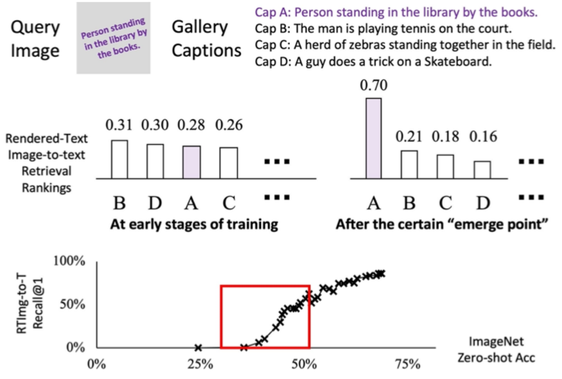
Emergence of Text Readability in Vision Language Models

Jaeyoo Park — SNU
Sanghyuk Chun — Naver AI Lab
Wonjae Kim† — Naver AI Lab (†Now at TwelveLabs)
Sangdoo Yun — Naver AI Lab
Bohyung Han — SNU

ComputerVisionLab, Seoul National University
NAVER AI LAB

CVPR Nashville JUNE 11-15, 2025

TL;DR



VLMs learn to read text suddenly after reaching a certain level of semantic understanding, not gradually.

Experimental Setting

Dataset (3,000 Triplets)

T (Raw Text) — **R** (Rendered-Text Image) — **I** (Natural Image)

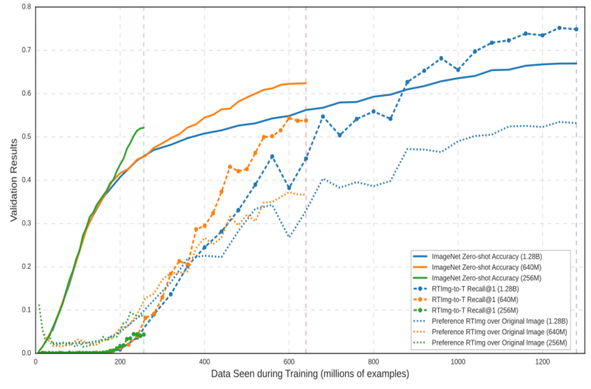
- **T** (Raw Text): Diverse captions (NoCaps).
- **R** (Rendered-Text Image): Generated via FLUX.1 dev (diffusion).
  - Varied parameters: Font, color, background, position.
- **I** (Natural Image): Corresponding original image.

Key Metrics

- Text Recognition (**RTImg-to-T Recall@1**):
  - Ability to match rendered text image (R) with its raw text (T).
- Deeper Semantic Understanding (**RTImg-to-I Recall@1**):
  - Ability to match rendered text image (R) with the corresponding natural image (I), indicating understanding of rendered text's meaning.
- Preference for Rendered Text (**Similarity (R, T) > Similarity (I, T)**)
- General Semantic Understanding (**ImageNet Zero-shot Accuracy**)

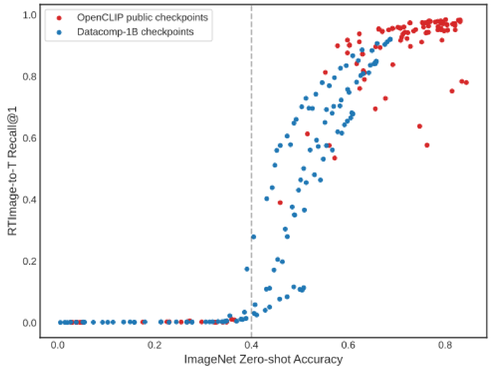
When and How Does Text Readability Emerge?

