Seeing the Unseen: Visual Similarity for Pixel Language Model Adaptation

Ran Zhang, Miryam de Lhoneux, Wessel Poelman

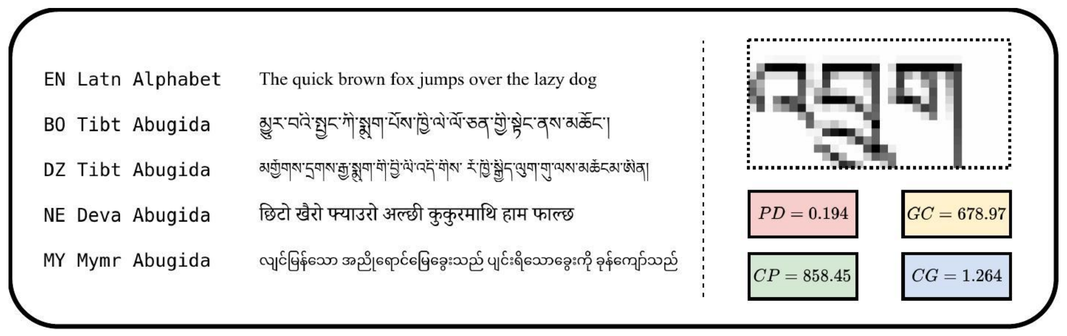
LAGoM NLP, Department of Computer Science, KU Leuven, Belgium

github.com/RAN-rz/seeing_the_unseen

Email: zrran0601@gmail.com

MOTIVATION

Pixel-based language models replace tokenizers with rendered images of text, making cross-lingual transfer depend on the visual and structural properties of writing systems. Yet how these models adapt to low-resource languages with unique scripts remains unexplored, and prior work only classifies "visual similarity" into coarse, binary categories (e.g. "Similar" vs. "Not-Similar", or shared script family). Using Tibetan, a morphologically rich, script-unique, and severely under-resourced language, as a case study, we ask: does a finer-grained, patch-level measure of visual similarity actually contribute to cross-lingual transfer in pixel language models, and how feasible is continued pre-training (CPT) into an entirely new script?



FOUR RENDERING-LEVEL VISUAL SIMILARITY METRICS

PD — **Pixel Density**
proportion of non-background (ink) pixels, a proxy for glyph coverage and visual complexity

GC — **Grapheme Clusters / block**
number of human-perceived characters encoded per rendered block

CP — **Code-points / block**
raw Unicode encoding length of each rendered block

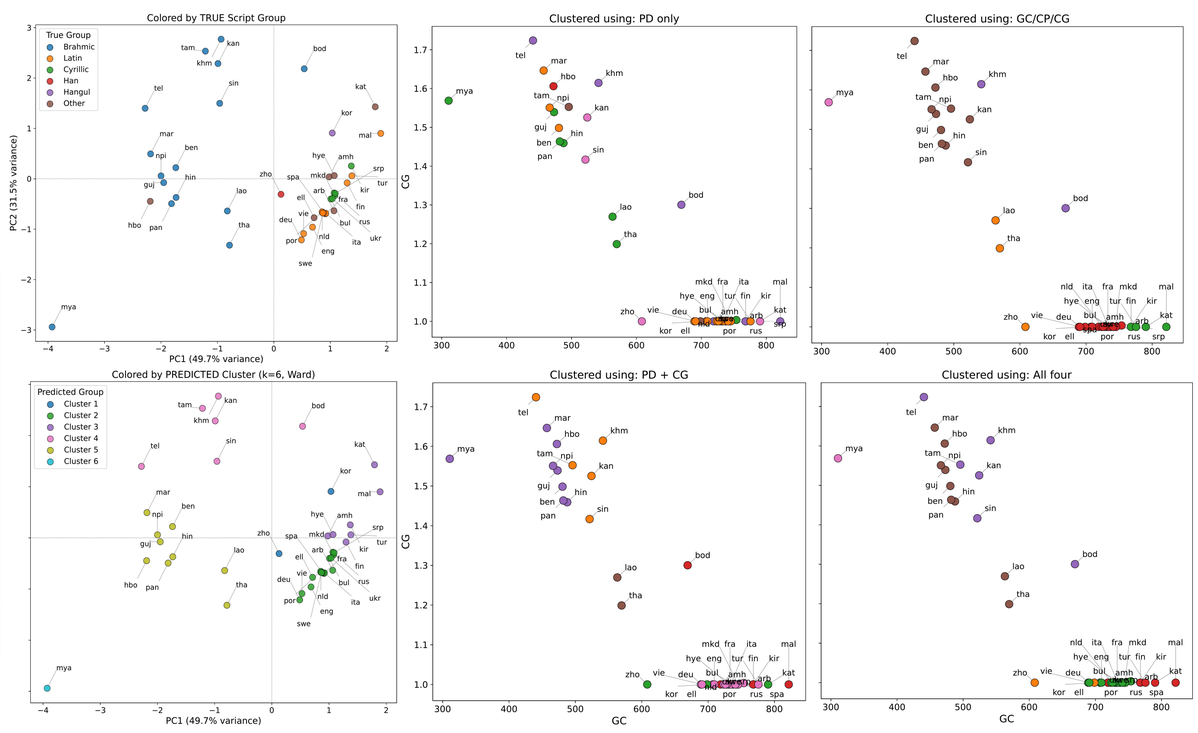
CG — **Code-point-to-Grapheme ratio**
how many code-points compose one visible character, captures stacking / conjunct compositionality

