SCORE: Semantic Collage by Optimizing Rendered Elements

Zefan Shao, Jin Zhou, Hongliang Yang, Pengfei Xu*

College of Computer Science and Software Engineering, Shenzhen University, China

Abstract

Collage is a powerful medium for visual expression, traditionally demanding significant artistic expertise and manual effort. Existing methods often struggle with a trade-off between semantic expression and the visual fidelity of the constituent images. To address this, we introduce **SCORE** (**S**emantic **C**ollage by **O**ptimizing **R**endered **E**lements), a novel text-driven framework that automates the creation of semantically rich and structurally sound collages. Our key innovation is to shift the optimization process entirely into the image space. By employing a differentiable renderer, we can backpropagate gradients from a powerful, pre-trained text-to-image model directly to the spatial parameters, including position, rotation, and scale, of each image element. We leverage Variational Score Distillation (VSD) to provide robust semantic guidance from a text prompt, ensuring the final layout aligns with the desired concept. Crucially, our "minimal editing" principle preserves the integrity of the original elements by forgoing any content-level modifications. The layout is refined by a joint loss function that combines the VSD-based semantic loss with structural regularizers that penalize overlap and enforce boundary constraints. The output of **SCORE** is a parametric, structured representation that allows further editing and downstream use. Our work reduces the barrier to creative expression and provides a new, powerful paradigm for organizing visual contents.

Introduction

Assembling and collaging visual elements to encapsulate features or concepts provides a unified and intuitive representation, which has been instrumental in creating intriguing visual designs and artworks (Spielmann 1999; Wang et al. 2006). With the advent of digital tools, the technical barriers to creating collages have been reduced, allowing easier manipulation of pixels on a screen. Despite this convenience, the core artistic challenge persists: the process remains heavily reliant on the creator's experience and manual effort in both selecting materials and arranging them to convey a specific theme harmoniously. This still presents a significant barrier for non-experts. Consequently, automatically arranging a set of discrete image elements into a coherent and meaningful collage guided by high-level semantics has become a key problem that urgently needs to be solved.

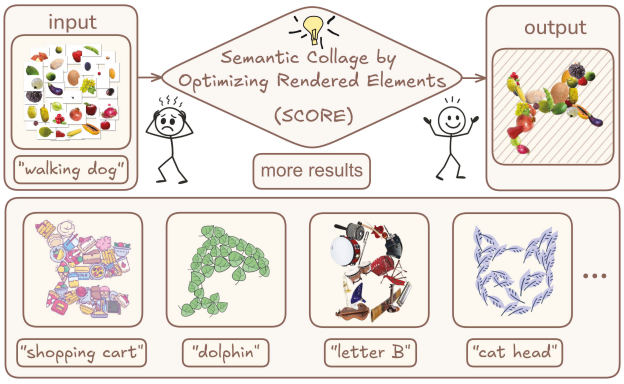

Figure 1: Overview of **SCORE**. Given a set of image elements and a text prompt, our method optimizes the spatial arrangement of visual elements to generate a semantically aligned collage.

Numerous techniques have been proposed to address similar tasks, with the majority of existing methodologies concentrating on object-space optimization (Kwan et al. 2016; Wang et al. 2019), a paradigm that operates directly on the mathematical descriptions (e.g., vertices and curves) of the geometric elements. In this paradigm, collage generation is often framed as a geometric constraint satisfaction problem, in which elements are filled within a fixed contour. However, this approach has inherent limitations. It often requires hand-crafting complex geometric descriptors and energy functions, which may lack generalizability across different shapes. When we want to express a certain concept with collage, a fixed contour can sometimes become a limitation. Furthermore, learning-based methods, such as Neural Collage Transfer (Lee et al. 2023), have explored artistic reconstruction; however, they typically rely on a reference style image and can introduce severe, undesirable deformations to the source image elements. While other text-driven methods, such as CLIP-CLOP (Mirowski et al. 2022), have emerged, they may not sufficiently prioritise the visual fidelity of the original image elements, thereby weakening their individual expressive ability, which goes against the

*Corresponding author, e-mail: xupengfei.cg@gmail.com

concept of collage.

In this work, we propose **SCORE** (**S**emantic **C**ollage by **O**ptimizing **R**endered **E**lements), a novel framework for generating text-driven image collages in image space. The core idea is to leverage the powerful prior knowledge of large-scale text-to-image models and translate a high-level text prompt into an optimization objective. We employ a differentiable renderer to construct the collage, which allows gradients from our objective function to backpropagate directly to the spatial transformation parameters (i.e., position, rotation, and scale) of each image element.

The key contributions of our work are threefold:

- **Semantic-driven Layout over Template Constraints.** We introduce a paradigm that directly uses text to drive the layout optimization. By utilizing Variational Score Distillation (VSD) as a semantic loss, our method bypasses the need for intermediate representations such as contours or reference images, thereby significantly enhancing creative freedom and the precision of semantic alignment.
- **Image element Fidelity via Minimal Editing.** Our optimization process is strictly confined to adjusting the position, rotation, and scale of the image elements. No distortion, cropping, or content-level modifications are performed. This "minimal editing" strategy maximally preserves the integrity and visual clarity of the original images, resulting in highly interpretable final compositions that respect traditional collage practices.
- **Joint Optimization of Semantics and Structure.** We design a composite loss function that synergistically combines the semantic guidance from the pre-trained generative model with a suite of structural regularizers. These regularizers, which include penalties for overlap, size, and boundary constraints, ensure the resulting layout is not only semantically meaningful but also structurally sound and visually organized.

