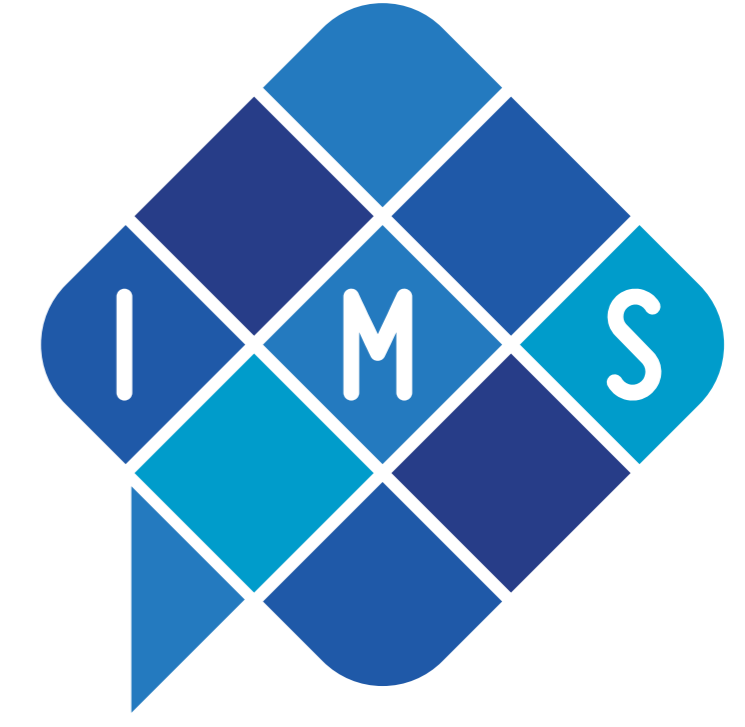


Evaluating Textual and Visual Semantic Neighborhoods of Abstract and Concrete Concepts



Sven Naber¹ Diego Frassinelli² Sabine Schulte im Walde¹
¹IMS, University of Stuttgart ²MaiNLP, LMU Munich

Background and Experimental Setup

Conceptual Similarity Across Modalities

- Humans intuitively understand conceptual similarity, e.g. we perceive *cat* as more similar to *dog* than to *table*.
- Different modalities → different semantic spaces.
 - Textual representations: distributional patterns in language.
 - Visual representations: perceptual similarity in images.
- How are these spaces organized, and how do they align across modalities and levels of conceptual concreteness?**

Research Questions

- RQ1:** How aligned are semantic neighborhoods within and across modalities?
- RQ2:** How does concreteness (abstract–concrete) affect alignment?
- RQ3:** What are the effects of neighborhood size and image aggregation?

Experimental Setup

- Candidate Concepts:** 5,448 frequent nouns from Brysbaert concreteness ratings.
- Target Concepts:** Three 500-noun subsets from the candidates.

Level	Range	Concepts
abstract	1.0–2.0	<i>idea, justice</i>
mid-scale	3.0–4.0	<i>story, election</i>
concrete	4.8–5.0	<i>apple, car</i>

- Distributional word representations:** Count-based, GloVe, Word2Vec, FastText.
- Sentence representations:** MpNet, Gemma, Qwen3 (mean of 35 ENCOW16AX sentences).
- Visual representations:** ViT, DINOv2, HierA, CLIP (mean of top 35 Bing images).
- Metric:** Normalized Alignment Score (NAS) based on $n-k$ nearest-neighbor overlap.

