gyrus, exhibits selective responsiveness to visual stimuli depicting environments. The retrosplenial cortex (RSC) is a critical but often overlooked region located at the posterior cingulate gyrus, bridging the parietal and medial temporal lobes. Functionally, it serves as a key hub in the brain's navigation and memory systems, integrating spatial information from the hippocampal formation with perceptual inputs from visual and parietal cortices.

Considering the importance and spatial relevance, we provided the case study on floc-place ROIs, especially the RSC. In Fig.4e of main paper, we presented the bubble chart of R^2 sensitivity to positional features of different ROIs of S1. It is noteworthy that the vertices in PPA are mostly and consistently sensitive to positional features, showing high mean value and low variance. In contrast, the vertices in RSC are with high mean R^2 but also high variance. To investigate this, we provided an additional study of investigation. In Fig. VIII, the distribution of R^2 values of OPA, PPA and RSC from two subjects (S1, S2) are presented. The key observations include the following: (1) Different from other functional occipital areas such as face-related OFA, visual-word-related OWFA, the OPA shows tolerance to the elimination of positional representations. Specially, OPA is a functional area for depicting scenes and environments. The main reason is that the setting of *w/o positional features* in controlled experiment mainly eliminates the object-associative spatial information. In deep learning, MAX-POOLING is a common network module to extract the global visual information, widely-used in scene recognition of deep vision tasks. The scene information has been involved, yielding the low sensitivity of OPA. From the anatomical standpoint, the OPA locates in occipitoparietal area (dorsal stream), not close to the occipital ROIs (OPA and OWFA). The regional difference in cerebral cortex provides another explaination of OPA-tolerance. (2) The vertices within PPA are consistently with high R^2 value, especially for S1. The low variance indicates the functional segregation of PPA. (3) Although the mean R^2 of PPA and RSC is close, the histogram distribution is different. About 85% of vertices have high-concentrated R^2 values, while the other 15% vertices has uniformly-distributed R^2. It suggests the functional

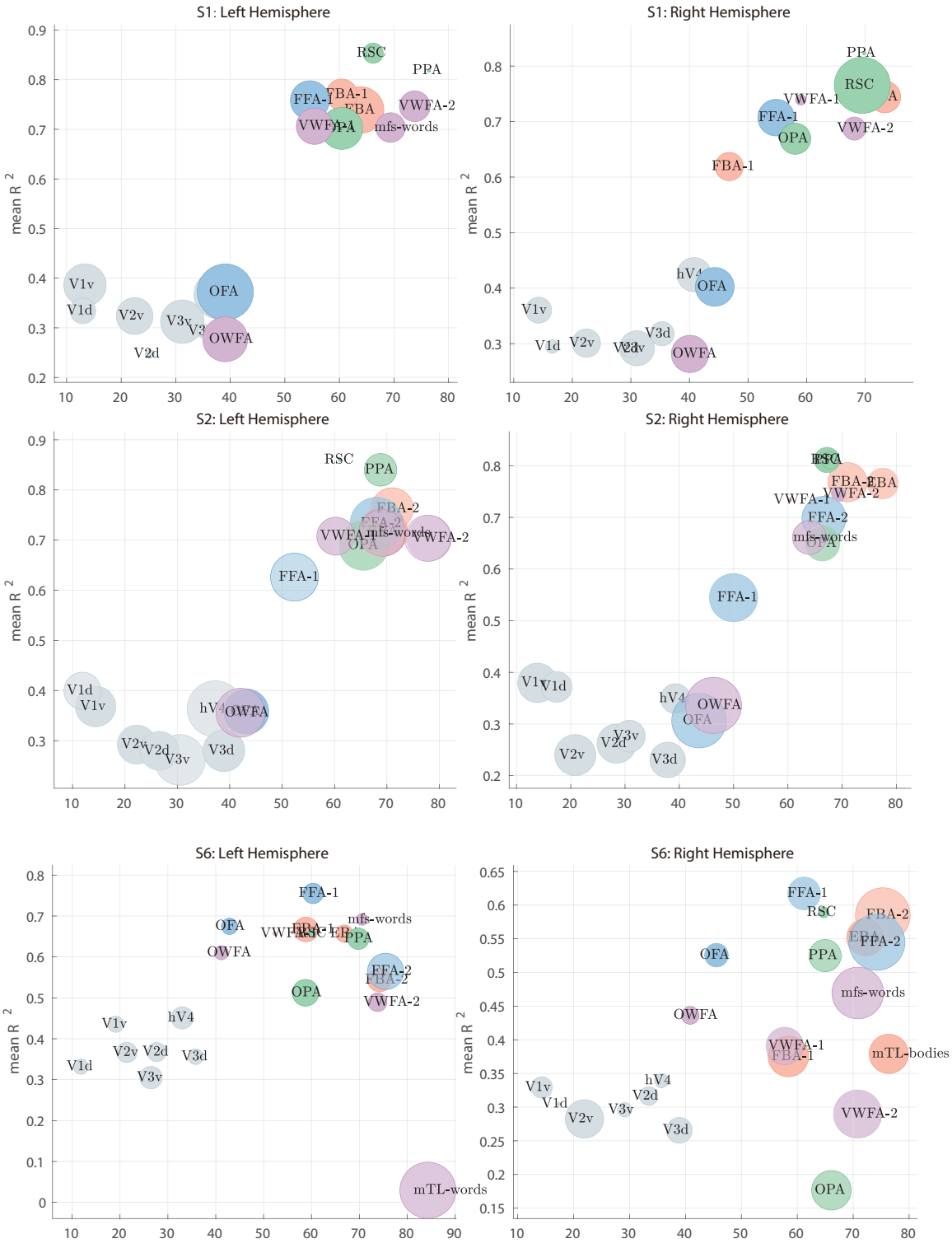

Fig. VII: For divided visual ROIs, the sensitivity to positional representations (measured by R^2 values between generated fMRI signals) is different, yet in alignment with visual hierarchy.