SCORE can generate diverse, high-quality collages that are highly consistent with the input text prompt. Our final high-resolution rendering stage further produces a clean, transparent-background asset that is ready for downstream applications. Moreover, the output of our method is not a monolithic image but a form of quantized imagery: a structured, parametric representation that facilitates further editing. Our work not only further reduces the barrier for creative expression but also provides a new, powerful paradigm for intelligent and automated visual content organization.

Related Work

Prior research broadly falls into three lines: geometry-based packing, image-based collages, and semantic-guided generation. We discuss these works as follows.

Geometry-Based Packing

Traditional geometry-driven approaches seek compact, non-overlapping layouts by abstracting elements into shapes and optimizing space efficiency. Classic strategies span regular and irregular packing (Wang et al. 2006; Itoh et al. 2004; Kwan et al. 2016; Saputra, Kaplan, and Asente 2018, 2019; Yao et al. 2025), with related work on letter or calligraphic packing using similar formulations (Xu and Kaplan 2007; Zou et al. 2016). Another branch partitions the target region first, e.g., via segmentation or circle packing, and then fills the parts with primitives or images (Kim, Pellacini et al. 2002; Yu et al. 2014). These structure-aware techniques sometimes blur into collage variants when images are used as fill units. Recent efforts also avoid explicit object reasoning by optimizing directly in image space with differentiable rendering and repulsion constraints (Wang and Lu 2025). **SCORE** optimizes placements directly in image space with semantics as the objective, without relying on contour priors, yielding greater flexibility.

Image-Based Collages

Collage methods operate on visual content itself (cutouts or patches) rather than purely on geometric proxies. Early systems such as AutoCollage (Rother et al. 2006) select and compose salient photos, while other work fuses significant regions (Goferman, Tal, and Zelnik-Manor 2010) or preserves semantic structure (Liu et al. 2017). A closely related thread is image mosaics, which approximate a target image by tiling and replacing tiles with best-matching images; advances in partitioning and replacement improve fidelity (Pavić, Ceumern, and Kobbelt 2009; Zhang, Ma, and Yu 2016; Xu et al. 2019). Reinforcement learning has been explored to simulate collage assembly (Lee et al. 2023), and Arcimboldo-style collages divide cropped objects into regions (Huang, Zhang, and Zhang 2011), both of which can be viewed as mosaic variants. Despite their visual appeal, prior image-based collages typically approximate a target by tiling or by hand-crafted constraints, with limited capacity to express text-specified semantics and with frequent distortions to source image elements. **SCORE** instead couples text-driven objectives with differentiable layout optimization, achieving semantic alignment without modifying image element content.

Semantic-Guided Generation

With large vision-language models, spatial relationships can be conditioned on text. Some methods predict layouts directly from language (Tang et al. 2023; Srivastava et al. 2025): LayoutDiffusion discretizes sequences of bounding boxes and learns their distribution (Zhang et al. 2023), while LayoutGPT treats layout as sequence generation over boxes or symbolic graphs (Feng et al. 2023). Final images are produced via layout-to-image pipelines (Li et al. 2023). These systems excel at structured scenes but generally output symbolic layouts rather than visual compositions, and they seldom account for the appearance or fidelity of image elements. A complementary line optimizes placements of given components to match target semantics. CLIPCLOP (Mirowski et al. 2022) formulates differentiable collage placement to maximize CLIP image-text scores and also tunes color mappings, which can distort sources. Such approaches may struggle with few inputs and often rely on dense stacking to accumulate semantic cues. **SCORE** can

perform semantic-driven placement while preserving element appearance, without relying on overlap or color remapping, and can express the semantics with a few elements.

Summary

Across these categories, there is a persistent trade-off: preserving the appearance of input image elements tends to weaken semantic alignment, while strong semantic control often edits or regenerates content, harming fidelity. We introduce **SCORE** to mitigate this tension. Given image elements, our method optimizes a parameterized collage that follows semantic cues without heavy editing or excessive overlap, achieving competitive results with fewer elements.

Preliminaries

In this section, we review the key foundations of our method: text-to-image diffusion models and the score distillation techniques used to guide optimization.

Diffusion Models

Diffusion Probabilistic Models (DPMs) (Ho, Jain, and Abbeel 2020) are a class of generative models that synthesize images by reversing a gradual noising process. The process consists of a fixed **forward process** and a learned **reverse process**.

In the forward process, Gaussian noise is added to a clean image x_0 over T steps, resulting in a progressively noisier sample x_t. The closed-form of this process at time t is:

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}), \tag{1}$$

where $\bar{\alpha}_t$ is a precomputed noise schedule.

The reverse process is learned by a neural network ϵ_θ, which predicts the added noise ϵ from a noisy image x_t, conditioned on a text prompt y and timestep t. The training objective minimizes the denoising error:

$$\mathcal{L}_{\text{DPM}} = \mathbb{E}_{t, x_0, \epsilon, y}\left[\|\epsilon - \epsilon_\theta(x_t, y, t)\|^2\right]. \tag{2}$$

In this work, We leverage DeepFloyd-IF (Shonenkov et al. 2023) to guide our collage generation process.

Score Distillation Sampling (SDS)

Originally proposed in DreamFusion (Poole et al. 2022) for text-to-3D generation, Score Distillation Sampling (SDS) enables the use of a pretrained 2D diffusion model as a loss function to optimize a differentiable scene representation, such as a NeRF or, in our case, a set of collage parameters.

