Parts of Speech–Grounded Subspaces in Vision-Language Models:

Supplementary Material

Anonymous Author(s)
Affiliation
Address
email

Contents

A Definitions and derivations **2**
A.1 Closed-form solution 2
A.2 The Logarithmic and Exponential Maps 2

B Additional results **2**
B.1 Qualitative results 2
B.2 Quantitative results 5

C Ablation studies **7**
C.1 Relationship to alternative component analyses 7
C.2 Subspaces vs Submanifolds 8
C.3 Role of k 8
C.4 Role of λ 9

D Experimental details **12**

Submitted to 37th Conference on Neural Information Processing Systems (NeurIPS 2023). Do not distribute.

A Definitions and derivations

A.1 Closed-form solution

Detailed steps for the expansion of the original objective into the trace maximisation form of the main objective are given as follows:

$$
\begin{aligned}
&(1-\lambda)||\mathbf{W}_i^\top \mathbf{X}_i||_F^2 - \sum_{j\in\mathcal{C}\setminus\{i\}} \lambda||\mathbf{W}_i^\top \mathbf{X}_j||_F^2 && (1)\\
=&(1-\lambda)\text{tr}\left((\mathbf{W}_i^\top \mathbf{X}_i)^\top(\mathbf{W}_i^\top \mathbf{X}_i)\right) - \sum_{j\in\mathcal{C}\setminus\{i\}} \lambda\text{tr}\left((\mathbf{W}_i^\top \mathbf{X}_j)^\top(\mathbf{W}_i^\top \mathbf{X}_j)\right) && (2)\\
=&\text{tr}\big((1-\lambda)\mathbf{X}_i^\top \mathbf{W}_i\mathbf{W}_i^\top \mathbf{X}_i - \sum_{j\in\mathcal{C}\setminus\{i\}} \lambda\mathbf{X}_j^\top \mathbf{W}_i\mathbf{W}_i^\top \mathbf{X}_j\big) && (3)\\
=&\text{tr}\big((1-\lambda)\mathbf{W}_i^\top \mathbf{X}_i\mathbf{X}_i^\top \mathbf{W}_i - \sum_{j\in\mathcal{C}\setminus\{i\}} \lambda\mathbf{W}_i^\top \mathbf{X}_j\mathbf{X}_j^\top \mathbf{W}_i\big) && (4)\\
=&\text{tr}\big(\mathbf{W}_i^\top((1-\lambda)\mathbf{X}_i\mathbf{X}_i^\top - \sum_{j\in\mathcal{C}\setminus\{i\}} \lambda\mathbf{X}_j\mathbf{X}_j^\top)\mathbf{W}_i\big) && (5)\\
=&\text{tr}\left(\mathbf{W}_i^\top \mathbf{C}_i\mathbf{W}_i\right), && (6)
\end{aligned}
$$

where $\mathbf{C}_i = \left((1-\lambda)\mathbf{X}_i\mathbf{X}_i^\top - \sum_{j\in\mathcal{C}\setminus\{i\}} \lambda\mathbf{X}_j\mathbf{X}_j^\top\right)$. Here we have used $||\mathbf{X}||_F^2 = \text{tr}\left(\mathbf{X}^\top\mathbf{X}\right)$ and the linearity and cyclic properties of the trace.

A.2 The Logarithmic and Exponential Maps

For mapping to the hypersphere's tangent space at a reference point in the main paper, we use well-known explicit formulas. Concretely, the *Logarithmic Map* $\text{Log}_{\mathbf{p}} : \mathbb{S}^{d-1} \rightarrow \mathcal{T}_{\mathbf{p}}\mathbb{S}^{d-1}$, which maps points on the sphere to the tangent space at a reference point $\mathbf{p} \in \mathbb{S}^{d-1}$ is defined as

$$
\text{Log}_{\mathbf{p}}(\mathbf{z}) = \arccos(\mathbf{z}^\top\mathbf{p})\frac{(\mathbf{I}_d - \mathbf{p}\mathbf{p}^\top)(\mathbf{z}-\mathbf{p})}{||(\mathbf{I}_d - \mathbf{p}\mathbf{p}^\top)(\mathbf{z}-\mathbf{p})||_2}. \tag{7}
$$

Its inverse, the *Exponential Map* $\text{Exp}_{\mathbf{p}} : \mathcal{T}_{\mathbf{p}}\mathbb{S}^{d-1} \rightarrow \mathbb{S}^{d-1}$ mapping points back onto the sphere is given by

$$
\text{Exp}_{\mathbf{p}}(\mathbf{z}) = \cos\left(||\mathbf{z}||_2\right)\mathbf{p} + \sin\left(||\mathbf{z}||_2\right)\frac{\mathbf{z}}{||\mathbf{z}||_2}. \tag{8}
$$

B Additional results

B.1 Qualitative results

Visual disentanglement Firstly, we include many more examples of the visual disentanglement with the 'adjective' and 'noun' PoS subspaces and TTIM of Rampas et al. [1] in Figures 1 and 2.