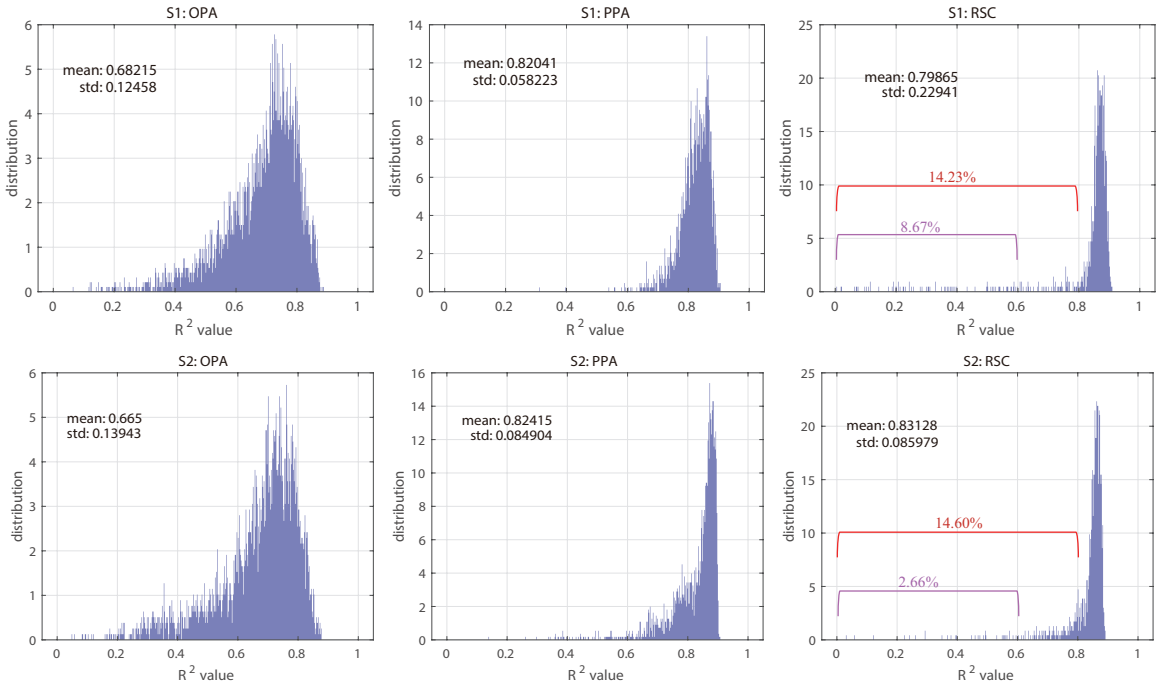

Fig. VIII: Histogram of position-sensitivity (measured by R^2 values) within `floc-places` ROIs.

divergence of `RSC`, supporting the importance and functional complexity of `RSC`.

Fig. IX provides the zoomed visualization of R^2 values within `RSC`. The R^2 values are calculated between the visual codes of two settings (as illustrated in Fig.4a of main paper, the first setting is *with positional representations* and the other setting is *without positional representations*), measuring the prediction difference. In comparison, Fig. VI visualizes the prediction performance difference, computed by the prediction accuracy (measured by Pearson correlation coefficients between generated signals and real-captured ground truth) of *with positional representations* ρ_p minus that of *without positional representations* ρ_n. As observed in Fig. VI, the prediction accuracy shows marginal difference, *i.e.*, $\rho_p - \rho_n$ is relatively smooth and low within `RSC`. However, as found in Fig. IX, the R^2 values exhibit significant difference. The fMRI voxels with high R^2 (high similarity, low difference) are colored red, while those with low R^2 (low similarity, high difference) are colored blue. The comparative visualization indicates that the fMRI vertices within `RSC` perform differently along with the positional representations, suggests the fine-grained functional specialization.

1.3 More visualization on hemisphere specificity impacts

In Fig.5 of the main paper, we present the visualization of the differences between separating and sharing the prompt codes. The performance changes caused by replacing the hemisphere-specific codes with sharing codes are recorded as the sensitivity. We

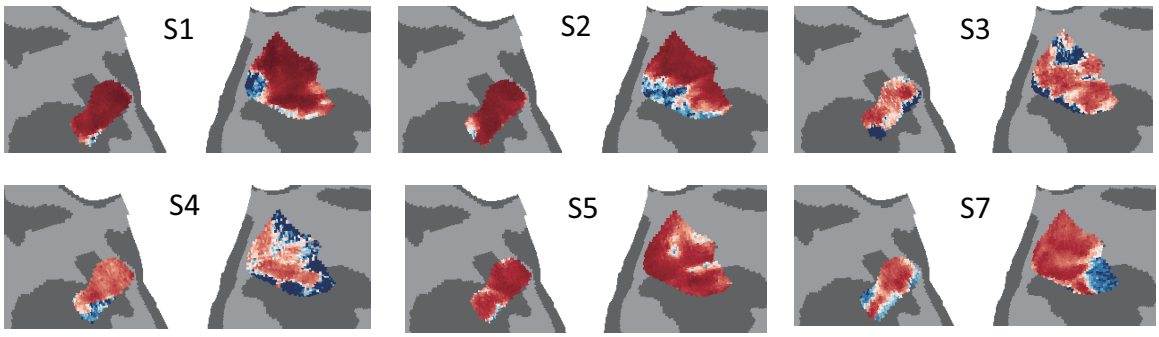

Fig. IX: Effects of positional representations in RSC ROIs.