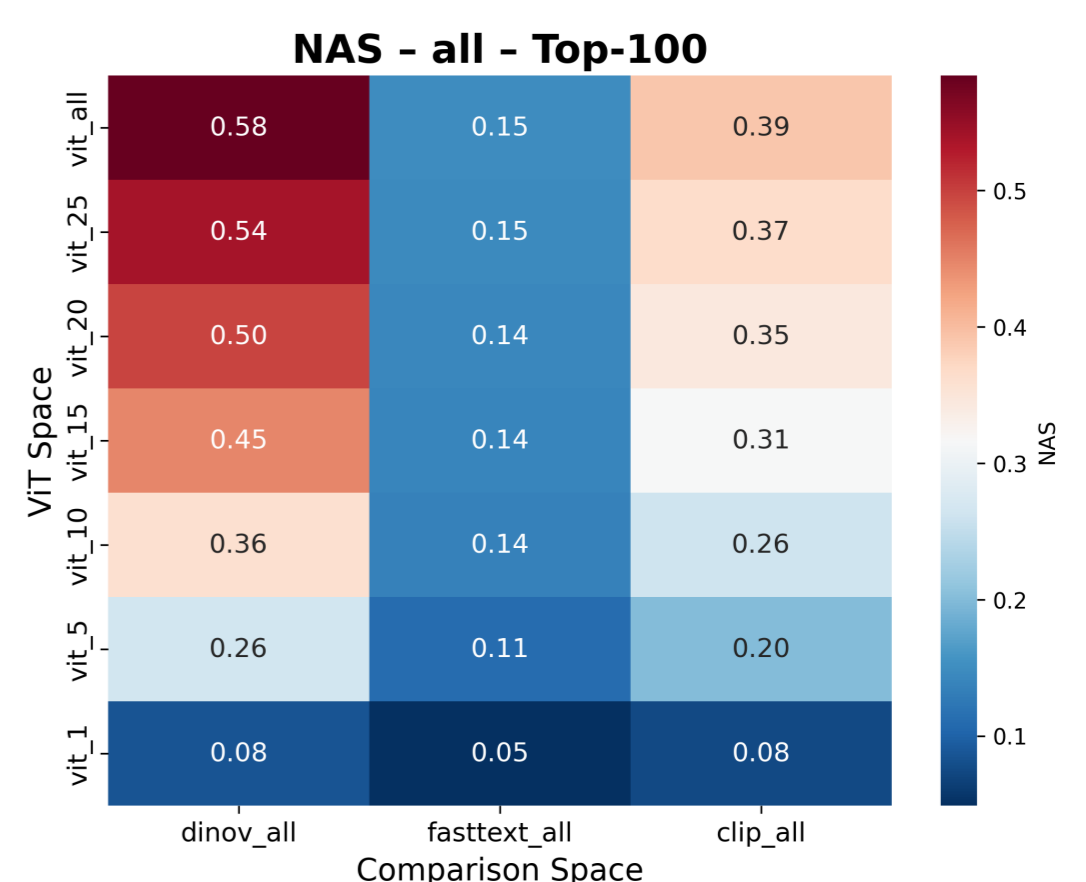
$$O_c^{n,k} = \frac{|N_1(c)[n:k] \cap N_2(c)[n:k]|}{k - n + 1}, \quad O_{\text{obs}} = \frac{1}{|C|} \sum_{c \in C} O_c^{n,k}$$

$$\text{NAS} = \frac{O_{\text{obs}} - O_{\text{rand}}}{1 - O_{\text{rand}}}, \quad O_{\text{rand}} = \frac{k - n + 1}{N - 1}$$

