# Supplemental Material for FontCLIP: A Semantic Typography Visual-Language Model for Multilingual Font Applications

Yuki Tatsukawa[1] I-Chao Shen[1] Anran Qi[1] Yuki Koyama[2] Takeo Igarashi[3] Ariel Shamir[4]

[1] {tatsukawa-yuki537, ichaoshen, annranqi1024}@g.ecc.u-tokyo.ac.jp , The University of Tokyo, Japan
[2] koyama.y@aist.go.jp, National Institute of Advanced Industrial Science and Technology (AIST), Japan
[3] takeo@acm.org, The University of Tokyo, Japan
[4] arik@runi.ac.il, Reichman University, Israel

## 1. Statistics for Correlation Experiments

We provide the detailed statistics of the correlation experiments for *in-domain* attributes in Table 1 and *out-of-domain* attributes in Table 2.

| attribute name | CLIP | FontCLIP (w/o CDP) | FontCLIP |
|---|---|---|---|
| “angular” | 0.102 | 0.765 | 0.685 |
| “artistic” | 0.537 | 0.912 | 0.847 |
| “attention-grabbing” | 0.060 | 0.851 | 0.873 |
| “attractive” | 0.015 | 0.784 | 0.819 |
| “bad” | −0.066 | 0.655 | 0.598 |
| “boring” | −0.276 | 0.850 | 0.860 |
| “calm” | −0.091 | 0.806 | 0.830 |
| “capitals” | 0.343 | 229 | 0.579 |
| “charming” | 0.432 | 0.649 | 0.565 |
| “clumsy” | 0.318 | 0.769 | 0.748 |
| “complex” | −0.048 | 0.809 | 0.791 |
| “cursive” | 0.396 | 0.629 | 0.519 |
| “delicate” | 0.489 | 0.882 | 0.828 |
| “disorderly” | 0.239 | 0.773 | 0.751 |
| “display” | 0.301 | 0.723 | 0.579 |
| “dramatic” | 0.504 | 0.887 | 0.853 |
| “formal” | −0.198 | 0.500 | 0.608 |
| “fresh” | −0.110 | 0.259 | 0.316 |
| “friendly” | 0.137 | 0.674 | 0.647 |
| “gentle” | 0.376 | 0.565 | 0.584 |
| “graceful” | 0.341 | 0.723 | 0.827 |
| “happy” | 0.238 | 0.834 | 0.845 |
| “italic” | 0.410 | 0.826 | 0.828 |
| “legible” | −0.449 | 0.396 | 0.536 |
| “modern” | 0.18 | 0.849 | 0.843 |
| “monospace” | 0.538 | 0.410 | 0.439 |
| “playful” | 0.375 | 0.862 | 0.850 |
| “pretentious” | 0.337 | 0.766 | 0.875 |
| “sharp” | 0.159 | 0.670 | 0.632 |
| “serif” | −0.046 | 0.807 | 0.711 |
| “sloppy” | 0.137 | 0.619 | 0.684 |
| “soft” | 0.190 | 0.848 | 0.861 |
| “strong” | 0.150 | 0.699 | 0.855 |
| “technical” | −0.180 | 0.621 | 0.628 |
| “thin” | 0.147 | 0.955 | 0.911 |
| “warm” | 0.047 | 0.498 | 0.847 |
| “wide” | −0.13 | 0.683 | 0.672 |
| mean | 0.159 | 0.704 | 0.723 |
| std | 0.242 | 0.172 | 0.143 |

Table 1: The detailed correlation results for *in-domain* attributes experiment of CLIP, FontCLIP trained without using compound descriptive prompts (CDP), and FontCLIP trained with CDP (Ours).

