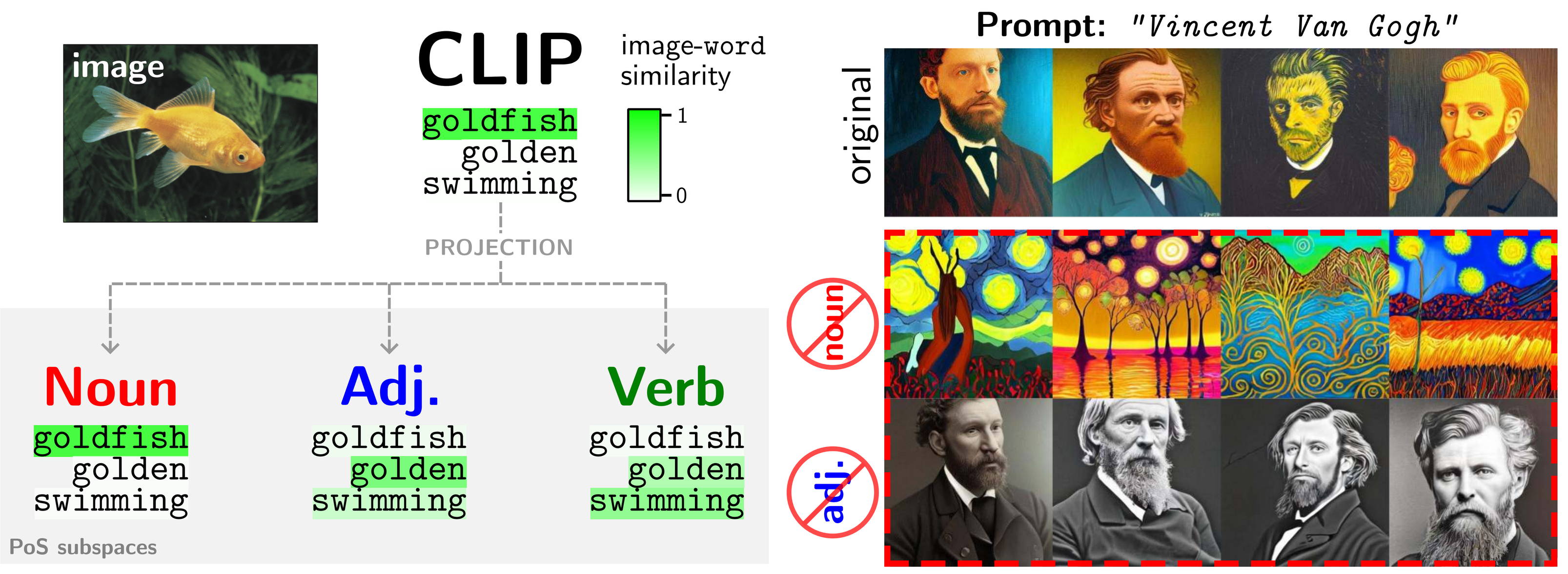
Parts of Speech–Grounded Subspaces in Vision-Language Models

James Oldfield [1] Christos Tzelepis [1] Yannis Panagakis [2,3] Mihalis A. Nicolaou [4] Ioannis Patras [1]

[1]Queen Mary University of London [2]National and Kapodistrian University of Athens [3]Archimedes/Athena RC [4]The Cyprus Institute



Summary

CLIP represents multiple visual properties in its embedding [2]. We leverage the association between PoS and specific visual modes of variation (e.g. nouns relate to objects, adjectives their appearance) to learn geometry-aware subspaces that better separate the constituent components.