further provide additional visualization maps across all the 8 subjects. The sensitivity maps of layout prompting are provided in Fig. X(a) and those of visual prompting are provided in Fig. X(b). For visual promptable generation, the sensitivity maps are similar across most of the subjects. Only S8 displays more disperse vertices and slightly weak difference in primary areas. Generally, the sensitivity to feature sharing is higher in ROIs corresponding to primary processing, especially V1, V2, validates the hemisphere lateralization deriving from the receptive fields of monocular vision. Compared to positional sensitivity, the impacted regions are more concentrated. For layput prompting, the impacted ROIs are dispersely distributed. But it is interesting that the left and right hemisphere show quite asymmetric sensitivity. For example, the fMRI prediction of S2 obtains accuracy gains in the right hemisphere but shows performance drops in the left hemisphere. It is opposite for S8. More exploitation is required to explain such individual differences.

Fig.XI provides the detailed performance degeneration of divided cortical ROIs over the tested subjects. The performance degeneration is measured by the mean Pearson coefficients of separate prompts minus that of shared prompts, denoted as $\Delta\rho_c$. Fig.XI(a) presents the barred accuracy difference of partitioned ROIs by visual prompting. The prf-visualrois ROIs show consistent generative degeneration and most of prf-visualrois ROIs inside left hemisphere obtains more gains with prompt segregation. The degeneration caused by ignoring hemisphere specificity gradually weakens along the processing streams, exhibits the decrease trends. Despite the individual differences in visual promptable generation, the ROIs identified by functional localizer (floc- ROIs) also show consistent performance drop. It supports the viewpoint that the brain modularity is a matter of degrees [5]. Against the other ROIs, the fMRI generation of mTL-words ROI benefits from prompt sharing, yet only one subject provides this ROI partition (S6). Besides, the ROIs in occipital cortex areas (OFA, OPA, OWFA) consistently suffer from the prompt-sharing generation. The hemisphere specificity with the layout promptable generation is more complicated. Compared to visual prompting, the average performance drop caused by prompt sharing is less and pronounced individual diversity over subjects can be observed. For

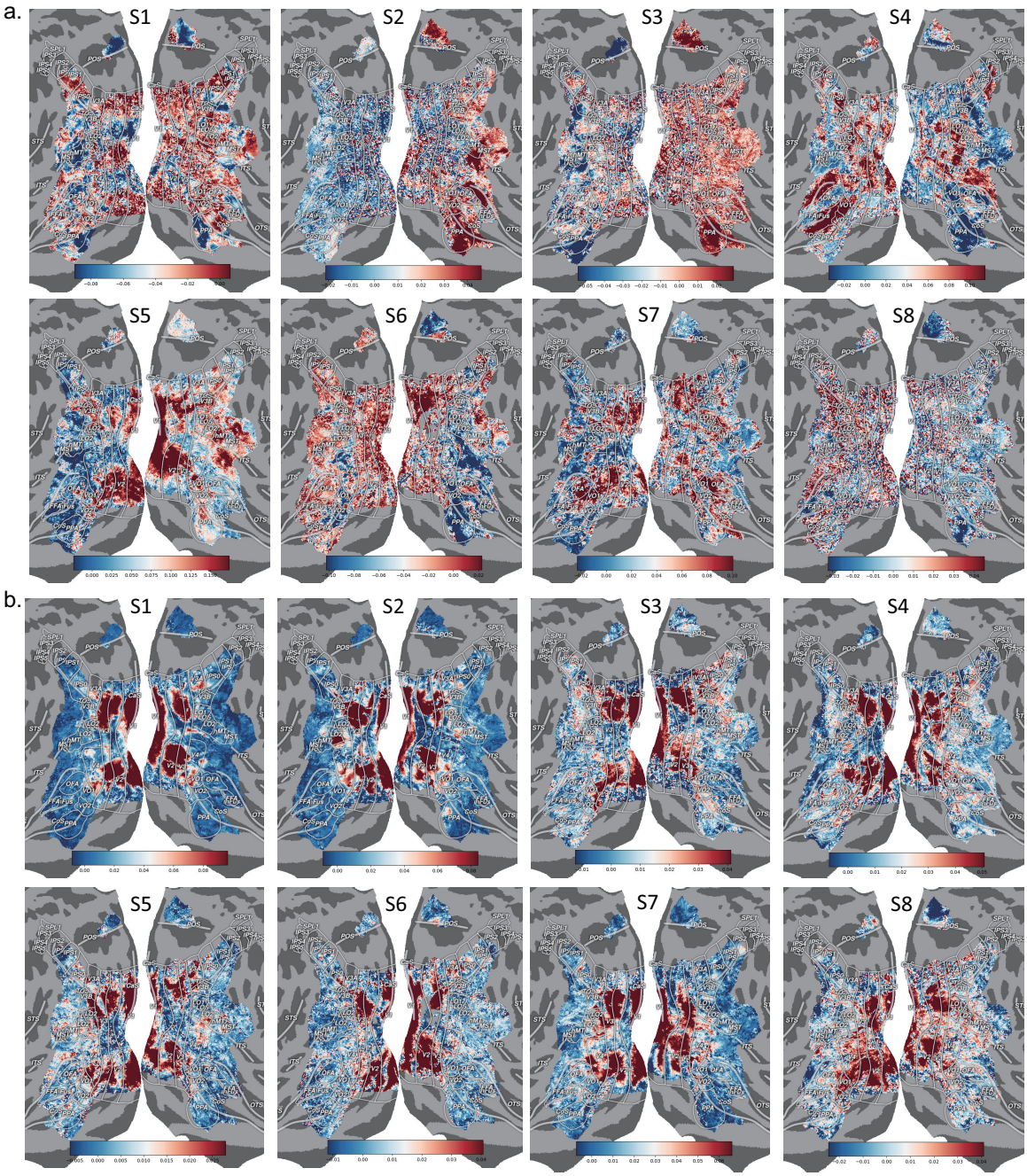

Fig. X: More visualization for 8 subjects to illustrate the hemisphere specificity in fMRI generation. (a) For layout promptable generation, the sensitive regions are distributed widely. (b) For visual promptable generation, compared to sharing prompt codes, separating the prompt codes significantly promote the fMRI generation in low-tier processing regions.

