Table R1: Analyzing the contributions of $\hat{z}_c$ and a_r with and without VAE Features. Stable Diffusion 2-1, using text-only conditions, outperforms random $\hat{z}_c$. Using both $\hat{z}_c$ and a_r yields the best performance, showing their complementarity.

Method	# Models	Low-Level				High-Level			
		PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	EffNet-B ↓	SwAV ↓
		w/o VAE Feature							
$\hat{z}_c \sim N(0,1)$	1	.016	.203	58.6%	70.6%	87.0%	90.5%	.839	.455
$\hat{z}_c = 0$	1	.033	.209	67.5%	83.1%	93.1%	94.7%	.717	.359
a_r only, SD-2-1	1	.046	.264	72.3%	86.4%	**93.8%**	**96.4%**	.693	.414
w/o VAE feature	1	**.093**	**.263**	**84.5%**	**90.6%**	93.6%	95.7%	**.684**	**.398**
		w/ VAE Feature							
$\hat{z}_c \sim N(0,1)$	1	.203	.324	91.6%	96.3%	95.3%	93.9%	.713	.378
$\hat{z}_c = 0$	1	.216	.336	91.8%	96.9%	96.1%	95.3%	.694	.339
a_r only, SD-2-1	1	.257	**.358**	92.9%	**97.3%**	96.6%	96.1%	.656	.332
Our Method	1	**.265**	.357	**93.1%**	97.1%	**96.8%**	**97.5%**	**.633**	**.321**

Table R2: Performance on the QA task declines without fMRI embeddings, notably in Brain Caption and Detail Description, and to a lesser extent in Complex Reasoning. This highlights the importance of fMRI embeddings despite some contextual information leakage in Complex Reasoning.

Method	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE	CIDEr	SPICE	CLIP-S
	Brain Caption								
a_r Only	39.06	21.87	12.36	08.01	11.90	31.64	03.32	03.18	27.88
Original Model	**57.19**	**37.17**	**23.78**	**15.85**	**18.60**	**36.67**	**49.51**	**12.39**	**65.49**
	Detail Description								
a_r Only	27.15	11.57	4.40	1.42	12.21	21.72	1.17	2.56	25.98
Original Model	**38.91**	**24.02**	**15.24**	**12.41**	**18.44**	**27.83**	**42.58**	**18.41**	**56.16**
	Complex Reasoning								
a_r Only	55.70	43.52	32.25	24.61	21.32	38.41	136.41	43.21	63.24
Original Model	**65.41**	**59.61**	**50.68**	**36.46**	**34.46**	**62.60**	**217.83**	**60.29**	**80.96**

Table R3: Top: Different cross-subject alignment methods minimally impact stimulus reconstruction, showing our method's robustness. Bottom: Comparison with contemporary work, MindEye2. Our method outperforms MindEye2's cross-subject baseline and is compatible with MindEye2's subject-specific models.

Method	# Models	Low-Level				High-Level			
		PixCorr ↑	SSIM ↑	AlexNet(2) ↑	AlexNet(5) ↑	Inception ↑	CLIP ↑	EffNet-B ↓	SwAV ↓
		Comparison of different cross-subject alignment methods							
Nearest	1	.259	.354	93.2%	96.7%	96.6%	97.3%	.636	.334
Area	1	.264	.358	92.8%	97.1%	96.4%	**97.6%**	.634	**.318**
Nearest-Exact	1	.262	.353	93.1%	96.9%	96.7%	97.3%	.636	.336
Trilinear (Original)	1	**.265**	**.357**	**93.1%**	**97.1%**	**96.8%**	97.5%	**.633**	.321
		Comparison with MindEye2							
MindEye2	4	**0.322**	**0.431**	**96.1%**	98.6%	95.4%	93.0%	**0.619**	0.344
MindEye2 (unrefined)	1	0.278	0.328	95.2%	**99.0%**	96.4%	94.5%	0.622	0.343
Our Method	1	0.265	0.357	93.1%	97.1%	**96.8%**	**97.5%**	0.633	**0.321**