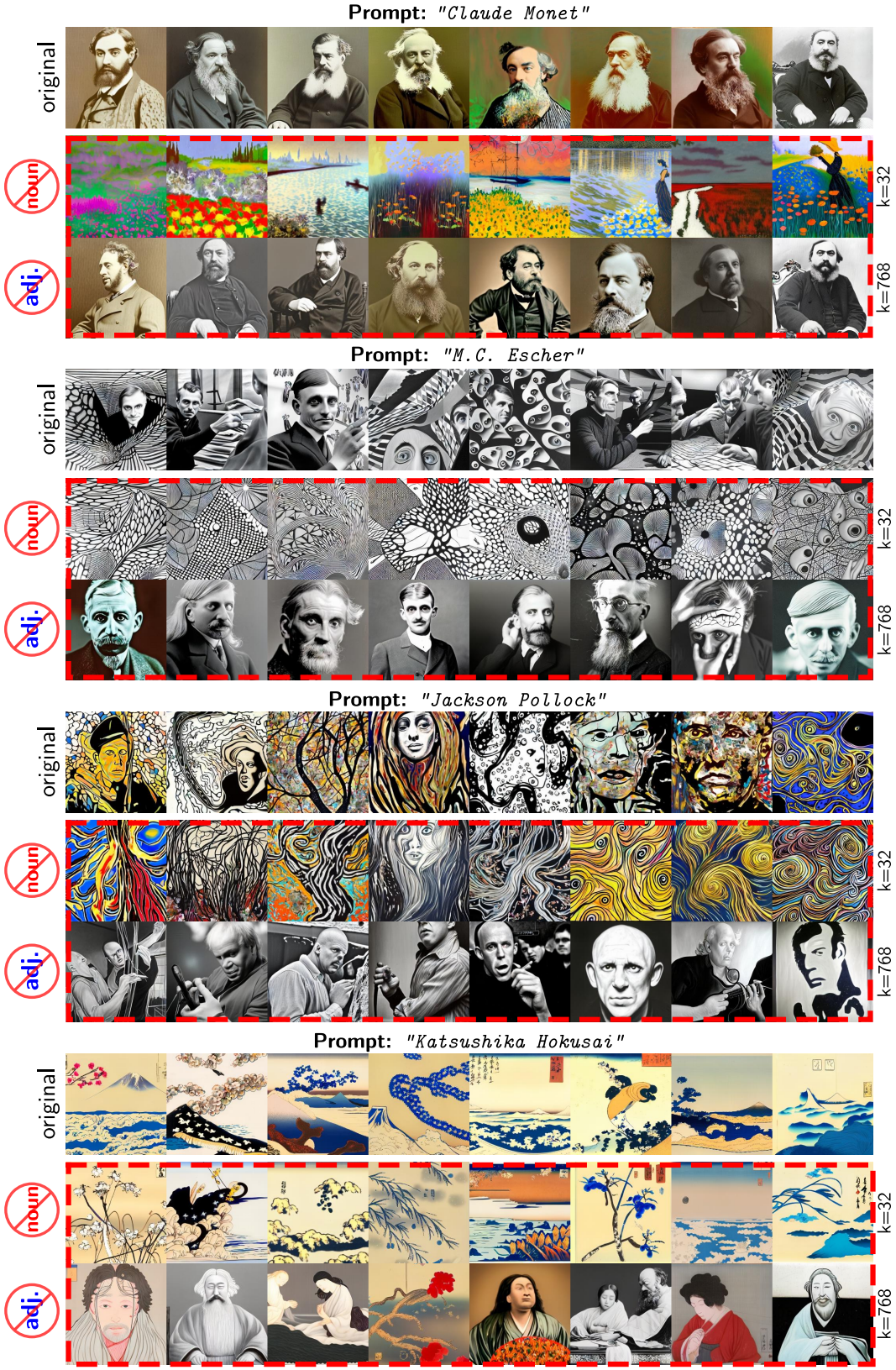


Figure 1: Additional results for visual disentanglement of text prompts using the learnt PoS subspaces.

Figure 2: Additional results for visual disentanglement of text prompts using the learnt PoS subspaces.

Visual theme removal As shown in the main paper, the adjective subspace works remarkably well for preventing the imitation of artists' styles in the CLIP-based text-to-image model Paella. Some additional examples projecting text prompt's CLIP embedding onto the orthogonal complements of the adjective subspace with $\Pi_A^{\perp}(\mathbf{z}_T)$ are shown in Figure 3.

Figure 3: Additional results blocking the imitation of artists using only the adjective subspace orthogonal complement projection $\Pi_A^{\perp}(\mathbf{z}_T)$.

However, as stated in the main paper, the adjective subspace orthogonal projection for the task of 'style blocking' is overly restrictive in also preventing the description of visual appearance with adjectives. We provide in Figure 4 some examples of this–for example, the 'stormy' and 'red' visual appearances are removed in Figure 4 after projection. On the other hand, the custom visual theme subspaces can target *specific* visual appearances more precisely–two examples are shown in Figure 5 (following the experimental setup outlined in Appendix D).

Figure 4: 'Style blocking' with the adjective subspace is overly restrictive (here blocking 'stormy' and 'red' visual appearances).

B.2 Quantitative results

Visual theme subspace invariance Here we show quantitative results for the visual theme-specific subspaces. We have evaluated this visually via text-to-image models in Figure 5, however here we wish to demonstrate that large variation in the CLIP representations directly is captured for only the themes of interest. Concretely, in Figure 6 we compute the same class invariance metric for 200

Figure 5: Additional results for the two custom visual theme orthogonal complement subspace projection for 'gore' and artists' styles.

random WordNET words and random theme-specific words in the learnt custom theme subspaces through $\frac{1}{200}||(\hat{\mathbf{W}}_i^\top \mathbf{Y})||_F^2$. Here, $\mathbf{Y} \in \mathbb{R}^{d \times 200}$ contains in its columns the CLIP word embedding mapped to the tangent space, and i denotes the specific custom visual theme of interest for a particular subfigure. As can be seen from the high magnitude in the orange bars and low magnitude of the blue bars in Figure 6, these subspaces map the 200 random other words much closer to the zero vector than the theme-specific words. This indicates the variance in just the words of interest has indeed been captured.