At each step, the current output of the generator $g(\theta)$ is rendered into an image and noised like Equation 1.The SDS loss computes the difference between the true noise ϵ and the frozen diffusion model's prediction ϵ_ϕ, and backpropagates this difference into the generator:

$$\nabla_\theta \mathcal{L}_{\text{SDS}} = \mathbb{E}_{t,\epsilon}\left[w(t)\left(\epsilon_\phi(x_t; y, t) - \epsilon\right)\frac{\partial g(\theta)}{\partial \theta}\right], \tag{3}$$

where $w(t)$ is a weighting function that balances the contribution of different timesteps. This approach allows the model to optimize its parameters using only a text prompt, without requiring paired supervision. However, SDS often exhibits instability and sensitivity to randomness during the optimization procedure.

Variational Score Distillation (VSD)

Variational Score Distillation (VSD) (Wang et al. 2023) addresses the limitations of SDS by reframing optimization as a variational inference problem. It introduces a trainable "lifted" distribution within the diffusion model to adapt to the specific scene being optimized.

Concretely, VSD fine-tunes a set of parameters $\Delta\theta$ (e.g., via LoRA (Hu et al. 2022)) within the pretrained diffusion model, resulting in an adapted noise predictor $\epsilon_{\theta+\Delta\theta}$. The VSD loss compares the score from the adapted model to the original frozen model:

$$\nabla_\theta \mathcal{L}_{\text{VSD}} = \mathbb{E}_{t,\varepsilon,c}\Big[w(t)\big(\epsilon_{\theta+\Delta\theta}(x_t, y, t) - \epsilon_\phi(x_t, y, t)\big)\frac{\partial g(\theta)}{\partial \theta}\Big]. \tag{4}$$

This formulation encourages the generator $g(\theta)$ to produce samples that are more favored by the adapted model than by the original one, thereby avoiding the mode collapse seen in SDS. Meanwhile, the LoRA parameters $\Delta\theta$ are also updated to provide a scene-specific refinement of the diffusion prior.

Due to its greater stability and semantic flexibility, our **SCORE** framework builds upon the VSD technique as its core guidance mechanism.

Methodology

Our proposed framework, **SCORE**, transforms a collection of images into a semantically coherent collage guided by a single text prompt P. Before entering the core optimization pipeline, we first preprocess the input images. The process then proceeds through two main stages: (1) Layout Parameter Initialization and (2) Image-Space Optimization. The entire pipeline operates primarily in the image space, as illustrated in Figure 2. The output of our method is a set of transformation parameters, offering flexibility for subsequent editing and applications.

Problem Formulation

The core task is to determine the optimal layout of a given set of images on a 2D canvas to visually represent a concept described by a text prompt.

Input. An image element set $\mathcal{E} = \{e_1, \ldots, e_N\}$ and a text prompt P.

Output. A set of 2D affine transformation parameters $\Theta = \{\theta_1, \ldots, \theta_n\}$, where each $\theta_i = \{t_i, s_i, r_i\}$ corresponds to the translation, scale, and rotation for image e_i. A collage work that follows the text prompt P can be rendered with image elements and these parameters.

Input Preparation

We assume that the input elements are a collection of foreground cutouts. To prepare them for composition, we perform a pre-processing step to remove the backgrounds and isolate the salient content of images. This is typically achieved using automated tools like Rembg (Gatis et al. 2020), resulting in a set of RGBA image elements with

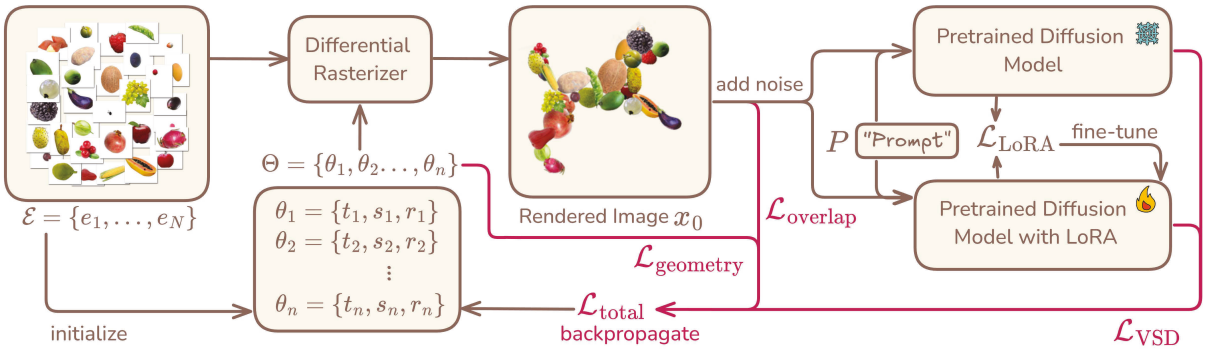

Figure 2: Overview of the **SCORE** pipeline. Given a set of foreground images and a text prompt P, we first perform background removal to isolate each object. Then, we initialize a coarse layout by assigning initial translation t, rotation r, and scale parameters s. Collage result is iteratively optimized in the image space using a differentiable renderer and a joint loss function, comprising a semantic loss guided by a diffusion model and structural losses that enforce visual coherence. The final output is a set of transformation parameters enabling high-quality, editable collage composition.

transparent backgrounds, which are essential for seamless blending in the subsequent rendering stages.