Table 1 — Visual Similarity Metrics on the Patch Level

Language	Script	PD	GC	CP	CG	Dist. to Tibetan
Tibetan (bo)	*Tibt*	0.19	678.97	858.45	1.26	– (target)
Dzongkha (dz)	*Tibt*	0.19	636.83	815.11	1.28	**0.52 (closest, same script)**
Khmer (km)	*Khmr*	0.21	541.49	874.24	1.62	2.02
Nepali (ne)	*Deva*	0.16	502.89	782.18	1.56	2.32
Chinese (zh)	*Hans*	0.25	608.31	608.31	1.00	2.70
Burmese (my)	*Mymr*	0.14	378.02	562.08	1.49	**3.97 (farthest)**

PD = Pixel Density · GC = Grapheme Clusters/block · CP = Code-points/block · CG = Code-point-to-Grapheme ratio
Distance = standardized Euclidean distance to Tibetan over the four metrics. For bo/dz/ne/my, computed on CPT pre-training data; for zh/km, on the 41-language PBC validation set — rankings are consistent across both corpora.

41-Language Validation of the Metrics



TRUE Script Group: Clusters broadly match script families, as expected.

PREDICTED Cluster: Recovers true groups well above chance (Purity=0.68, p<0.0001), suggesting the metrics capture finer-grained orthographic similarity beyond script-group labels

***PD* only**: Weakest single metric (Purity=0.585); *PD* alone isn't significant (p=0.637).

***PD* + *CG*:** Although *PD* shows weak discriminative power alone, its signal is distributed across interactions with the other metrics. The *PD*+*CG* pair alone already nearly matches the full four-metric result (Purity=0.683).

***GC*/*CP*/*CG*:** *GC* and *CG* are individually significant (p<0.0001); *CP* is not (like *PD*), but including all three still improves clustering quality (Purity=0.659).

All four: Best overall quality (Purity=0.683, ARI=0.325, NMI=0.548), final configuration.