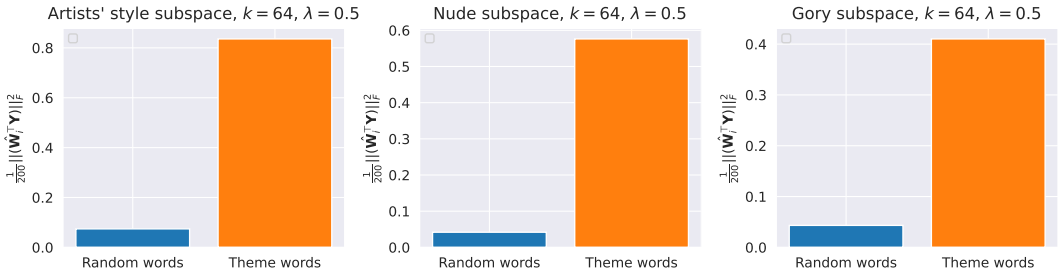


Figure 6: $\frac{1}{200}||(\hat{\mathbf{W}}_i^\top \mathbf{Y})||_F^2$ of CLIP representations (of both 200 random words and 200 words from the 'training' set describing the theme of interest) projected onto theme i-specific subspaces.

Additional zero-shot classifcation We show in both Table 1 and Table 2 additional results for the baseline zero-shot classification protocol (following the exact same setup in the main paper) with the similarity metric $S(\Pi_N(\mathbf{z}_I), \mathbf{z}_T)$ after the noun subspace projection. As can be seen, the proposed subspace leads to improved zero-shot classification on a wide range of datasets, for multiple CLIP architectures.

Table 1: Top-1 ZS classification accuracy with CLIP `ViT-B-16`.

	ImageNET	MIT-states	UT Zap.	DomainNET	StanfordCars	Caltech101	Food101	CIFAR10	CIFAR100	OxfordPets	Flowers102	Caltech256	STL10	MNIST	FER2013
CLIP	61.60	47.80	89.10	54.40	62.50	82.00	81.50	88.80	62.40	76.10	65.20	81.70	96.90	48.00	41.90
PoS PCA	61.60	47.90	89.00	54.40	62.50	82.10	81.50	88.80	62.30	76.00	65.30	81.80	**97.00**	47.70	41.70
PoS PGA	61.80	47.70	**89.60**	54.30	60.00	82.20	81.70	87.90	62.70	**78.60**	64.60	82.30	96.60	50.00	42.70
Noun Submanifold	**62.80**	**48.30**	88.10	**54.70**	**62.70**	**82.90**	**82.10**	**89.10**	**63.90**	77.00	**65.40**	**83.40**	**97.00**	**55.00**	**48.60**

Table 2: Top-1 ZS classification accuracy with CLIP `ViT-L-14`.

	ImageNET	MIT-states	UT Zap.	DomainNET	StanfordCars	Caltech101	Food101	CIFAR10	CIFAR100	OxfordPets	Flowers102	Caltech256	STL10	MNIST	FER2013
CLIP	69.00	51.80	90.40	59.60	74.70	80.90	87.40	**91.70**	72.70	82.50	74.60	86.80	**97.70**	**59.70**	35.70
PoS PCA	69.00	51.80	90.40	59.60	74.70	80.90	87.40	**91.70**	72.70	82.50	74.60	86.80	**97.70**	**59.70**	35.70
PoS PGA	69.10	51.90	**90.50**	59.60	74.60	80.80	87.50	91.60	72.80	82.60	**74.80**	86.80	**97.70**	**59.70**	35.80
Noun Submanifold	**70.10**	**52.50**	90.10	**60.00**	**76.00**	**82.80**	**87.70**	90.60	**73.90**	**83.10**	74.70	**88.10**	96.90	58.90	**46.40**

C Ablation studies

C.1 Relationship to alternative component analyses

We first compare 1D subspaces learnt with our method, FDA [2], and FKT [3] shown in Figure 7 on toy data chosen to illustrate the qualitative differences in the properties of the subspaces. Given the very different goals of FDA in minimising intra-class variation, the resulting FDA subspace is near-orthogonal to that of FKT and the proposed method. Whilst the FKT-given subspace is very close to ours, for this particular illustrative toy data our subspace better kills the variance in the red data (as illustrated in Figure 7 b.iii). This figure also provides a visualisation of the learnt subspace's proximity to the principal component of the target class and the bottom principal component of the red class' datapoints–the proximity to the two extremes being controlled by λ.