Layout Initialization

A proper initialization is crucial for the stability and efficiency of the optimization process. To avoid trivial local minima and provide a reasonable starting point, we initialize the layout with the following strategy:

- **Translation** (t): The initial (x, y) positions of all elements are uniformly distributed across the canvas to provide a diverse and unbiased starting layout.
- **Rotation** (r): Each image element is given a small random rotation, sampled from a narrow range (e.g., ± 0.1 radians), to introduce slight angular diversity.
- **Scale** (s): The initial scale of all elements is determined by the total number of image elements (N), providing a reasonable starting density:

$$s_{\text{init}} = (1/N)^{0.5} + 0.01. \tag{5}$$

Image-Space Optimization

This stage is the core of our framework, where we iteratively refine the layout parameters Θ by minimizing a joint loss function in the image space.

Differentiable Renderer. The differentiable renderer is the bridge between the geometric parameters Θ and the loss of the raster image space. Traditional affine transformations are non-differentiable when applied to a discrete pixel grid. As shown in the Figure 3, inspired by DiffVG (Li et al. 2020), we adopt a Gaussian Splatting-like approach to enable smooth, differentiable rasterization.

For each image element e_i, its pixels are treated as a point cloud and transformed using parameters θ_i. The color and alpha of each point are then splatted onto the canvas using a Gaussian kernel, so that nearby pixels receive proportionally higher weights. This process smooths discrete shifts and ensures a differentiable path from the rendered image to θ_i.

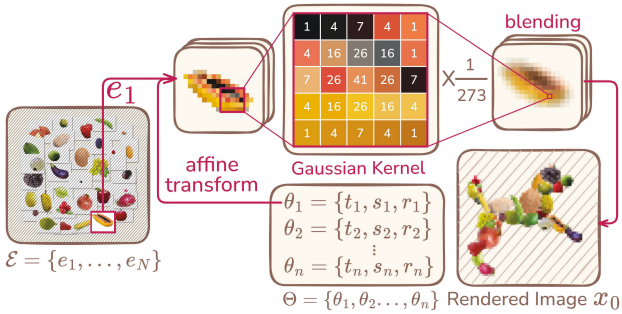

Figure 3: Differentiable renderer module. Pixels of each transformed element are treated as points and splatted onto the canvas with Gaussian kernels, yielding smooth, differentiable rasterization. (The shown 5×5 discrete Gaussian is only for visualization, not a fixed kernel. In practice, we inverse-sample, gather a square $p\times p$ neighborhood, compute per-pixel Gaussian weights, then composite layers.)

After rendering each element into its layer, we use standard alpha blending to composite the final image for loss calculation. The order of composition is fixed after a random initialization. During the optimization phase, overlapping parts are blended, so occlusions are not considered. The rendered canvas for supervision is a low-resolution raster image, which is sufficient to capture the global semantics conveyed by the current parameter composition.

Joint Loss Function. The overall loss used in our optimization pipeline consists of two components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{VSD}} + w_r \cdot \mathcal{L}_{\text{render}}. \tag{6}$$

Variational Score Distillation Loss ($\mathcal{L}_{\text{VSD}}$). This loss guides the collage to be semantically aligned with the text prompt P by leveraging a pre-trained diffusion model. For each optimization step, a temporary LoRA model is trained to approximate the noise prediction of the rendered image.

At each VSD step, the LoRA adapter is trained for a single iteration using the current noisy latents z_t and text embeddings. The LoRA loss is formulated as:

$$\mathcal{L}_{\text{LoRA}} = \mathbb{E}_{t,\epsilon}\left[\|\epsilon - \epsilon_{\theta+\Delta\theta}(z_t, t, c)\|_2^2\right], \tag{7}$$

where $\Delta\theta$ represents the LoRA parameters and $\epsilon_{\theta+\Delta\theta}$ is the adapted denoising model. This rapid specialization allows the LoRA model to become an expert denoiser for the specific content and style of the current collage state.

The VSD loss is implemented as a weighted difference between the base diffusion model's output and the LoRA model's output, after adding noise to the current canvas image (see Equation 4 for details).Encourages the rendered collage to evolve towards semantic consistency with the prompt P. The use of LoRA enables efficient adaptation without catastrophic forgetting of the pre-trained knowledge, while the dynamic retraining ensures that the guidance remains relevant throughout the optimization process.

Render Loss ($\mathcal{L}_{\text{render}}$). This term regularizes the geometric and visual quality of the layout. It comprises penalties on the translation and scale of individual elements, as well as a soft overlap penalty that prevents excessive stacking:

$$\mathcal{L}_{\text{render}} = \mathcal{L}_{\text{geometry}} + w_\alpha \cdot s \cdot \mathcal{L}_{\text{overlap}}, \tag{8}$$

where s is the normalized optimization step index, and w_α is a predefined weight. $\mathcal{L}_{\text{geometry}}$ and $\mathcal{L}_{\text{overlap}}$ are the **Geometry Loss** and **Overlap Loss**, respectively. The **Geometry Loss** is defined as:

$$\mathcal{L}_{\text{geometry}} = \sum_{i=1}^{N}\Bigg(\sum_{j\in\{x,y\}}\left[\text{ReLU}(|t_{i,j}| - t_{\text{lim}})\right]^2 + \left[\text{ReLU}(|s_i - s_{\text{ref}}| - s_{\text{lim}})\right]^2\Bigg), \tag{9}$$

which penalizes element centers t_i from drifting beyond the allowed boundary limit t_{lim} and regularizes the scale s_i of each element to be within the bounds $[s_{\text{ref}} - s_{\text{lim}}, s_{\text{ref}} + s_{\text{lim}}]$.