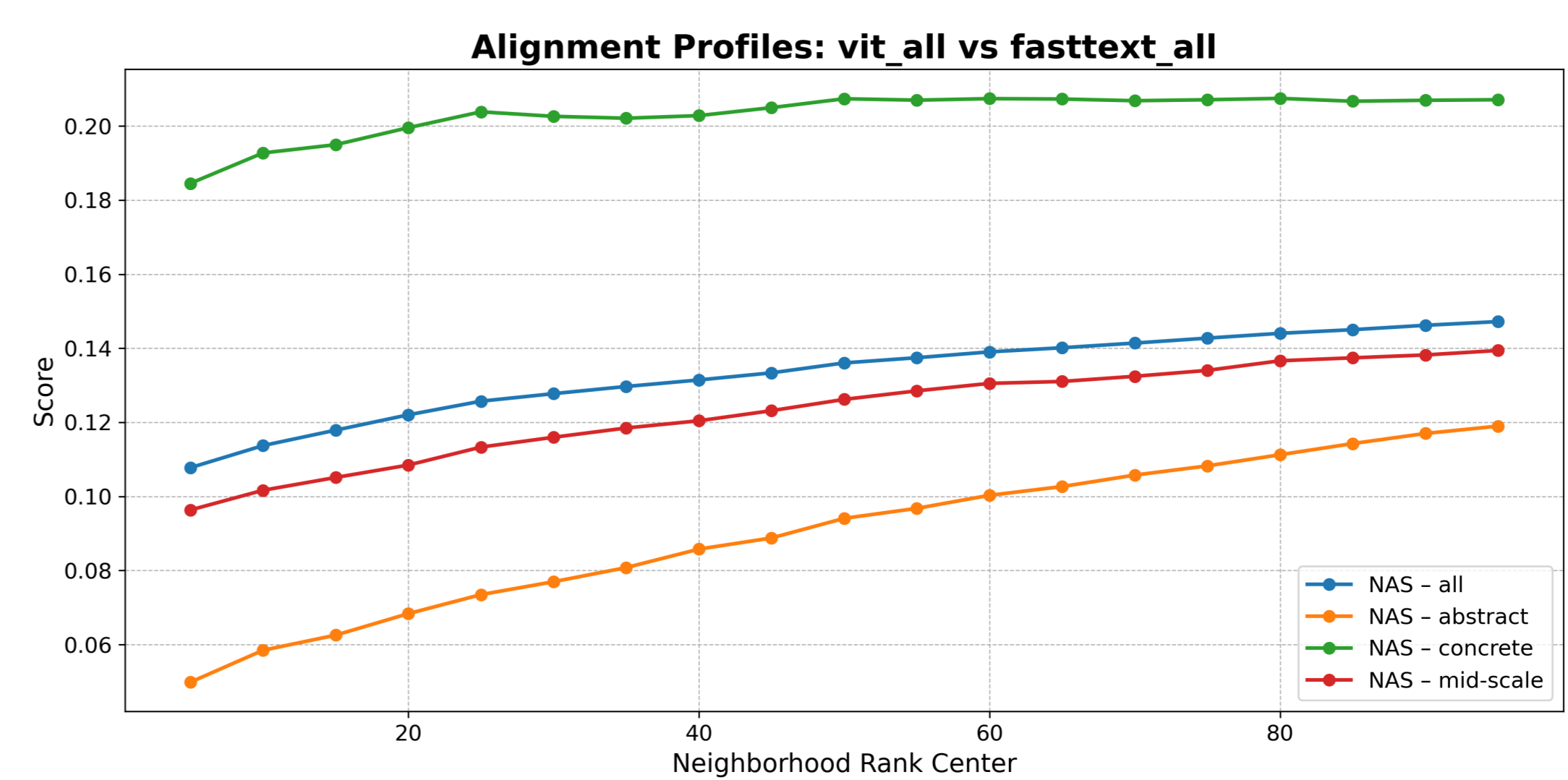
Results and Discussion

Key Findings

- Modality is the dominant factor:** Semantic neighborhoods align more strongly within a modality than across modalities.
- Concreteness enhances alignment:** Concrete concepts exhibit higher and more consistent within- and cross-modal alignment than abstract concepts.
- Image aggregation stabilizes representations:** Mean aggregation across multiple images per concept (best at 25) increases alignment.



- Neighborhood size influences alignment:** Larger k -neighborhoods capture broader semantic similarity and mostly yield higher alignment.



Nearest-Neighbor Examples

- Both vision and text models retrieve plausible neighbors for concrete, mid-scale and most abstract concepts.
- Non-depictable abstract terms such as *ethos* pose a challenge for image models.

Concept	ViT (top-5)	FastText (top-5)
<i>eye</i>	pupil; eyelid; macro; cataract; eyesight	eyelid; nose; eyesight; ear; forehead
<i>goal</i>	competition; greatness; op-timism; determination; confidence	effort; scorer; goalkeeper; striker; ball
<i>probability</i>	interpolation; denominator; differentiation; combination; fraction	likelihood; variance; estimation; prediction; approximation
<i>ethos</i>	competence; outcome; competency; analysis; epidemiology	ethic; attitude; commitment; tradition; individuality

Neighborhood Overlap across Representations