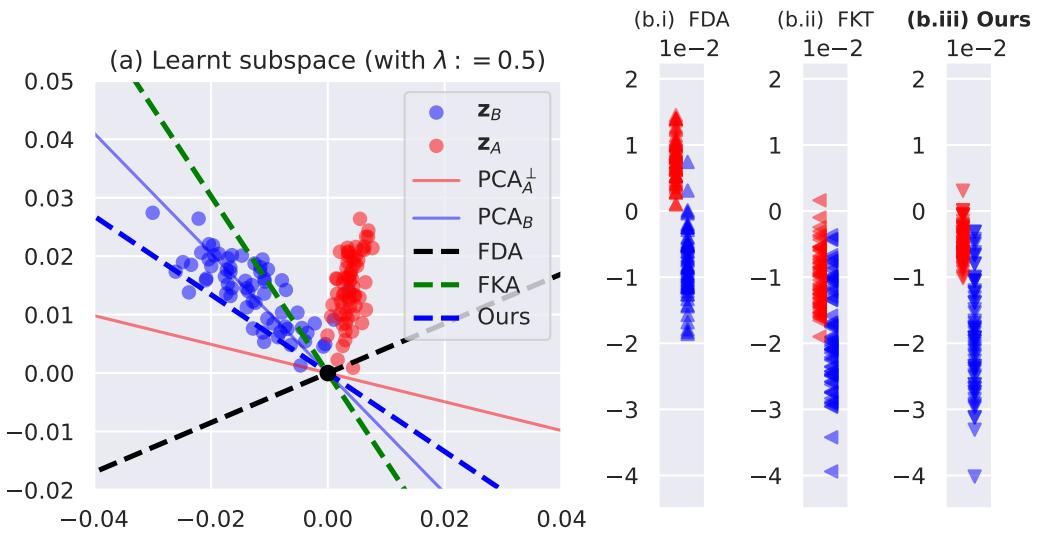


Figure 7: (a) A visual comparison of the leading eigenvector of $\mathbf{C}_i$ to the first FDA component (centred for comparison), to the first FKT component. Shown in (b) are the points' coordinates in the three subspaces. As can be seen, the learnt $\mathbf{w}_{B1}$ captures large variance in the blue target class and is close to orthogonal to data points of the red class.

C.2 Subspaces vs Submanifolds

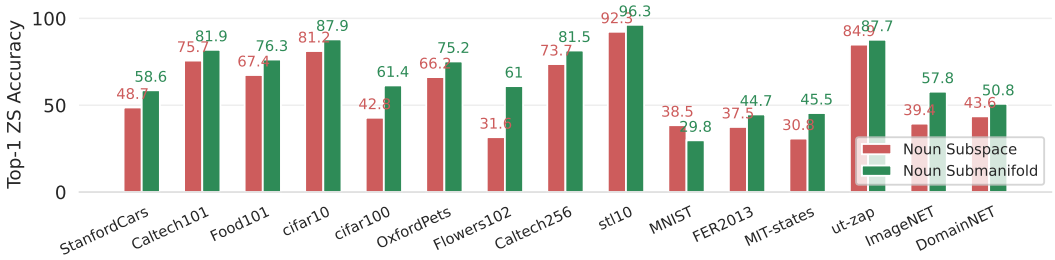


Figure 8: Ablation study on the zero-shot Top-1 performance comparing the subspaces to the submanifolds on CLIP `ViT-B-32`; the submanifolds perform almost stricly better than the subspaces.

Here, we show the benefit of using the geometry-aware formulation (learning subspaces in the *tangent space* to the CLIP hypersphere's intrinsic mean) over the subspaces of the ambient Euclidean space of $\mathbb{R}^d$. We show in Figure 8 the zero-shot classification accuracies on all datasets considered in the main paper, first projecting the images onto both the regular Euclidean subspaces and instead the submanifolds which better respect the geometry of the sphere. As can be seen, the geometry-aware subspaces almost strictly outperform the regular subspaces, on almost all datasets.

C.3 Role of k

We show the impact of various choices of k (dimensionality of the subspaces) as visualised with text-to-image models in Figure 9 with the projections onto the orthogonal complements to kill the undesired variation. As can be seen, increasing k removes more and more visual information relevant to the particular visual mode. For example, we see the image feature increasingly less snow on the right, whilst increasing less of London on the left.

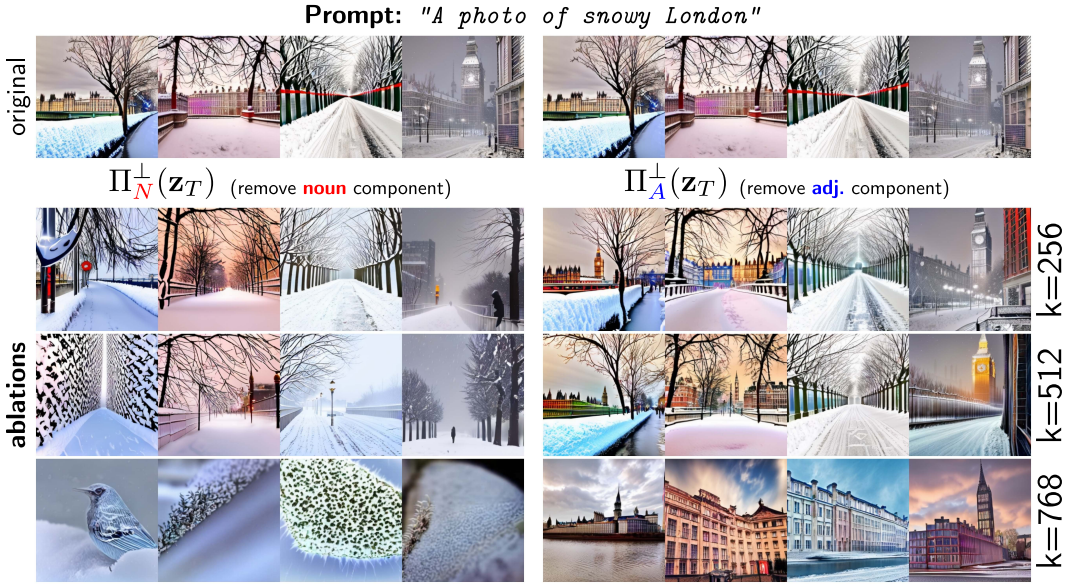


Figure 9: Ablation study on the value of k on the $d = 1024$ OpenCLIP VL space when projecting onto the orthogonal complements of k-dimensional adjective and noun subspaces.