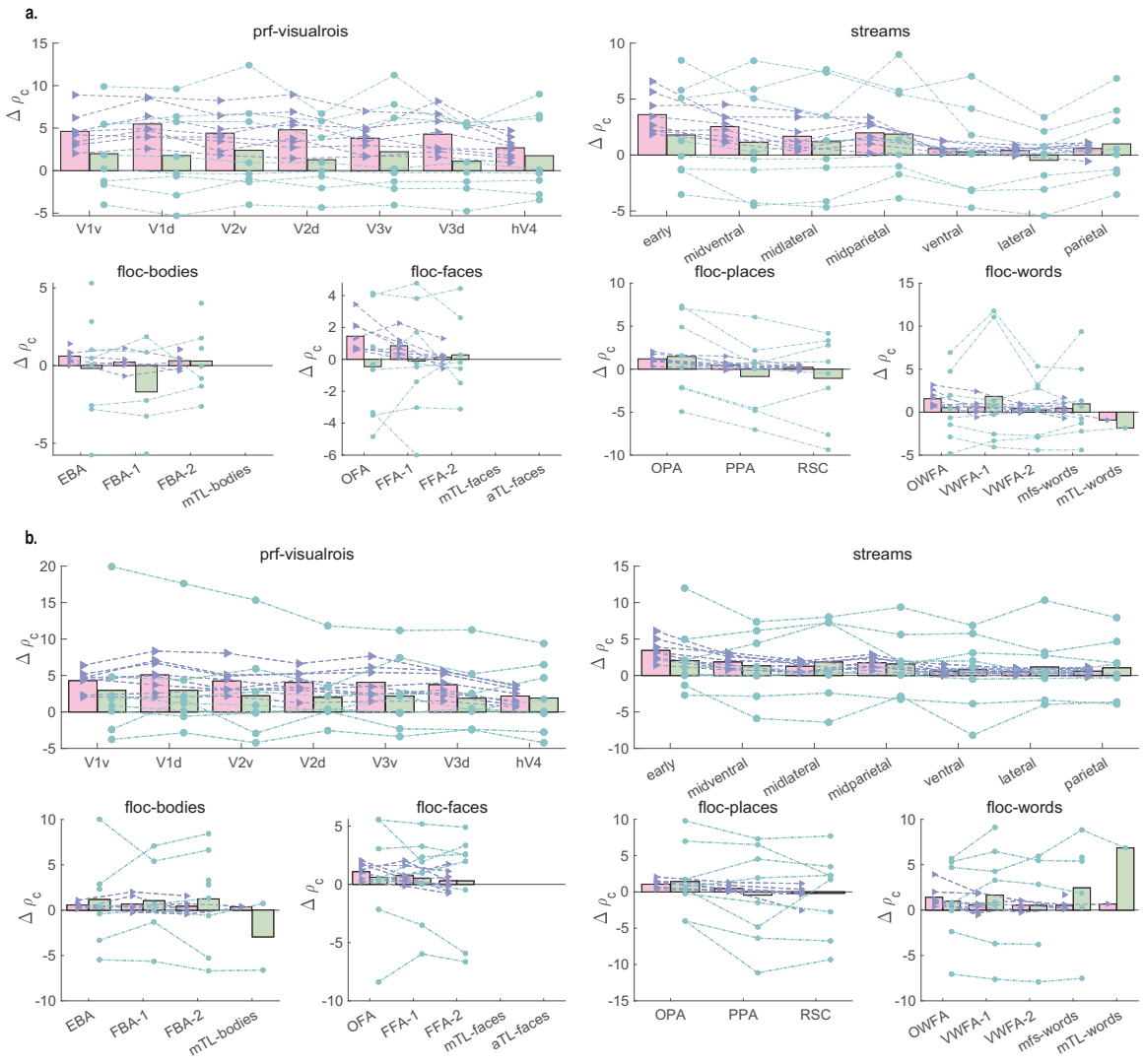

Fig. XI: Effects of sharing or separating conditions for different hemispheres. **a.** For visual prompt, using hemisphere-separate prompt codes consistently increases the prediction accuracy over visual ROIs. Notably, the visual areas attached to the primary visual processing are more required with separate prompt codes. **b.** For layout prompt, separate prompt codes obtains average prediction improvement. But whether the performance improves varies over the subjects.