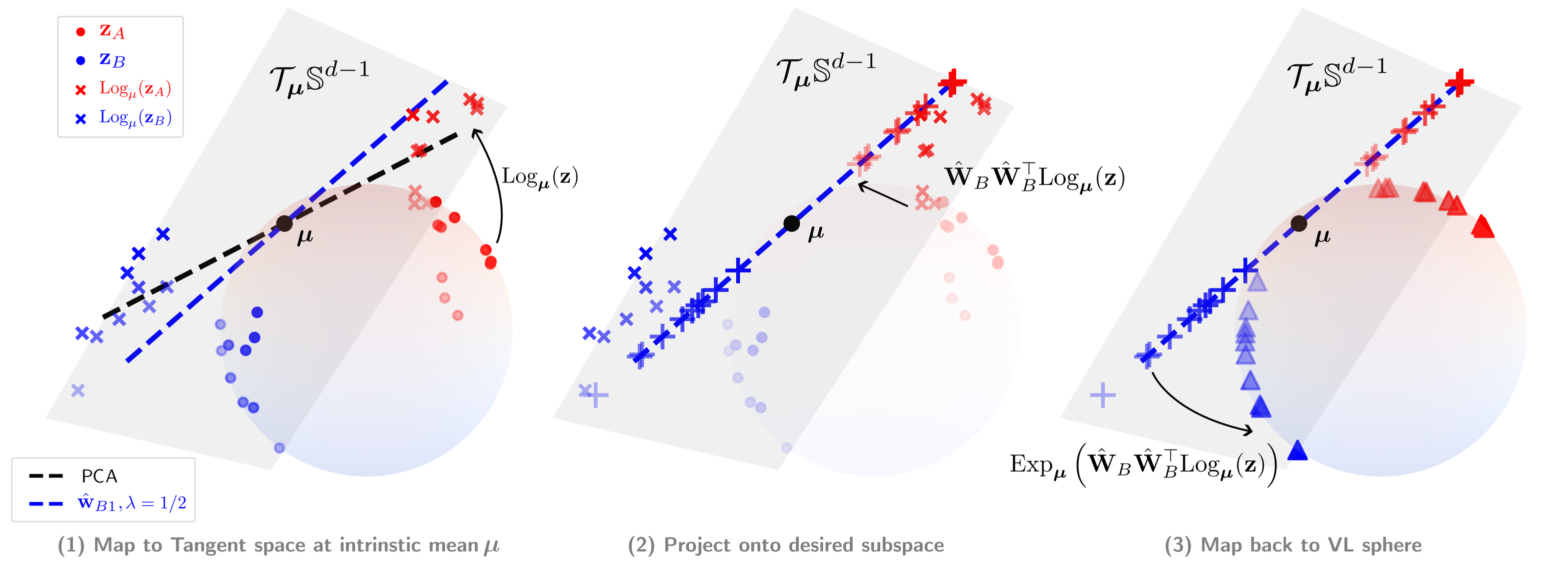
Method

We learn PoS-specific subspaces in the joint vision-language space:

$$\mathbf{W}_i = \underset{\mathbf{W}_i^\top \mathbf{W}_i = \mathbf{I}_k}{\arg\max} \Big\{ (1-\lambda)||\mathbf{W}_i^\top \mathbf{X}_i||_F^2 - \sum_{j \in \mathcal{C}\setminus\{i\}} \lambda ||\mathbf{W}_i^\top \mathbf{X}_j||_F^2 \Big\},$$

where $\mathbf{X}_i \in \mathbb{R}^{d \times n}$ contain in their columns n CLIP embeddings of examples of PoS i (solution is given in closed-form).

Isolating/removing representations visually associated with PoS i is then achieved by projecting onto the subspaces/their orthogonal complements, respectively.



Geometry-aware subspaces are learnt *in the tangent space* to CLIP's VL hypersphere's intrinsic mean–better respecting the geometry of the manifold on which the representations lie [1].

Visual disentanglement

We visualise the ability of subspace orth. complement projection $(\mathbf{I}_d - \mathbf{W}_i \mathbf{W}_i^\top)\mathbf{x}$ to remove visual variation from the CLIP embeddings associated with a specific PoS with LAION's CLIP-based T2IM Paella [3]:

Style-blocking projections

By removing style-based variation from the CLIP representations, the projection onto the orth. complement of the adjective subspace provides a way to block the imitation of artists' styles:

References

[1] P.T. Fletcher et al. "Principal geodesic analysis for the study of nonlinear statistics of shape". In: *IEEE Trans. Med. Imag.* 23.8 (2004), pp. 995–1005.

[2] Sachit Menon et al. "Task Bias in Vision-Language Models". In: *ArXiv* (2022).