C.4 Role of λ

We next provide an ablation study on the value of λ. We provide both a visual qualitative ablation and a quantitative one.

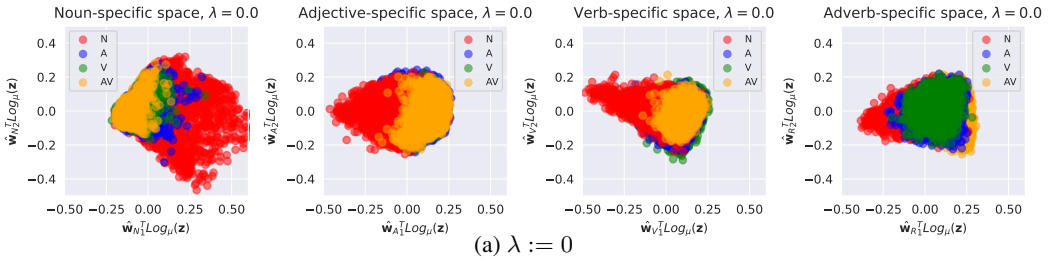
Qualitative Concretely, in each subfigure (row) of Figure 11, we take the first 5000 WordNET text strings from each part of speech and compute their CLIP text embeddings $\mathbf{z}_T \in \mathbb{R}^d$. We then calculate $\hat{\mathbf{W}}_i^\top \text{Log}_{\boldsymbol{\mu}}(\mathbf{z}_T)$, plotting the first coordinate along the x-axis and the second along the y-axis of each subfigure in Figure 11. Ideally, the data points in the target class should be the only ones with a large norm if this hyperplane captures visual variation that is unique to a particular word class. We see that $\lambda := 0$ preserves the most variance in the target class' embeddings but the different categories' projections are clearly entangled–the other classes' datapoints also have large norm. Conversely, $\lambda := 1.0$ maps all points effectively to the zero vector–killing the variance in all categories. As can be seen in Figure 11c, $\lambda := 0.5$ offers a reasonable balance of both properties in this 4-class setting. The exact same experiments are run on the larger version of CLIP, shown in Figure 12, where similar conclusions can be drawn about the practical impact of λ.

Quantitative For quantifying this in more dimensions, we compute the class invariance metric (used in the main paper) in Figure 13 and Figure 14 for various values of λ, where we observe that $\lambda := 0.5$ is a sensible choice for multiple CLIP architectures.

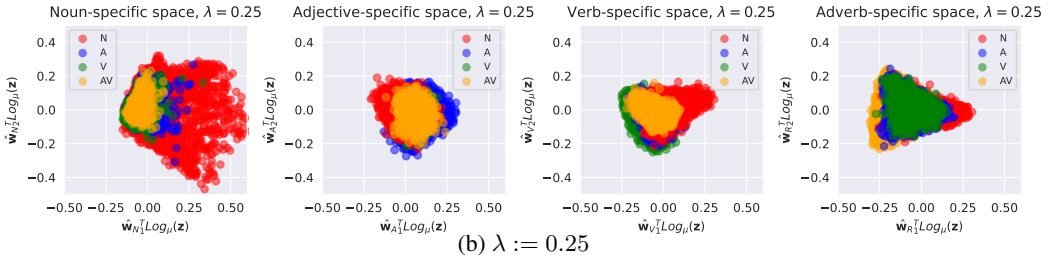
Visual subspaces Finally, we demonstrate the importance of using PoS as 'negative examples' in the summation in the main objective when learning visual theme-specific subspaces. Intuitively, whilst we want to maximise the variation for phrases of a particular theme (such as 'gory'), we also want to preserve the ability to generate other concepts with the TIIM, which is what the objective provides through the hyperparameter λ.

In particular, we show in the second row of Figure 10 the visual results when projecting onto the orthogonal complement of a 'gory' subspace learnt when we do *not* use the PoS as 'negative guidance' (i.e. when $\lambda := 0$). As can be seen in comparison to the third row of Figure 10, using the PoS is critical in this instance for retaining the ability to synthesise existing related concepts (here ensuring e.g. 'cranberry juice' can still be synthesised even though visually 'gory' appearances are removed).

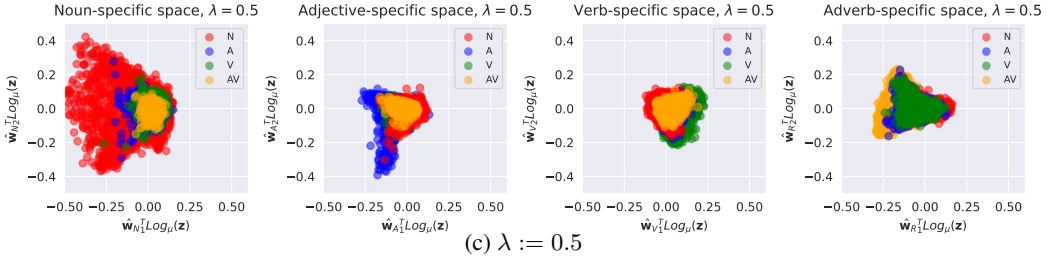
Figure 10: Without using PoS as 'negative guidance' (i.e. when $\lambda := 0$), related concepts (e.g. 'cranberry juice') can be visually removed to a much greater extent than when using the PoS guidance ($\lambda := 0.5$).



