Pacific Graphics 2026
Y. He, N. Thürey, and L. Liu
(Guest Editors)

Volume 45 (2026), Number 7

Telligram: Text-Driven Calligram Generation via Diffusion-Guided Skeleton Optimization (Supplementary Material)

Tianci Shi and Pengfei Xu†

CSSE, Shenzhen University, China

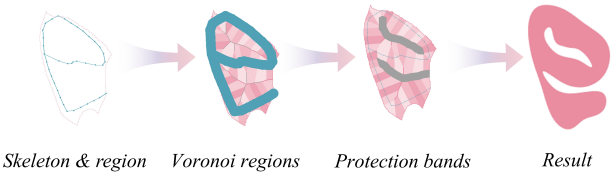

Figure 1: *Illustration of the Voronoi partition used in Section 4. Skeleton-point samples induce local Voronoi regions, and their boundaries provide candidate separating lines for constructing the protection band.*

This supplementary document provides additional details that were omitted from the main paper for space reasons.

1. Voronoi-Based Constraint and Protection Band

In the main paper, Section 4 uses Voronoi-assisted filling to preserve local letter details during region thickening. Here we explain more concretely how the Voronoi-based constraint and its protection band are constructed and when they are applied.

As illustrated in Fig. 1, suppose that the k-th letter has a fitted skeleton $\mathcal{S}_k$ and an assigned support region M_k. A direct filling strategy would simply thicken $\mathcal{S}_k$ inside M_k until the region is covered as much as possible. However, this direct strategy may remove narrow separations between nearby letter parts. In practice, the first failure often appears near region boundaries induced by neighboring skeleton branches.

To reduce this problem, we first sample points along the fitted skeletons and build a Voronoi partition from these skeleton points. This partition divides the local space according to nearest skeleton-point distance. The resulting Voronoi boundaries indicate where two nearby skeleton branches compete for the same area. These boundaries are therefore a useful signal for where direct thickening may merge structures too early.

Not every Voronoi boundary should be turned into protection. Some boundaries lie far from the actual letter strokes and have little effect on readability. We only keep the boundaries that stay within the skeleton protection band, that is, within a narrow neighborhood of the fitted skeleton. Intuitively, these are the boundaries that are close enough to real letter structure to matter during filling, rather than distant partitions caused only by global geometry.

After selecting such boundaries, we thicken them into a narrow band and use this band as the protection band. In other words, this band is removed from the fillable region before the final dilation step. This creates a thin local barrier near sensitive branch boundaries, so the thickened glyph is less likely to erase fine gaps or merge adjacent parts too early.

In the actual implementation, this protection band is not applied everywhere by default. We only allow the protection band to overwrite a local area when the current fill in that area already exceeds 90% coverage. The reason is practical: when the local region is still far from filled, enforcing the protection too early may leave obvious holes and reduce overall glyph completeness. By activating it only after the local filling ratio is already high, the method uses the protection band mainly as a late refinement step. This keeps most of the region fillable at early stages, while still preserving thin local structures near the end of the realization process.

Overall, this mechanism plays a selective role. The Voronoi partition identifies candidate separating lines, the skeleton protection band filters them to keep only structurally relevant ones, and the 90% coverage rule prevents premature blocking. Together, these steps make the final filling more stable and more readable than direct thickening alone.

2. Multimodal Experiments

We also tested whether a general multimodal image model could directly generate calligram-like results from prompting alone. In particular, we used prompts of the form “a silhouette of a bunny made of word ‘bunny’” and asked GPT Image 2.0 to produce semantic text-shaped images.

In practice, the generated results mainly fall into two failure modes. In the first mode, the model uses many repeated copies of the same word to pile up a recognizable semantic silhouette, as shown on the left side of Fig. 2. This can match the target concept at the image level, but it does not produce a readable calligram in which the word is organized as one coherent glyph structure.

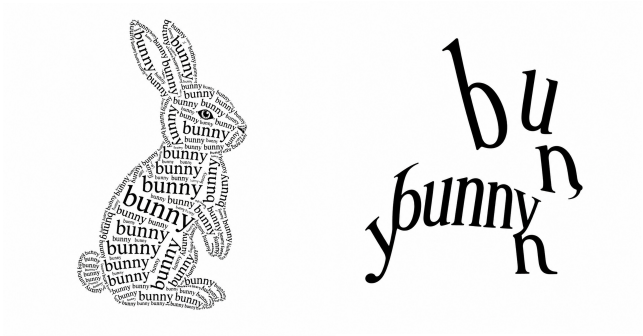

Figure 2: *Examples from multimodal prompting with text-shaped silhouette requests. The left side shows semantic shapes formed by many repeated copies of the same word, while the right side shows simple local letter deformation without a clear global semantic figure.*

In the second mode, the model uses only a few letters and applies simple local deformation to them. Although these outputs may preserve letter appearance, they usually fail to form a recognizable semantic shape at the global level. As a result, the output is neither a clear semantic figure nor a readable whole-word calligram.

These observations further support the motivation of our method. The difficulty is not only to match a target concept but to coordinate semantic shape formation with structured letter occupancy and readable glyph realization. This is why our method separates semantic occupancy formation from later geometric realization instead of relying on prompt-only image generation.

3. Per-Category Semantic Recognition Results

To support the limitation-oriented semantic recognition analysis in the main paper, we report the per-category results of the balanced CLIP zero-shot top-5 evaluation here. Each category contains 15 generated samples. A sample is counted as a top-5 semantic match if the predictions contain either the target category itself or a more specific label with the same target noun. As discussed in the main paper, this analysis is intended only as a coarse category-level indicator for silhouette-like outputs.

Category	Total	Top-5 Matches
camel	15	15
bear	15	15
butterfly	15	15
cattle	15	15
dinosaur	15	15
kitten	15	15
duck	15	15
elephant	15	15
goose	15	15
horse	15	15
mammoth	15	15
sheep	15	15
submarine	15	15
teapot	15	15
crown	15	14
dog	15	14
giraffe	15	14
umbrella	15	14
bunny	15	13
chicken	15	13
rat	15	13
shark	15	11
kangaroo	15	10
rhino	15	10
crocodile	15	9
banana	15	9
pig	15	7
pufferfish	15	7
plane	15	6
swan	15	6
owl	15	5
bicycle	15	4
eagle	15	0
flamingo	15	0
motorcycle	15	0
violin	15	0

Table 1: *Per-category results for the balanced 36-category CLIP zero-shot top-5 semantic recognition analysis. Each category contains 15 generated samples.*