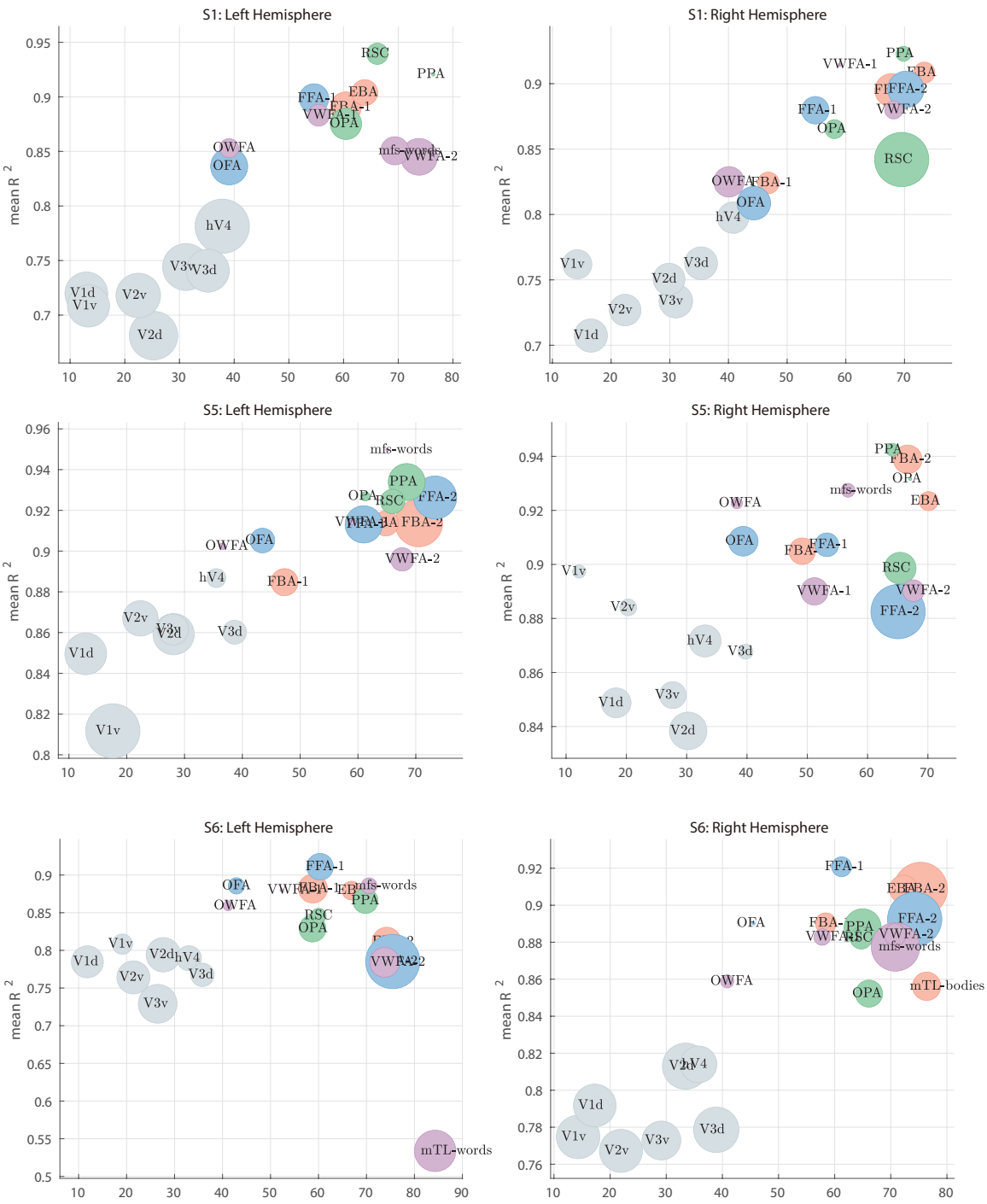

Fig. XII: Sensitivity to prompt code sharing, measured by R^2 values between generated fMRI signals based on sharing or seperate visual promptable codes.

prf-visualrois ROIs and streams-partitioned ROIs, generative degeneration can be generally observed. Despite of the individual diversity, the generation increase or reduction is roughly consistent for each subject across distributed ROIs, reveals consistently positive or negative $\Delta\rho_c$. There are several floc-ROIs obtain the generative gains with sharing layout prompts, which is reasonable from data learning perspective (abundant data benifit data learning). Combining Fig.XI(a)-(b), there is transitional changes of performance gain. For layout prompting, the right hemisphere exhibits higher generative gain by code separation. Oppositely, the left hemisphere obtains higher performance gain with separate guidance in visual prompting. This might stem from the causality and information preference inherently consisted in brain computation. For example, the verbal cognition is left-lateralized for most people but its required low-level processing such as character focusing is right-lateralized [6, 7].

Similar to positional features analysis, we provided the bubble chart of R^2 distribution for divided visual ROIs. The channel-wise R^2 values are calculated between the generated fMRI signals with sharing or seperate visual codes. For each divided ROI, the mean and standard derivation are computed over the R^2 values for included vertices. As in Fig.XII, the sensitivity to code sharing is less significant than that to positional features, yielding high R^2 scores. Despite of the range difference, the distribution of visual ROIs resembles with the distribution of positional sensitivity. For S1 and S5, the R^2 values of primary visual ROIs (from V1d, V1v to hV4) are smaller, indicating that the primary visual regions require specific guidance.

1.4 Comparison over influential factors

The proposed NeoDiffuser works as a computational model for fMRI generation. Several influential factors such as prompt settings, positional features, and code sharing are explored. We further compare the effects of the above-mentioned three influential factors in Fig. XIII. The mean R^2 values are computed between the generated fMRI signals and the reference ones with visual prompts within divided visual ROIs. In Fig. XIII(a), the mean R^2 values of functional ROIs of left hemisphere (LH) and right hemisphere (RH). The tendency of R^2 changes is generally consistent, validating that the prompt settings, positional features and code sharing are external manifestations of visual hierarchy at different levels. For primary visual ROIs, the hV4 is less impacted, showing high R^2 values. For visual ROIs related to faces (floc-faces), the OFA from both hemispheres shows higher sensitivity (lower R^2) than FFA-1, FFA-2. For visual ROIs related to places (floc-places), the OPA from both hemispheres show higher sensitivity (lower R^2) with prompt changes (from *visual prompts* to *layout prompts*) and positional features, yet the PPA is stable with all the three test factors. For ROIs related to visual words (floc-words), the OWFA show higher sensitivity under the changing conditions of prompt settings and positional features, but relatively stable with seperate or sharing codes.