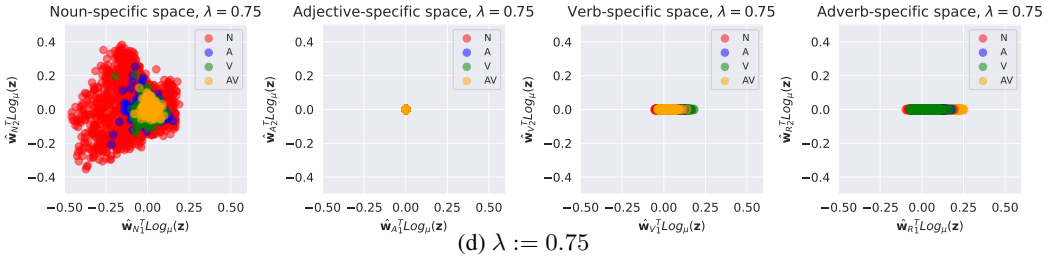
(a) $\lambda := 0$



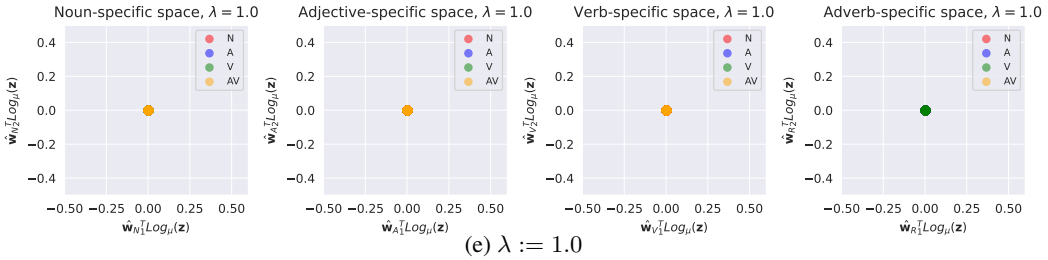
(b) $\lambda := 0.25$



(c) $\lambda := 0.5$

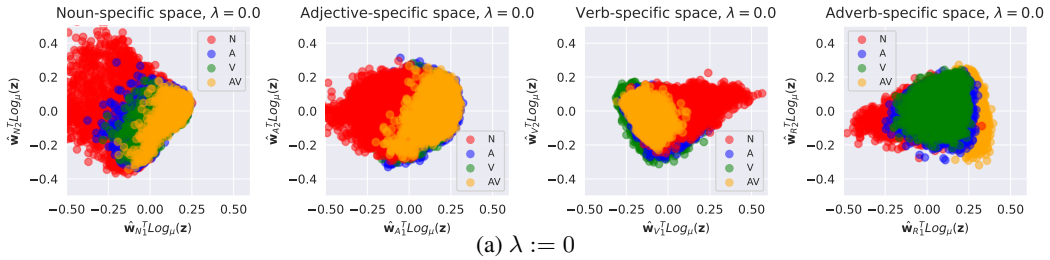


(d) $\lambda := 0.75$

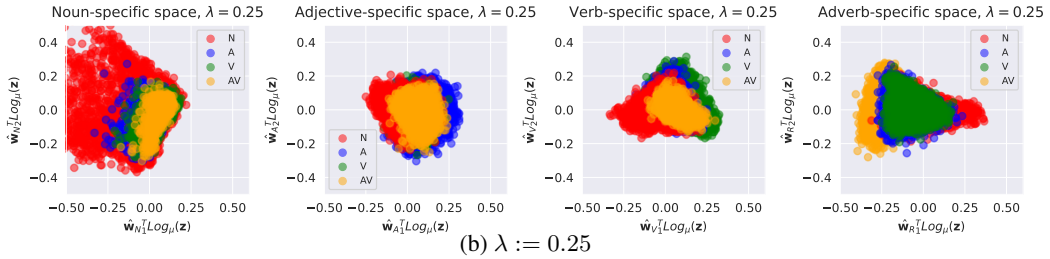


(e) $\lambda := 1.0$

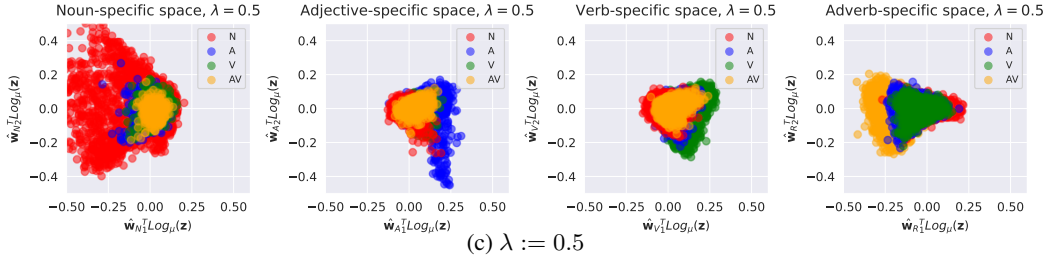
Figure 11: Embeddings' first two coordinates in the tangent space(es), with various values of λ in the main objective (axis limits are fixed to compare length of vectors across values of λ). The base CLIP model `clip-vit-base-patch32` is used here.