Let $A_{\text{accum}}(u, v)$ denotes the accumulated alpha value at pixel (u, v):

$$A_{\text{accum}}(u, v) = \sum_{i=1}^{N} \alpha_i(u, v). \tag{10}$$

The **Overlap Loss** is then defined as:

$$\mathcal{L}_{\text{overlap}} = \mathbb{E}_{(u,v)\in\Omega^+}\left[\text{ReLU}(A_{\text{accum}}(u, v) - \tau_\alpha)\right], \tag{11}$$

where Ω^+ is the set of pixels where $A_{\text{accum}}(u, v) > 0$, and τ_α is the threshold (typically 1.0).

Overall, this joint loss balances semantic alignment with layout regularity and enables a stable optimization path toward producing semantically coherent and visually pleasing image collages.

Optimization and Output. After a fixed number of optimization iterations, we get a structured set of optimal transformation parameters Θ_{final}. We render the elements on a high-resolution canvas based on Θ_{final}, and get the results shown in the next section.

Results and Experiments

To evaluate the effectiveness of **SCORE**, we designed a comprehensive set of experiments. Our evaluation framework is intentionally unified, relying on high-level semantic and aesthetic models to assess performance. This allows for a fair and direct comparison across all methods and prompt types, moving beyond traditional geometric metrics to capture the perceptual quality of the final collage.

Experimental Setup

Dataset. We conducted experiments on three thematic image collections: "Fruit", "Food Icons" and "Single Icon", which cover three common collage forms. The images were collected from the Internet. To ensure a standardized evaluation, we create fixed test sets with 16, 32, and 64 images by random selection for each theme.

Baselines. We compare **SCORE** against three carefully selected baselines: CLIP-CLOP (Mirowski et al. 2022), ShapeCollage (ShapeCollage 2025), and Image-Space Collage (Wang and Lu 2025). CLIP-CLOP is the most relevant prior method. It also performs text-driven optimization to arrange a fixed image set, so it enables a direct, like-for-like comparison on the core task. Packing algorithms can approach an upper bound when a target contour is known. We therefore include ShapeCollage and the recent Image-Space Collage as packing-based references: for concrete shape prompts, we give them the ground-truth contours and let them optimize image element placement inside them. These two baselines show what a mature, specialized packing method can achieve with perfect geometric priors.

Text Prompts. To ensure a fair and direct comparison with the packing-based baselines, including ShapeCollage and Image-Space Collage, we source our prompts from the MPEG-7 Core Experiment CE-Shape-1 Test Set (Sikora 2002). We use the official contours from this dataset as the ground truth for the packing-based methods and use the corresponding object names (e.g., "butterfly", "helicopter") as text prompts for our method and CLIP-CLOP. This setup allows for a rigorous evaluation of shape formation capabilities.

Evaluation Metrics

We evaluate with five metrics covering semantics, shape, aesthetics, human-proxy judgment, and structure: **CLIP-RGB** (semantic alignment) computes the CLIP similarity between the rendered collage and the text prompt (Radford et al. 2021); **CLIP-Contour** (shape) is our metric that converts the collage to a binary contour map and measures its CLIP similarity to the prompt, isolating shape from color; **Aesthetic Score** uses a LAION-trained aesthetic predictor (Schuhmann et al. 2022); **GPT-4o Score** asks GPT-4o to rate 1-10 on *Contour Expression*, *Layout Rationality*, and *Aesthetic Quality*, consistent with evidence that LLM-based judgments (e.g., GPT-4V) correlate with human preferences for compositional T2I tasks (Huang et al. 2023); **Overlap Ratio** (diagnostic) is the total area of overlapping regions between rendered elements divided by the sum of element

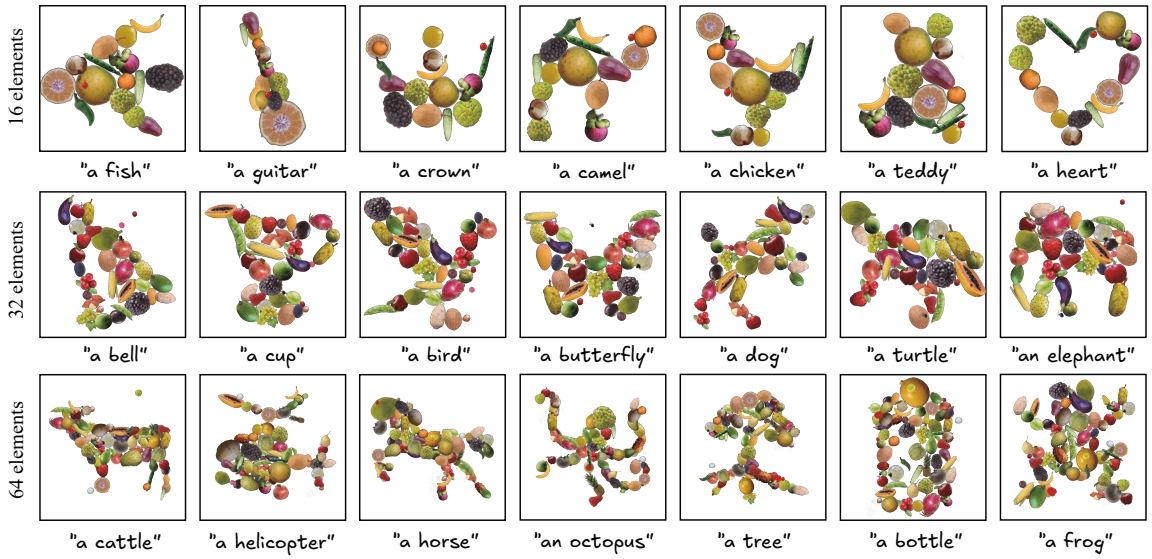

Figure 4: Results generated with our method in the theme of fruit. For easy comparison, the elements we use are three predetermined sets of three magnitudes. More results are provided in the supplementary.