In Fig. XIII(b), the sensitivity measured by R^2 values are computed along

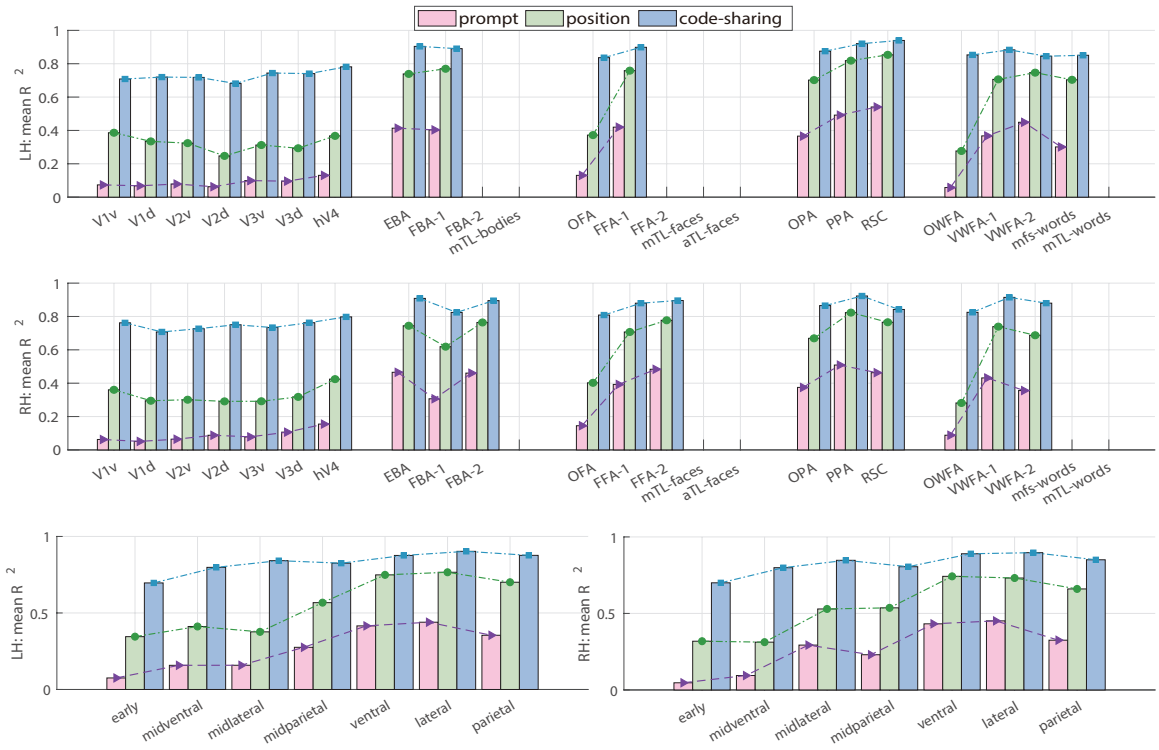

Fig. XIII: Comparison over different influential factors for visual promptable generation."LH" indicates left hemisphere, and "RH" indicates right hemisphere .

processing `streams`. There are three processing streams, denoted as V-stream (`early`→`midventral`→`ventral`), L-stream (`early`→`midlateral`→`lateral`), P-stream (`early`→`midparietal`→`parietal`). For the three processing streams, the R^2 values gradually increase along, indicating the decreasing impacts from the factors. For the test subject `S1`, the `parietal` regions are prone to be affected by the factors such as prompt settings.

1.5 Cross-modality alignment for multimodal fMRI generation

The proposed NeoDiffuser is built based on diffusion model, combined with multimodal machine learning (MMML). From the perspective of computer science, it is important to illustrate the impacts of aligning different modalities, *i.e.*, cross-modality alignment. In particular, since the model is required to generate fMRI signals with multiple kinds of prompts, we designed and explored the cross-modality alignment (CMA) by imposing the alignment loss on the projection vectors of brain activity data and prompt codes. Technically, cross-modality alignment is performed by producing the same-dimensional code with fMRI codes (from original fMRI signals) and imposing contrastive loss between the fMRI codes and the associated prompt codes for generation. We first perform the comparative study on visual promptable generation. The prediction