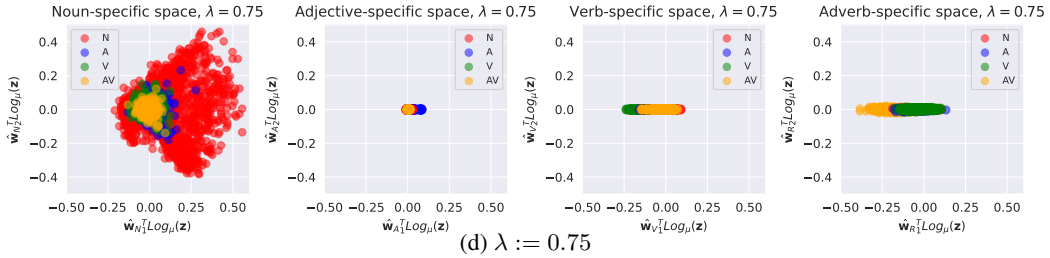
(a) $\lambda := 0$



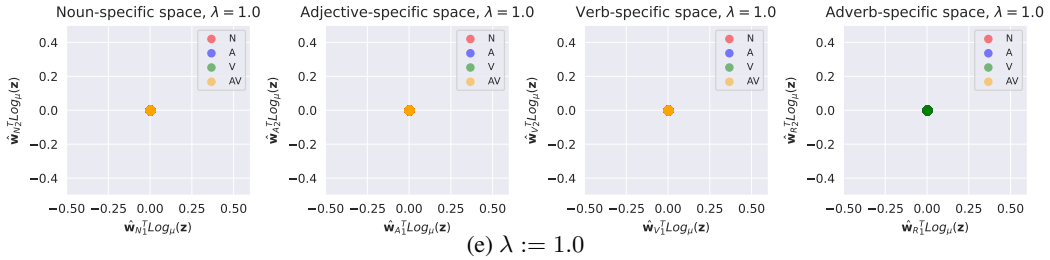
(b) $\lambda := 0.25$



(c) $\lambda := 0.5$

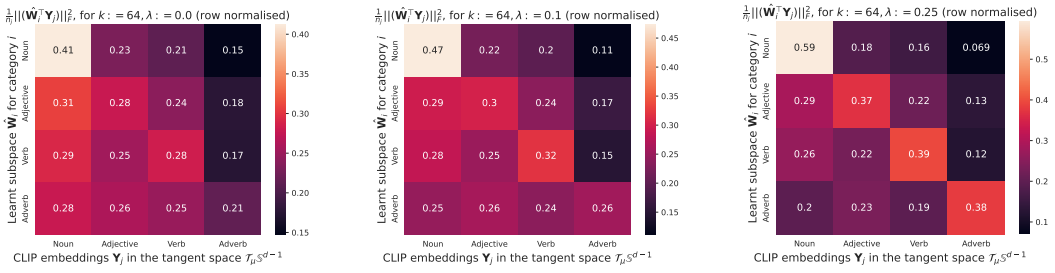


(d) $\lambda := 0.75$

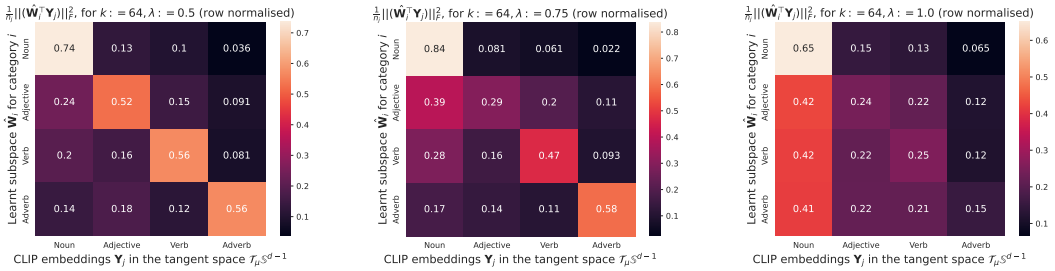


(e) $\lambda := 1.0$

Figure 12: Embeddings' first two coordinates in the tangent space(es), with various values of λ in the main objective (axis limits are fixed to compare length of vectors across values of λ). The larger CLIP model `clip-vit-large-patch14` is used here.