Method	Size	CLIP-RGB↑	CLIP-Contour↑	Aesthetic↑	VLM↑
CLIP-CLOP	16	20.3871	21.8230	4.4530	5.3
	32	21.9691	23.6119	4.4092	5.2
	64	22.1453	24.2226	4.5256	5.5
Shape Collage	16	19.1100	21.3674	4.0400	4.5
	32	19.3360	22.0394	4.2381	5.3
	64	19.6338	22.4230	4.1598	6.5
Image-Space Collage	16	20.3326	22.7278	4.6847	7.2
	32	21.5848	24.1972	4.9692	7.2
	64	21.8722	25.0503	4.9030	7.6
Ours (**SCORE**)	16	21.2721	26.4064	5.3359	8.2
	32	22.1625	27.1836	5.6324	**8.6**
	64	**22.2248**	**27.5407**	**5.6688**	8.3

Table 1: Quantitative comparison between baselines and our method on different numbers of elements.

areas. We report the Overlap Ratio only in ablations to quantify the effect of $\mathcal{L}_{\text{overlap}}$, not to rank overall quality.

Main Results

We summarize our main findings through both quantitative and qualitative comparisons.

Quantitative Comparison. Table 1 compares our method (**SCORE**) with the baselines across all metrics. For semantic methods (CLIP-CLOP, **SCORE**), each prompt is run three times and averaged. Performance improves as the number of input images increases, as expected; notably, **SCORE** leads in the low-input regime and retains its advantage as inputs grow. With 64 images, it achieves the best core scores and also ranks highest under the LLM-based evaluation. Overall, **SCORE** surpasses semantic and packing baselines in semantic, structural, and aesthetic quality, showing strong expressiveness even with few inputs.

Qualitative Comparison. In Figure 5, we provide a side-by-side visual comparison of results generated by our method and the baselines for a specific prompt (i.e., "a dog"). This figure will visually substantiate the quantitative findings, highlighting **SCORE**'s superior ability to generate semantically coherent and well-structured collages. More results generated by our method are shown in Figure 4.

User study. We conducted a user study with 50 participants to assess how well each method aligns with human semantic perception. In each trial, participants were shown four collages generated from the same textual theme and asked to select the result that best matched the description of the target. The order of the four images was randomized to avoid positional bias. Across all participants and prompts, our method was selected in 59% of the trials, compared to 20.8% for Image-Space Collage, 11.8% for Shape Collage, and 8.4% for CLIP-CLOP. These results indicate that users consistently judged our collages as most faithful to the intended semantics, confirming that our approach better captures text-guided layout intent from a human perspective.

Generation Diversity. In addition to producing high-quality single results, the **SCORE** framework benefits from the inherent stochasticity of the diffusion process to generate diverse layouts for the same text prompt. As shown in Figure 6, by using different random seeds, our method can explore multiple, equally plausible, and creative collage results in response to the prompt "a crown". This capability is crucial for creative applications, as it provides users with a range of options rather than a single, deterministic output.

Ablation Study

To validate the contribution of each term in our joint loss, we conduct an ablation by removing the render losses $\mathcal{L}_{\text{geometry}}$ and $\mathcal{L}_{\text{overlap}}$ in turn and comparing against the full model.

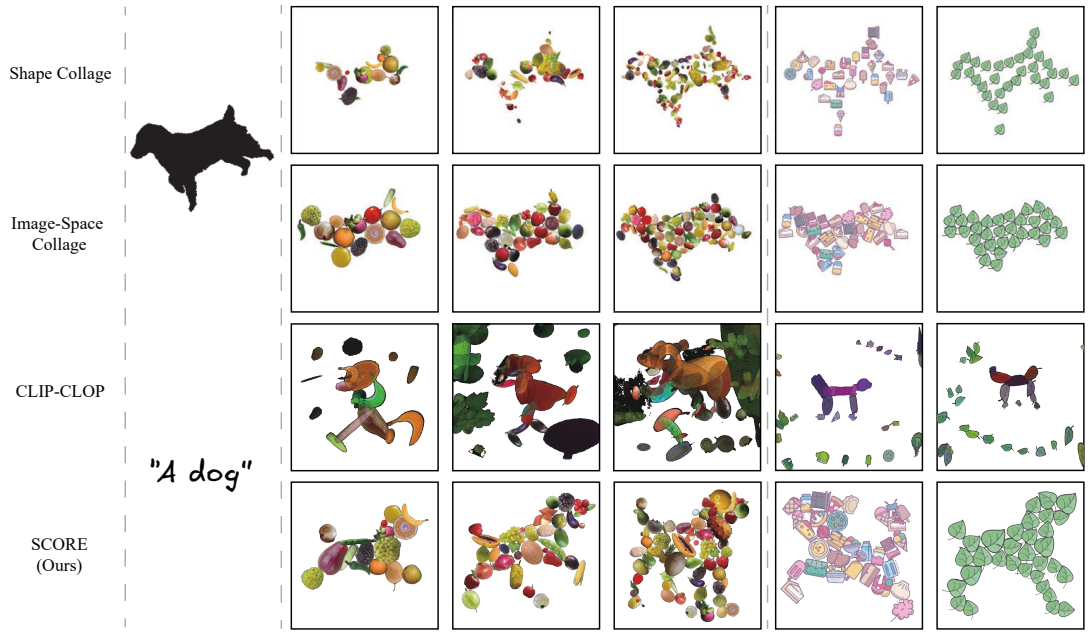

Figure 5: Qualitative comparison between baselines and our method on different numbers of elements and different themes.