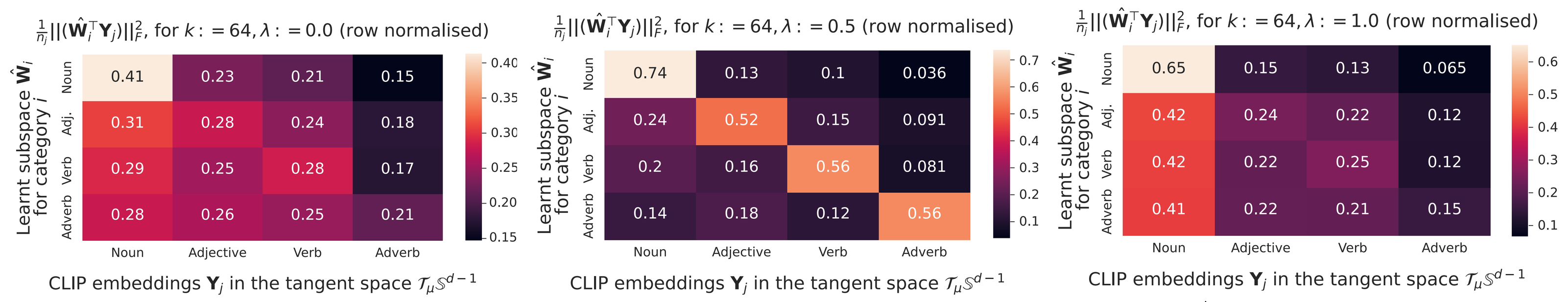
[3] Dominic Rampas et al. "Fast Text-Conditional Discrete Denoising on Vector-Quantized Latent Spaces". In: *ArXiv* (2022).

Custom theme subspaces

The main objective also allows one to learn subspaces corresponding to more specific visual themes (supervised with a dictionary of custom phrases), and thus provides a more targeted way to remove specific visual concepts (e.g. 'gore'):

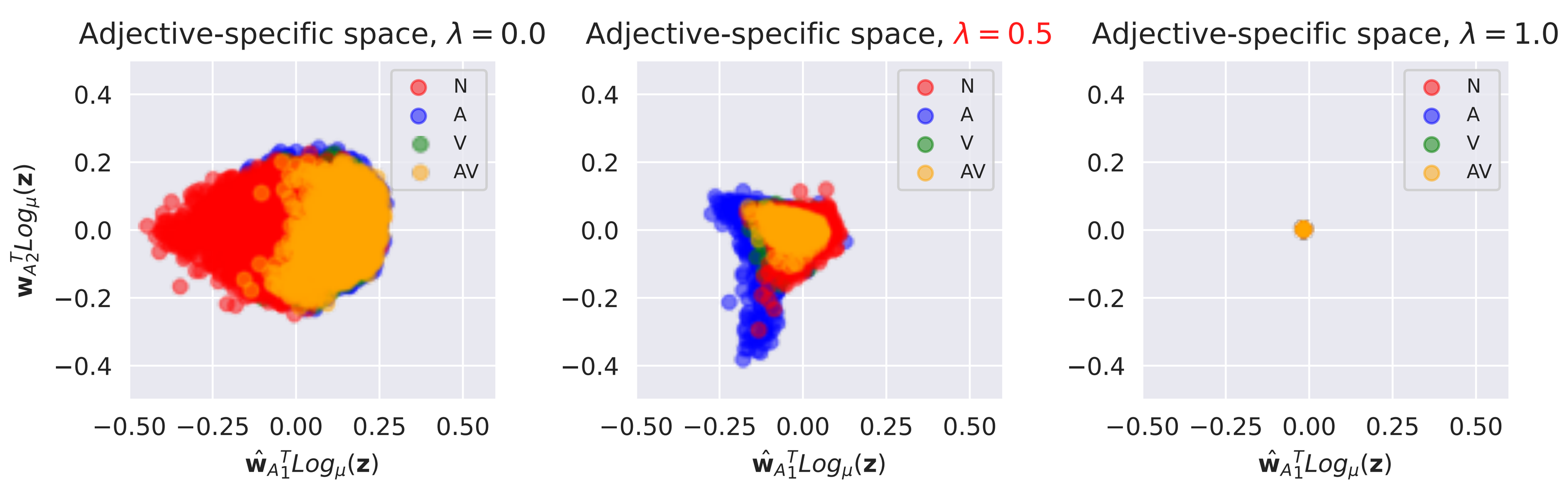
Quantitative & ablations

The quantity $\frac{1}{n_j}||\mathbf{W}_i^\top \mathbf{X}_j||_F^2$ measures the presence of PoS j's data in PoS i's subspace:



(a) PGA for PoS of interest. (b) Ours ($\lambda := 0.5$). (c) PGA$^\perp$ for nuisance PoS.

A choice of $\lambda := 0.5$ provides a reasonable balance between maximising the variance for the target PoS and killing that of the remaining:



SCAN ME for project page