| attribute name | CLIP | FontCLIP (w/o CDP) | FontCLIP |
|---|---|---|---|
| “angular” | 0.102 | −0.173 | 0.350 |
| “artistic” | 0.536 | 0.764 | 0.874 |
| “attention-grabbing” | 0.060 | 0.907 | 0.868 |
| “attractive” | 0.015 | 0.445 | 0.383 |
| “bad” | −0.066 | 0.610 | 0.480 |
| “boring” | −0.276 | −0.110 | −0.550 |
| “calm” | −0.091 | 0.102 | 0.206 |
| “capitals” | 0.343 | 0.181 | 0.339 |
| “charming” | 0.432 | 0.905 | 0.907 |
| “clumsy” | 0.318 | 0.256 | 0.611 |
| “complex” | −0.048 | 0.760 | 0.747 |
| “cursive” | 0.396 | 0.409 | 0.350 |
| “delicate” | 0.489 | 0.343 | 0.865 |
| “disorderly” | 0.239 | 0.739 | 0.607 |
| “display” | 0.301 | 0.476 | 0.575 |
| “dramatic” | 0.503 | 0.684 | 0.823 |
| “formal” | −0.198 | −0.442 | −0.194 |
| “fresh” | −0.110 | 0.031 | 0.416 |
| “friendly” | 0.137 | 0.377 | 0.227 |
| “gentle” | 0.376 | 0.554 | 0.909 |
| “graceful” | 0.341 | 0.812 | 0.745 |
| “happy” | 0.238 | 0.561 | 0.552 |
| “italic” | 0.410 | 0.471 | 0.661 |
| “legible” | −0.449 | -0.701 | −0.533 |
| “modern” | 0.183 | 0.125 | 0.442 |
| “monospace” | 0.538 | −0.127 | 0.295 |
| “playful” | 0.375 | 761 | 0.852 |
| “pretentious” | 0.337 | 0.373 | 0.648 |
| “serif” | −0.046 | −0.05 | −0.209 |
| “sharp” | 0.159 | −0.25 | −0.188 |
| “sloppy” | 0.137 | −0.08 | −0.191 |
| “soft” | 0.190 | 0.496 | 0.349 |
| “strong” | 0.150 | 0.175 | 0.597 |
| “technical” | −0.180 | 0.628 | −0.111 |
| “thin” | 0.147 | 0.393 | 0.489 |
| “warm” | 0.048 | 0.768 | 0.779 |
| “wide” | −0.139 | −0.221 | −0.018 |
| mean | 0.159 | 0.316 | 0.404 |
| std | 0.242 | 0.389 | 0.402 |

Table 2: The detailed correlation results for *out-of-domain* attributes experiment of CLIP, FontCLIP trained without using compound descriptive prompts (CDP), and FontCLIP trained with CDP (Ours).

## 2. Dual-Modal Multilingual Font Retrieval

### 2.1. Statistics for Chinese Character Pairwise Attribute Prediction

We provide the detailed statistics of the font retrieval task for Chinese characters in Table 3. The results presented in Table 3(b) clearly demonstrate that FontCLIP achieves higher accuracy with lower standard deviation. This indicates that FontCLIP consistently performs better in generalizing to *out-of-domain* attributes.

| attribute name | CLIP | FontCLIP |
|---|---|---|
| “thin” | 71.33% | 78.87% |
| “calm” | 61.44% | 50.98% |
| “sloppy” | 41.83% | 52.29% |
| “sharp” | 64.71% | 69.93% |
| “technical” | 35.29% | 68.62% |
| mean | 54.92% | 64.14% |
| std | 15.52% | 12.08% |

(a): Chinese characters with *in-domain* attributes

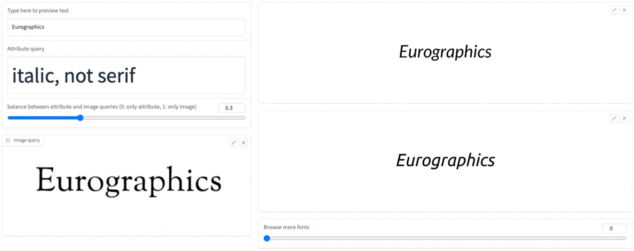
| attribute name | CLIP | FontCLIP |
|---|---|---|
| “traditional” | 49.02% | 62.74% |
| “Japanese style” | 33.33% | 63.40% |
| “robust” | 44.44% | 67.32% |
| mean | 42.26% | 64.48% |
| std | 8.07% | 2.48% |

(b): Chinese characters with *out-of-domain* attributes

Table 3: The accuracy of the pairwise attribute prediction task for Chinese characters with *in-domain* and *out-of-domain* attributes.

### 2.2. Dual-Modal Font Retrieval and User Interface

To perform multi-modal font retrieval, users can use our user interface as shown in Figure 1. This interface allows users to provide reference font images, specify desired attributes, and adjust the style weight $w$ using a slider.



Type here to preview text
Eurographics

Attribute query
italic, not serif

balance between attribute and image queries (0: only attribute, 1: only image) 0.3

Image query

Browse more fonts 0

Figure 1: The user interface for our multi-modal font retreival.