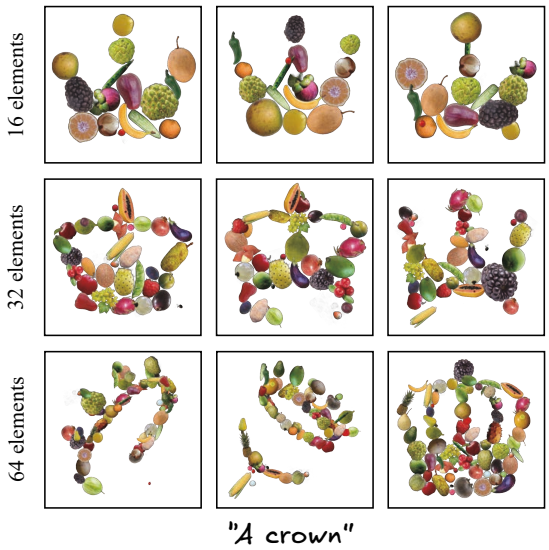

Figure 6: Results generated by our method with the same prompt. As the number of elements used increases, the diversity of the results also increases.

Without $\mathcal{L}_{\text{geometry}}$, some elements are driven off-canvas or shrunk toward a small scale rather than being repositioned for semantic alignment, which is not desirable in collage works. Related metrics also declined. Without $\mathcal{L}_{\text{overlap}}$, results can still be generated, but the Overlap Ratio in Table 2 becomes large, and substantial occlusion degrades perceptual clarity. Adding $\mathcal{L}_{\text{overlap}}$ may slightly reduce other metrics in some cases, yet it markedly suppresses excessive overlap and yields more readable collages.

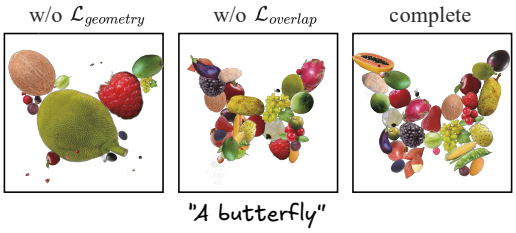

Figure 7: Ablation results generated by our method.

Method	CLIP-RGB↑	CLIP-Contour↑	Aesthetic↑	Overlap
w/o $\mathcal{L}_{\text{geometry}}$	22.1484	26.0668	5.3421	0.0564
w/o $\mathcal{L}_{\text{overlap}}$	**23.2130**	**27.7504**	5.3640	0.4088
Ours (SCORE)	22.1625	27.1836	**5.6324**	0.0797

Table 2: Ablation study on our key $\mathcal{L}_{\text{render}}$ components.

Conclusion & Discussion

We introduce **SCORE**, a text-driven collage framework that optimizes directly in image space. By coupling Variational Score Distillation (VSD) with structural regularizers, it yields collages that align semantically, remain structurally coherent, and preserve source appearance. Minimal editing and parametric output further enhance visual quality and post-editing flexibility. **SCORE** is sensitive to the pixel-space diffusion prior and relies on accurate foreground segmentation, especially for real images. Although it captures contours well, it struggles when objects have simple outlines but rich internal texture. Future efforts may focus on improving computational efficiency to support real-time applications and extending the framework from 2D collages to the structured arrangement of 3D objects.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback and the user study participants for their time. This work was partially supported by grants from NSFC (62472287), Guangdong Basic and Applied Basic Research Foundation (2023A1515011297), and Shenzhen Natural Science Foundation (JCYJ20250604181519025).

References

Feng, W.; Zhu, W.; Fu, T.-j.; Jampani, V.; Akula, A.; He, X.; Basu, S.; Wang, X. E.; and Wang, W. Y. 2023. Layoutgpt: Compositional visual planning and generation with large language models. *Advances in Neural Information Processing Systems*, 36: 18225–18250.

Gatis, D.; et al. 2020. Rembg. https://github.com/danielgatis/rembg. Accessed: 2025-05-30.

Goferman, S.; Tal, A.; and Zelnik-Manor, L. 2010. Puzzle-like collage. In *Computer graphics forum*, volume 29, 459–468. Wiley Online Library.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2): 3.

Huang, H.; Zhang, L.; and Zhang, H.-C. 2011. Arcimboldo-like collage using internet images. In *Proceedings of the 2011 SIGGRAPH Asia Conference*, 1–8.

Huang, K.; Sun, K.; Xie, E.; Li, Z.; and Liu, X. 2023. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *Advances in Neural Information Processing Systems*, 36: 78723–78747.

Itoh, T.; Yamaguchi, Y.; Ikehata, Y.; and Kajinaga, Y. 2004. Hierarchical data visualization using a fast rectangle-packing algorithm. *IEEE Transactions on Visualization and Computer Graphics*, 10(3): 302–313.

Kim, J.; Pellacini, F.; et al. 2002. Jigsaw image mosaics. *ACM Transactions on Graphics*, 21(3): 657–664.

Kwan, K. C.; Sinn, L. T.; Han, C.; Wong, T.-T.; and Fu, C.-W. 2016. Pyramid of arclength descriptor for generating collage of shapes. *ACM Trans. Graph.*, 35(6): 229–1.