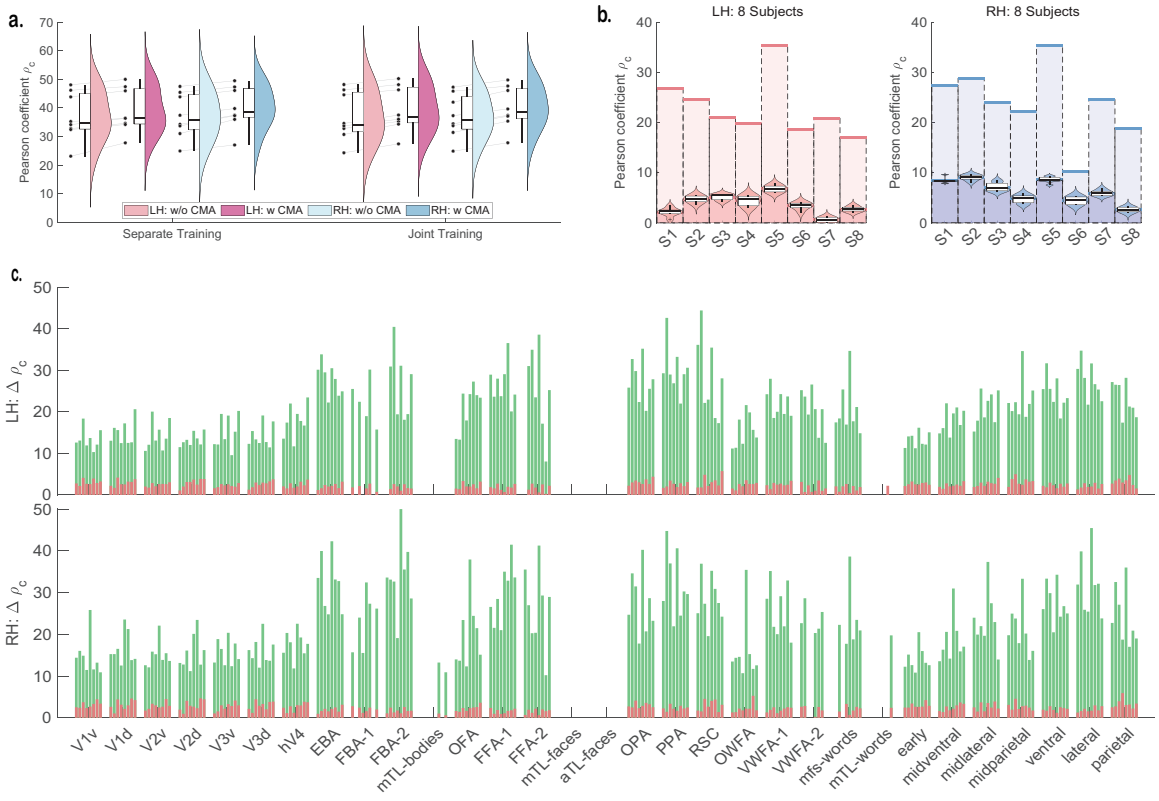

Fig. XIV: Effects of cross-modality alignment for fMRI generation. **(a).**With visual prompts, the cross-modality alignment boosts the generation performance. **(b).** With layout prompts, cross-modality alignment is essential to produce meaningful and stable prediction. **(c).** The difference of Pearson correlation coefficients ($\Delta\rho_c$) across 8 subjects show the importance of cross-modality alignment.

accuracy, measured by Pearson correlation coefficients between generated fMRI signals and ground truth ones, are computed for left and right hemispheres, respectively. The prediction accuracy across different subjects with/without cross-modality alignment are presented in Fig.XIV(a). We equip the CMA for both separate training and joint training on left and right hemispheres, respectively. For all the 8 subjects, the cross-modality alignment improves the channel-wise prediction accuracy by around 3%. It indicates that aligning the feature space between different modalities, as visual images and brain data, enables the generation of coherent and meaningful outputs, leading the performance of visual promptable fMRI generation.

Further, we performed the cross-modality alignment on layout prompts. The measured Pearson correlation coefficients are presented in Fig.XIV(b). The cross-modality alignment is essential for controllable generation with extremely sparse prompts. Compared to visual prompting, the layout prompting shows high dependency on cross-modality alignment. Without CMA, the Pearson coefficients measure the similarity between generated fMRI signals and ground truth ones are mostly distributed

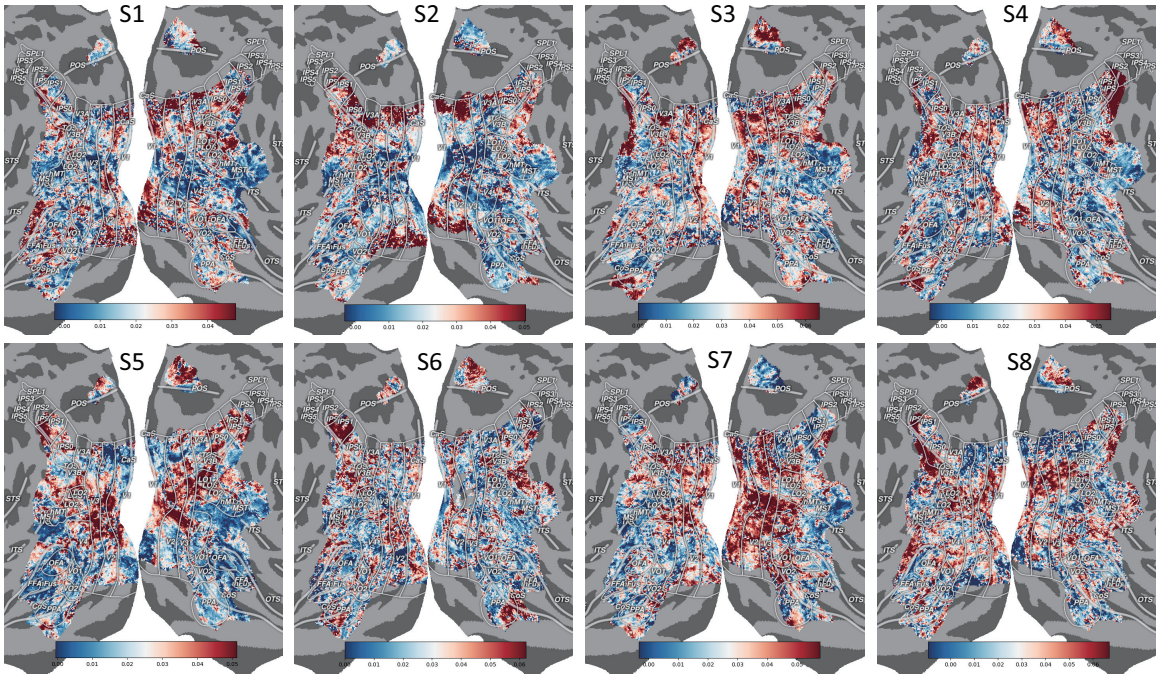

Fig. XV: Visualization of cross-modality alignment effects for visual promptable generation. Compared to sharing prompt codes, separating the prompt codes significantly promote the fMRI generation in low-tier processing regions. The regions which benefit from cross-modality alignment distribute in scatters in the visual cortex.

below 10%. With CMA, the prediction accuracy measured by Pearson correlation coefficients reach 16%∼33%. From the technical view of multi-modality learning, the cross-modality alignment helps the model to adapt the different desired conditions in generative learning. The cross-modality alignment can align the feature codes from different modalities of data, further alleviating the fitting burden of deep neural networks. By the process of aligning the representations of different modalities in a shared latent space, the cross-modality alignment facilitates the transfer of knowledge between modalities and enables generative models to leverage information from fMRI modality and improve the generation quality.