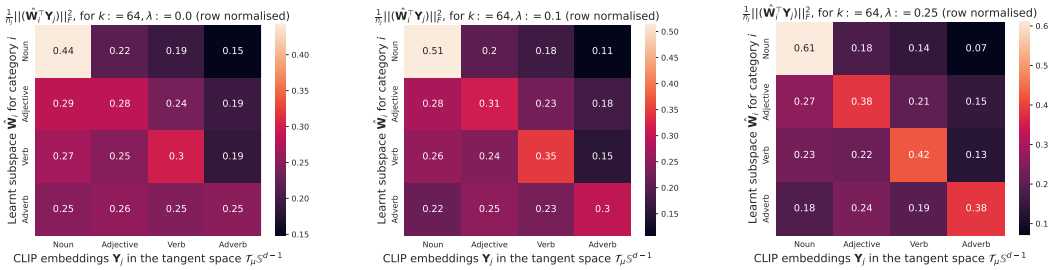


(a) $\lambda := 0$ (b) $\lambda := 0.1$ (c) $\lambda := 0.25$

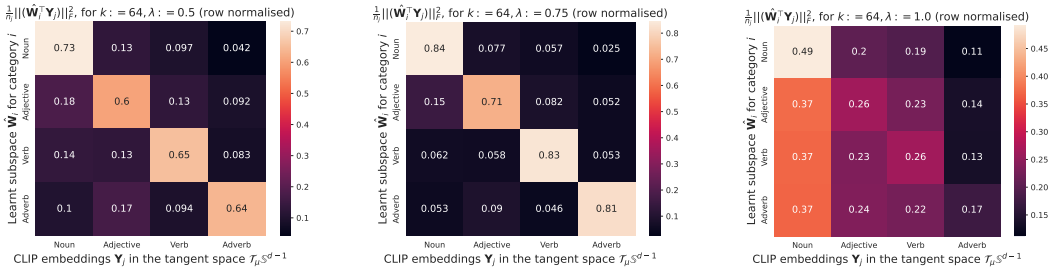


(d) $\lambda := 0.5$ (e) $\lambda := 0.75$ (f) $\lambda := 1.0$

Figure 13: Ablation on λ with the quantity $\frac{1}{n_j}||(\hat{\mathbf{W}}_i^\top \mathbf{Y}_j)||_F^2$ introduced in the main paper. Row-normalisation is performed to highlight the relative representation of each class' embeddings within each subspace. The base CLIP model `clip-vit-base-patch32` is used here.



(a) $\lambda := 0$ (b) $\lambda := 0.1$ (c) $\lambda := 0.25$



(d) $\lambda := 0.5$ (e) $\lambda := 0.75$ (f) $\lambda := 1.0$

Figure 14: Ablation on λ with the quantity $\frac{1}{n_j}||(\hat{\mathbf{W}}_i^\top \mathbf{Y}_j)||_F^2$ introduced in the main paper. Row-normalisation is performed to highlight the relative representation of each class' embeddings within each subspace. The larger CLIP model `clip-vit-large-patch14` is used here.

D Experimental details

For reference in this section, we first include the dimensionalities of the shared VL spaces and public links to implementations used of the three CLIP models in this paper in Table 3.

Table 3: Dimensionality of the VL space representations in the four CLIP models.

Model name	`clip-vit-base-patch32`	`clip-vit-base-patch16`	`clip-vit-large-patch14`	'OpenCLIP'
Dimensionality of $\mathbf{z}_I$, $\mathbf{z}_T$	512	512	768	1024
Public link	HuggingFace	HuggingFace	HuggingFace	Github

Visual disentanglement The Paella TTIM [1] used in the main paper adopts the 'OpenCLIP' [4] model with a larger $d = 1024$-dimensional VL representation. For all 'visual disentanglement' results throughout both the paper and supplementary material, we use $k = 768$ dimensional 'adjective' and 'noun' subspaces for all text prompts (apart from when removing the 'content' representations in visually polysemous phrases, where we find only $k = 32$ components are necessary).

Visual theme subspaces For the custom visual subspaces, we produce a list of phrases related to the visual theme of interest by asking ChatGPT [5] questions of the format: `Please give me a list of 250 words and phrases related to the concept of {x}`, where `x` is the visual concept of interest (such as 'gore'). For the case of the artist subspace, we ask: `Please give me a list of 250 of the most famous painters and visual artists of all time.` In each scenario, we follow up twice more asking for additional responses (given the limited response length), specifying that it tries not to repeat any of the previous answers in the list. For the experiments in the 'gory' and 'artist' custom subspaces, ChatGPT gave us 371 and 830 unique phrases respectively (taking just the provided artists' surnames as additional examples for the latter), and use a $k = 128$- and $k = 512$-dimensional subspace respectively (given the limited number of phrases for the gore subspace provided by ChatGPT).

Concurrent work Recent preprints [6, 7] explicitly address the task of ablating particular concepts in diffusion models specifically. However, in contrast to the proposed method, these preprints fine-tune Stable Diffusion–specific [8] submodules, and do not focus on the final CLIP vector representations. Thus, there is no straightforward way to compare the proposed method working in CLIP's shared vision-language space directly nor the alternative Paella [1] TIIM. One methodological benefit to [6] over the proposed method however (purely in the context of text-to-image synthesis) is in the requirement of only a single text prompt describing a concept, relative to our necessary collection[1]. On the other hand, our subspaces are learnt in closed form–for example, the 'gory' subspace takes only 0.28 seconds to compute on a V100 GPU, given the CLIP embeddings. This is in contrast to [6]'s models which are stated to require 1000 gradient descent steps to compute, and [7] taking 5 minutes per concept.

Compute time and harware To run the Paella model, we use a 32GB NVIDIA Tesla V100 GPU. Learning the subspaces is particularly fast given the closed-form solution, taking just 1.1 seconds to compute all 4 (noun, adjective, verb, and adverb) PoS subspaces. Encoding all WordNet PoS examples with CLIP takes 28.91 minutes, however, this is a fixed cost and only needs to be done once at the beginning (after which any number of additional subspaces can be computed very quickly).

References

[1] Dominic Rampas, Pablo Pernias, Elea Zhong, and Marc Aubreville. Fast text-conditional discrete denoising on vector-quantized latent spaces, 2022.

[2] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7 (2):179–188, 1936.

[3] Keinosuke Fukunaga. *Introduction to statistical pattern recognition*. Elsevier, 2013.

[4] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning, 2022.

[5] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *NeurIPS*, 35:27730–27744, 2022.

[1] In particular, n phrases' embeddings can span a subspace with a maximum of n dimensions

[6] Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. *arXiv preprint arXiv:2303.07345*, 2023.

[7] Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. *arXiv preprint arXiv:2303.13516*, 2023.

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Bjorn Ommer. High-resolution image synthesis with latent diffusion models. *CVPR*, Jun 2022.