LLM	ViT	Average nearest neighbor distance	Wasserstein distance	Mean overall distance
Qwen3-0.6B	Aimv2	1.42	4.86	10.78
Qwen3-0.6B	SAIL-ViT	1.15	3.88	9.52
Qwen3-1.7B	Aimv2	1.30	4.65	11.11
Qwen3-1.7B	SAIL-ViT	1.05	3.60	9.89
Qwen3-8B	Aimv2	0.78	3.59	11.24
Qwen3-8B	SAIL-ViT	0.66	2.63	10.06
Interlm2.5-1.8B	Aimv2	1.17	4.14	10.99
Interlm2.5-1.8B	SAIL-ViT	0.96	3.18	9.79

Table 7: Comparison of LLM and ViT models on three distance metrics.

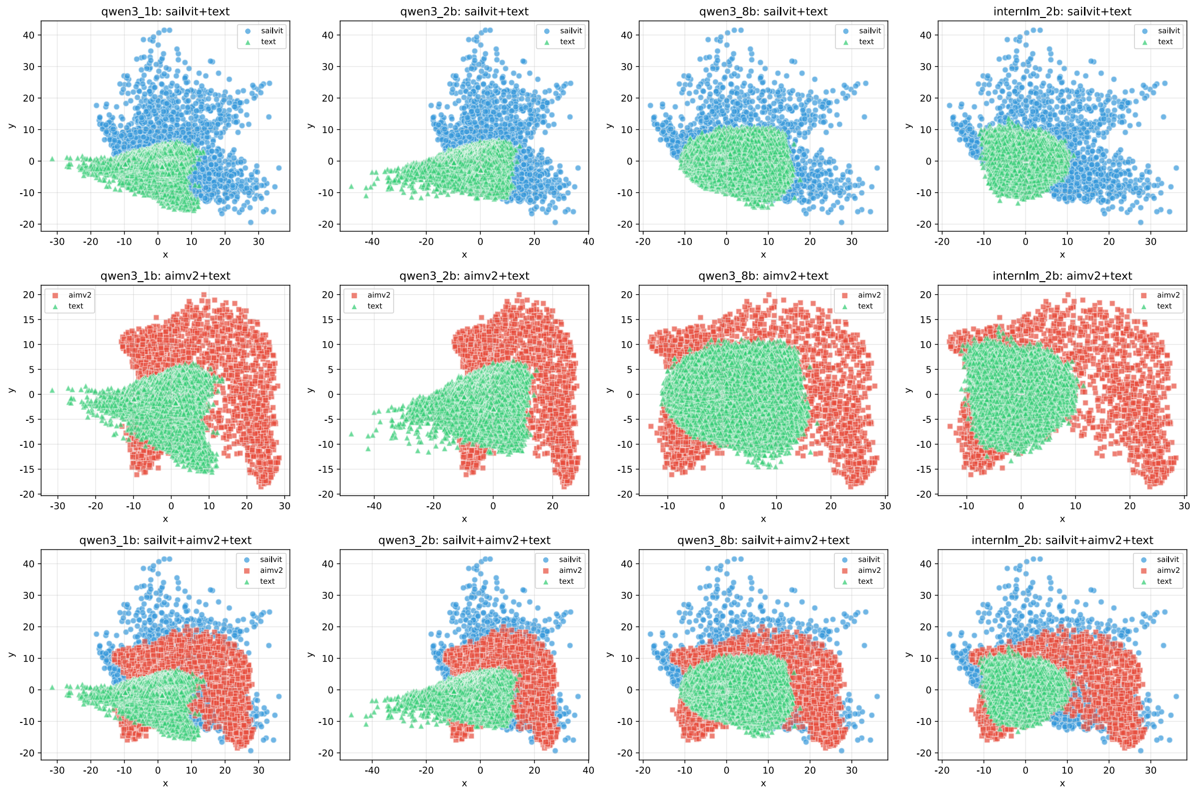

Figure 6: Visualization of token embedding distributions across different models and combination strategies. Each row corresponds to a model variant (Qwen3-0.6B, Qwen3-1.7B, Qwen3-8B, InternLM-1.8B), and each column illustrates a grouping strategy: combining SAIL-ViT with text embeddings, AIMv2 with text embeddings, or all three jointly. Colors and markers indicate 'SAIL-ViT', 'AIMv2', and 'text' tokens, highlighting spatial relationships and overlaps between embedding spaces.

Specifically, we calculate the distribution distance between visual features extracted by SAIL-ViT (and its baseline) and textual features from LLMs of different sizes and types. We randomly sampled five images from the internet, extracted their visual features using both SAIL-ViT and its baseline, and concatenated them to form a feature collection of size [5120, 1024]. Then, we applied Principal Component Analysis (PCA) to reduce the dimensionality to [5120, 2]. On the textual side, we extracted feature vectors from the lookup tables of different LLMs. We quantified the distribution distances by computing the average nearest neighbor distance, Wasserstein distance, and mean overall distance between these feature clusters. As shown in Figure 6, the visualization results show that the visual feature vectors extracted by SAIL-ViT are more compact and exhibit greater overlap with textual feature vectors, whereas the baseline model produces more dispersed visual features with less overlap with the textual space. Quantitatively, as shown in the table below, the visual feature space of SAIL-ViT is significantly closer to the textual feature space of LLMs across different sizes and architectures, as measured by multiple distance metrics. These results demonstrate that SAIL-ViT effectively reduces the gap between visual and textual feature spaces.