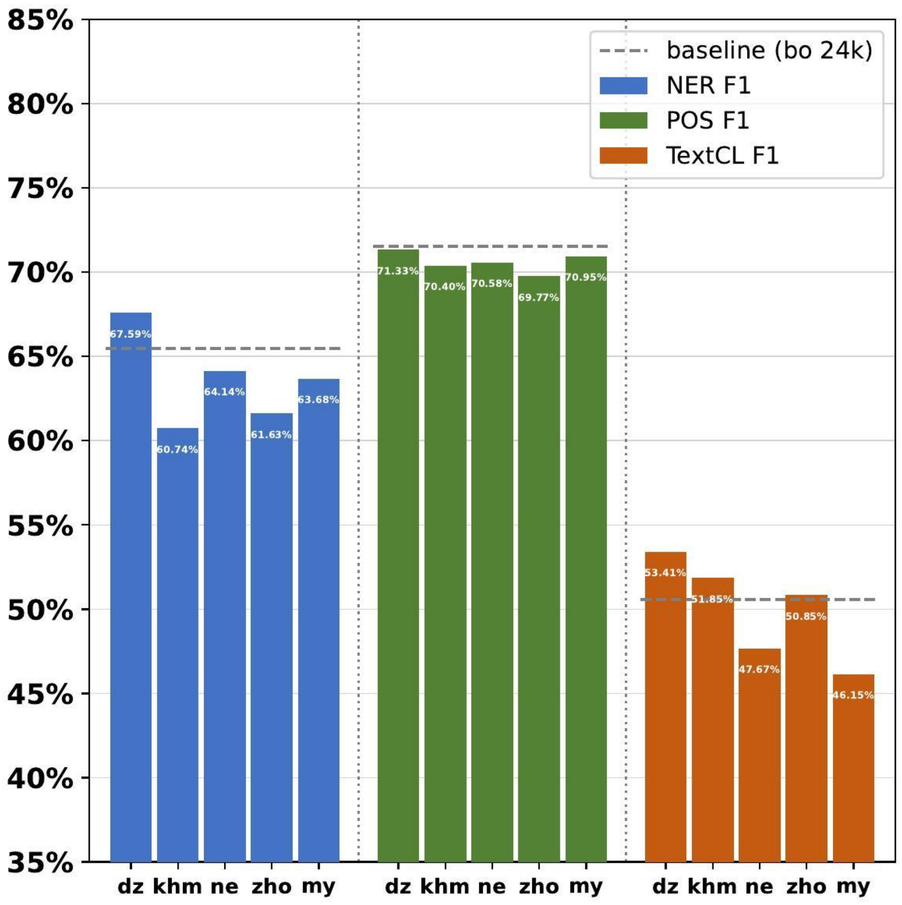
EXPERIMENTS & RESULTS

EXPERIMENTAL SETUP

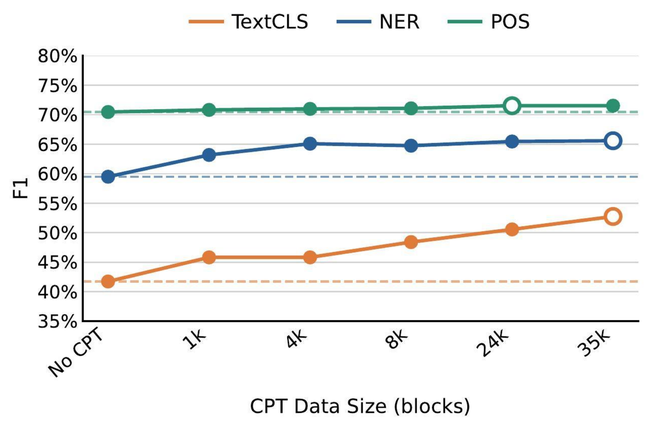
Two continued pre-training backbones: PIXEL-MONO — English-only pre-training (zero script overlap with Tibetan) and **PIXEL-M4** — multilingual (English, Hindi, Ukrainian, Chinese); Hindi also belongs to the Brahmic writing system.
Auxiliary CPT languages are chosen across the orthographic-proximity spectrum to Tibetan (see Table1).
CPT data scale is ablated at **0 / 1k / 4k / 8k / 24k / 35k** rendered blocks.
We evaluate on the same three Tibetan downstream tasks: **NER, POS tagging, and Text Classification;** and benchmarked against **Glot500 and CANINE-S** as reference points.

RESULTS

Orthographic Proximity → Semantic Transfer



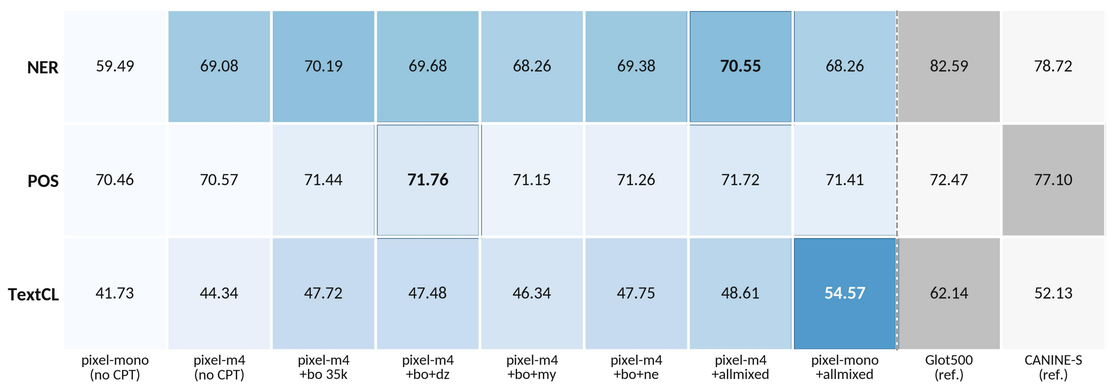
PIXEL-MONO + 24k CPT blocks. TextCL and NER broadly follow the proximity ranking: dz > km > ne > zh > my. Khmer, despite being the 2nd-closest language, gives the lowest NER score, showing the effect is task-dependent.



CPT Data Scale: Task-Dependent Sensitivity

Semantic tasks (TextCL, NER) improve steadily as Tibetan CPT data scales from 0 to 35k blocks; POS tagging stays nearly flat

CPT Results Across PIXEL-M4 Configurations (Table 2)



Among PIXEL-based conditions (left of dashed line), pixel-m4+allmixed gives the best NER, pixel-m4+bo+dz the best POS, and pixel-mono+allmixed the best TextCL (bold, boxed); yet Glot500 and CANINE-S (right, reference-only) still outperform most PIXEL configurations, especially on NER and POS.

CONCLUSION

- Four rendering-level metrics (*PD, GC, CP, CG*), computed on the exact patches PIXEL models process, give a fine-grained, empirically validated measure of visual similarity, verified across 41 languages and 6 script groups.
- Higher orthographic proximity to Tibetan generally improves semantic-task transfer under CPT, even under severe data constraints, but the effect is not uniform across languages or tasks (see the Khmer / NER counterexample).
- PIXEL-M4 (multilingual) starts stronger on Tibetan tasks, but PIXEL-MONO (monolingual) gains more from continued pre-training, suggesting a trade-off between initial representation strength and later adaptation capacity.

Limitations: *This case study prioritizes depth over breadth, Tibetan plus five auxiliary Brahmic/Han-script languages, under a controlled, matched data budget.*

This poster presentation is supported by the KU Leuven Master of Artificial Intelligence (MAI) programme.

Code & Models

KU Leuven MAI Programme