Robust Emergence Across Training Scales



- **Training Setting**
  - **ViT-B/16**
  - **Datacomp-1B (256M, 640M, 1.28B)**
- **Main Point** : The **abrupt emergence** of text readability (**RTImg-to-T**) is remarkably consistent across different training data scales.
- **Observations** :
  - For all scales, **RTImg-to-T** (*dashed lines*) performance **remains low initially**, then **sharply increases after around 200 million samples**
  - This indicates that a foundational **semantic understanding precedes the development of text-specific recognition**.
  - Interestingly, the model's **preference for rendered text** images (*dotted lines*) also **develops later**
- **Implication** : This consistent, **scale-invariant emergence pattern** strongly suggests it's a fundamental learning dynamic in these VLMs, not just an artifact of a specific training run.

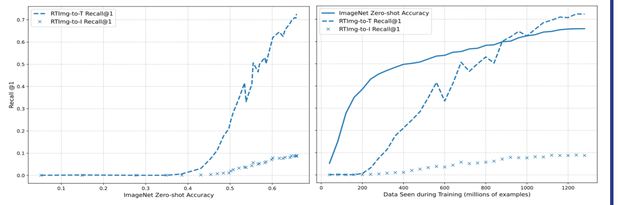
The "Emerge Point": A General Phenomenon



- **Extends to 114 OpenClip weights**
  - **Diverse Architectures (ViT-G, SigLIP, ConvNext, etc.,)**
  - **Diverse Pretraining Data (LAION, WebLi, DFN, etc.,)**
- **Critical Threshold (≈0.4 ImageNet Acc.)**: Text readability (RTImg-to-T) abruptly emerges when general semantic understanding (ImageNet Zero-shot Accuracy) reaches a critical threshold of approximately 0.4.
  - **Below** this threshold: Text readability is **near random**.
  - **Above** this threshold: Text readability **rapidly improves**.
- **General Phenomenon:** This "**emerge point**" is consistently observed across both our Datacomp-1B trained models (blue dots) and 114 diverse public OpenCLIP models (red dots), indicating it's a general characteristic.
- **Significance** : **A Shift in Capability**: This ≈0.4 threshold marks a crucial '**emerge point**' where VLMs **shift from primarily semantic processing to incorporating symbolic, text-based information**, effectively starting to 'read' and understand text within images.
- **Implication**: This delayed emergence suggests contrastive loss may prioritize general semantic learning first, with symbolic text understanding developing later through further refinement.

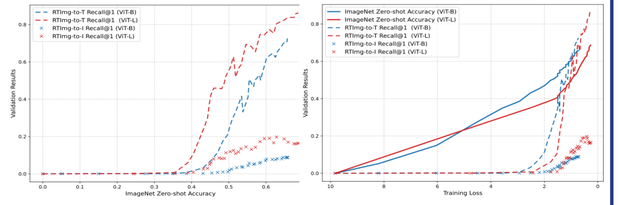
Beyond Pattern Matching: Deeper Understanding & Scaling

Deeper Understanding is Harder



- **True semantic understanding** (RTImg-to-I, *X-lines*) is **far harder** & emerges later than basic text recognition (RTImg-to-T, *dashed lines*).
  - Suggests **RTImg-to-T might be superficial**; RTImg-to-I requires deeper visual-semantic integration.
  - **Delayed RTImg-to-I Emergence**: This delay is likely because contrastive learning prioritizes direct image-text comparisons over rendered text-image alignment (despite both being visual inputs).

Scaling Helps, But Pattern Persists



- Larger models (ViT-L/16) improve RTImg-to-I, but the **delayed emergence pattern persists**.
  - Patterns of abrupt and delayed emergence hold across scales.

Conclusion & Future Works

Key Takeaways

- **Abrupt Emergence**: Text readability in VLMs is an *emergent capability*, not gradually learned.
- **Delayed Development**: It appears after general semantic understanding, around a consistent ~0.4 ImageNet Acc. threshold.
- **Deeper is Harder**: Semantic understanding of rendered text (RTImg-to-I) is even more challenging and emerges later, suggest text readability remains superficial.

Future Work

- Tailor training strategies for faster, robust text comprehension.
- Investigate underlying mechanisms of this emergence.