Lee, G.; Kim, M.; Lee, Y.; Lee, M.; and Zhang, B.-T. 2023. Neural collage transfer: Artistic reconstruction via material manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2394–2405.

Li, T.-M.; Lukáč, M.; Gharbi, M.; and Ragan-Kelley, J. 2020. Differentiable vector graphics rasterization for editing and learning. *ACM Transactions on Graphics (TOG)*, 39(6): 1–15.

Li, Y.; Liu, H.; Wu, Q.; Mu, F.; Yang, J.; Gao, J.; Li, C.; and Lee, Y. J. 2023. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 22511–22521.

Liu, L.; Zhang, H.; Jing, G.; Guo, Y.; Chen, Z.; and Wang, W. 2017. Correlation-preserving photo collage. *IEEE transactions on visualization and computer graphics*, 24(6): 1956–1968.

Mirowski, P.; Banarse, D.; Malinowski, M.; Osindero, S.; and Fernando, C. 2022. CLIP-CLOP: CLIP-Guided Collage and Photomontage. arXiv:2205.03146.

Pavić, D.; Ceumern, U.; and Kobbelt, L. 2009. GIzMOs: Genuine image mosaics with adaptive tiling. In *Computer Graphics Forum*, volume 28, 2244–2254. Wiley Online Library.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Rother, C.; Bordeaux, L.; Hamadi, Y.; and Blake, A. 2006. Autocollage. *ACM transactions on graphics (TOG)*, 25(3): 847–852.

Saputra, R. A.; Kaplan, C. S.; and Asente, P. 2018. RepulsionPak: Deformation-driven element packing with repulsion forces. In *Proceedings of the 44th Graphics Interface Conference*, 10–17.

Saputra, R. A.; Kaplan, C. S.; and Asente, P. 2019. Improved deformation-driven element packing with repulsionpak. *IEEE transactions on visualization and computer graphics*, 27(4): 2396–2408.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35: 25278–25294.

ShapeCollage. 2025. ShapeCollage. http://www.shapecollage.com. Accessed: 2025-05-08.

Shonenkov, A.; Konstantinov, M.; Bakshandaeva, D.; Schuhmann, C.; Ivanova, K.; and Klokova, N. 2023. If by deepfloyd lab at stabilityai. https://github.com/deep-floyd/IF. Accessed: 2025-05-30.

Sikora, T. 2002. The MPEG-7 visual standard for content description-an overview. *IEEE Transactions on circuits and systems for video technology*, 11(6): 696–702.

Spielmann, Y. 1999. Aesthetic features in digital imaging: collage and morph. *Wide Angle*, 21(1): 131–148.

Srivastava, D.; Zhang, X.; Wen, H.; Wen, C.; and Tu, Z. 2025. Lay-Your-Scene: Natural Scene Layout Generation with Diffusion Transformers. *arXiv preprint arXiv:2505.04718*.

Tang, Z.; Wu, C.; Li, J.; and Duan, N. 2023. Layoutnuwa: Revealing the hidden layout expertise of large language models. *arXiv preprint arXiv:2309.09506*.

Wang, W.; Wang, H.; Dai, G.; and Wang, H. 2006. Visualization of large hierarchical data by circle packing. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, 517–520.

Wang, Y.; Chu, X.; Zhang, K.; Bao, C.; Li, X.; Zhang, J.; Fu, C.-W.; Hurter, C.; Deussen, O.; and Lee, B. 2019. Shapewordle: tailoring wordles using shape-aware archimedean spirals. *IEEE Transactions on Visualization and Computer Graphics*, 26(1): 991–1000.

Wang, Z.; Lu, C.; Wang, Y.; Bao, F.; Li, C.; Su, H.; and Zhu, J. 2023. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in neural information processing systems*, 36: 8406–8441.

Wang, Z.; and Lu, M. 2025. Image-Space Collage and Packing with Differentiable Rendering. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, 1–11.

Xu, J.; and Kaplan, C. S. 2007. Calligraphic packing. In *Proceedings of Graphics Interface 2007*, GI '07, 43–50. New York, NY, USA: Association for Computing Machinery. ISBN 9781568813370.

Xu, P.; Ding, J.; Zhang, H.; and Huang, H. 2019. Discernible image mosaic with edge-aware adaptive tiles. *Computational Visual Media*, 5(1): 45–58.

Yao, S.-Y.; Wu, D.-Y.; Lee, T.-Y.; et al. 2025. Shape Cloud Collage on Irregular Canvas. *IEEE Transactions on Visualization and Computer Graphics*.

Yu, Z.; Lu, L.; Guo, Y.; Fan, R.; Liu, M.; and Wang, W. 2014. Content-Aware Photo Collage Using Circle Packing. *IEEE Transactions on Visualization and Computer Graphics*, 20(2): 182–195.

Zhang, J.; Guo, J.; Sun, S.; Lou, J.-G.; and Zhang, D. 2023. Layoutdiffusion: Improving graphic layout generation by discrete diffusion probabilistic models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7226–7236.

Zhang, L.; Ma, K.-L.; and Yu, J. 2016. Adaptively tiled image mosaics utilizing measures of color and region entropy. In *Proceedings of the 9th International Symposium on Visual Information Communication and Interaction*, 122–129.

Zou, C.; Cao, J.; Ranaweera, W.; Alhashim, I.; Tan, P.; Sheffer, A.; and Zhang, H. 2016. Legible compact calligrams. *ACM Transactions on Graphics (TOG)*, 35(4): 1–12.