In Fig. XIV(c) of this supplementary material, we present the performance drop (*i.e.*, sensitivity) caused by CMA removal, measured by the difference with and without CMA across divided ROIs. The performance drop in layout prompting is barred in green, while that in visual prompting is barred in red. With layout prompts, the performance drop caused by CMA removal is more significant in functional localizer (`floc-`) ROIs, corresponding to higher-tier visual processing. The performance drop shows compatibility with the visual processing levels. For visual prompting, the performance drop is relatively smooth along the divided ROIs. This is further verified by the visualization of CMA sensitivity in Fig. XV. The fMRI vertices of signals

benefiting from CMA are scattered and distributed in a disperse way, meanwhile show individual specificity.

1.6 Visualization on clustered groups

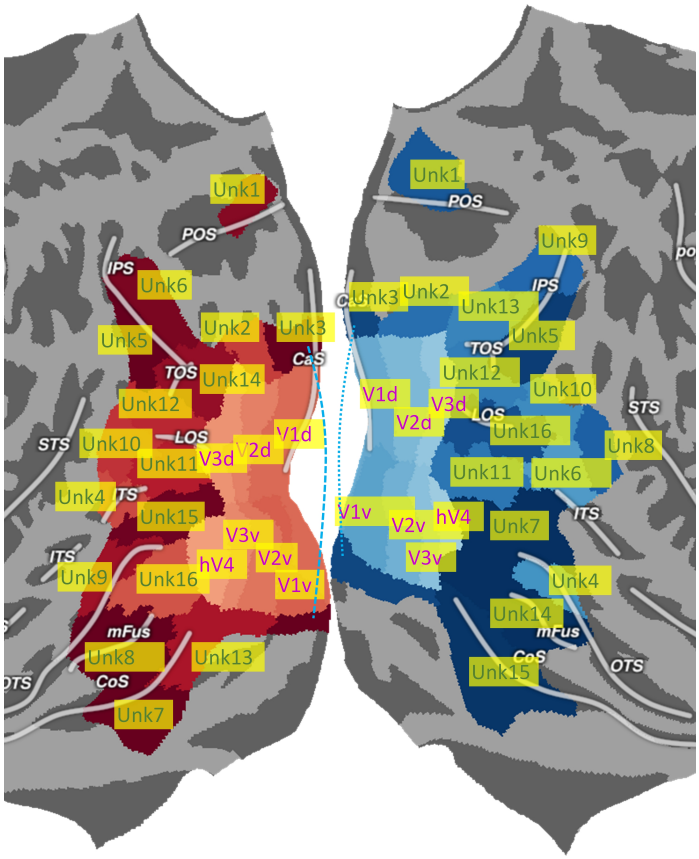

Fig. XVI: Visualization on the clustered regions for correlation analysis with balanced vertices.

Fig. XVI provides the clustered regions for correlation analysis in Fig.7c-d of the original manuscript, serves as the supplementary for the correlation analysis. For example, the `Unk1` which exhibits low association with other clustered ROIs inside non-primary areas is verified as overlapped with `RSC`. It has stronger connection with `Unk2` of left hemisphere (LH) and `Unk3` of right hemisphere (RH), which are near the primary areas `V1d`∼`V3d`. For the right hemisphere, `Unk7`∼`Unk10`, which manifest low association with primary ROIs, locate near the outside boundary of visual cortex, in accordance with the processing streams [8]. We also analyze the signal specificity in Fig.7d of the main paper, denoted as the recovery accuracy for other ROIs subtract that of being recovered by other ROIs. As depicted before, the primary cortex ROIs

carry out low data specificity, for which the recovery for other ROIs underweighs the being-recovered by other ROIs. Besides, the clustered ROIs with high data specificity locate in the middle regions, such as `Unk11`∼`Unk16`. As the middle connected regions, the fMRI data are informational to recover the data from the low-tier primary ROIs and high-tier functional regions.

References

[1] M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2:1019–1025, 1999.

[2] B. Rossion. Damasio's error—prosopagnosia with intact within-category object recognition. *Journal of Neurophysiology*, 12:357–388, 2018.

[3] L. Strother, A. M. Coros, and T. Vilis. Visual cortical representation of whole words and hemifield-split word parts. *Journal of Cognitive Neuroscience*, 28(2): 252–260, 2016.

[4] E. J. Allen and et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022.

[5] X. Chen, J. Necus, L.R. Peraza, and et al. The functional brain favours segregated modular connectivity at old age unless affected by neurodegeneration. *Communications Biology*, 4:973, 2021.

[6] M. M. Mesulam. Large-scale neurocognitive networks and distributed processing for attention, language, and memory. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 28(5):597–613, 1990.

[7] S. J. Gotts, H. J. Jo, G. L. Wallace, Z. S. Saad, R. W. Cox, and A. Martin. Two distinct forms of functional lateralization in the human brain. *Proceedings of the National Academy of Sciences*, 110(36):E3435–E3444, 2013.

[8] D. J. Kravitz, K. S. Saleem, C. I. Baker, and M. Mishkin. A new neural framework for visuospatial processing. *Nature Reviews Neuroscience*, 12:217–230, 2011.