# Stroke of Surprise: Progressive Semantic Illusions in Vector Sketching

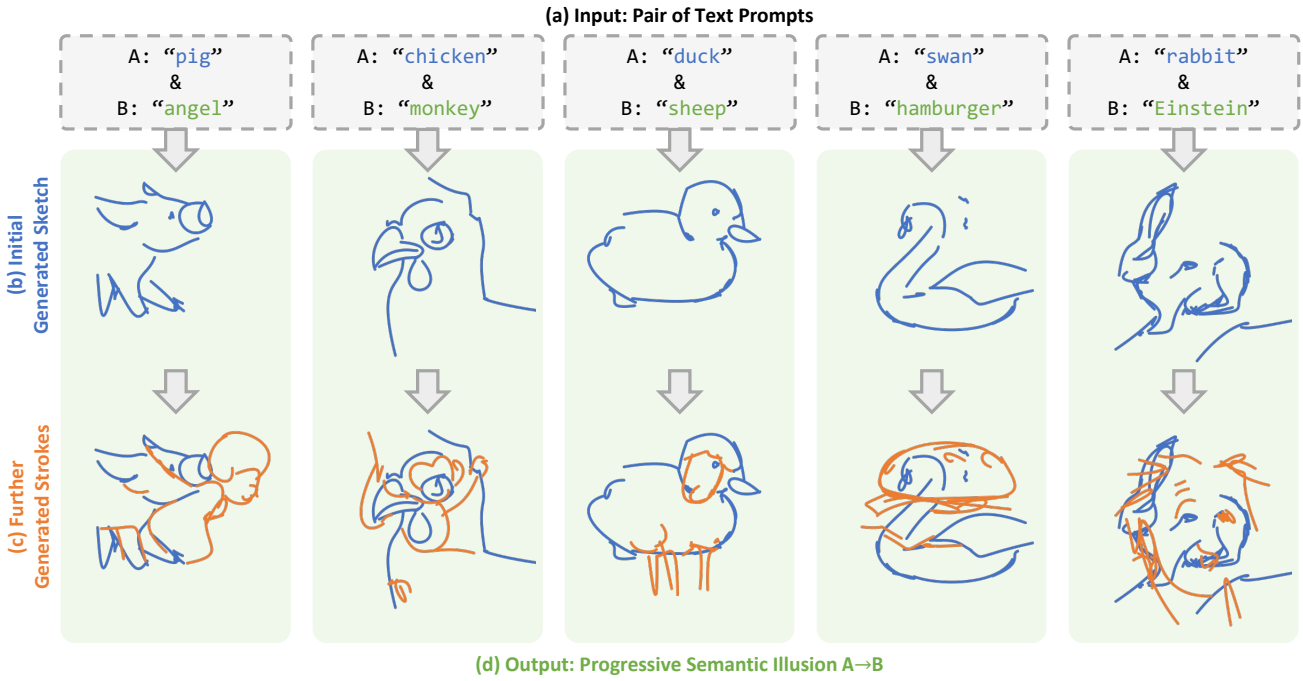



Figure 1. **Progressive semantic illusions from text.** Given a pair of text prompts (a), our method generates a vector sketch that evolves over time. The initial generated sketch (b) depicts the first concept (e.g., "pig"). By adding further generated strokes (c), the drawing is transformed into a totally different object (e.g., "angel"). This creates a *Stroke of Surprise*: the process **subverts the viewer's expectation** of the initial concept, triggering a dramatic **semantic reversal** as the final strokes re-contextualize the entire composition.


## Abstract

*Visual illusions traditionally rely on spatial manipulations such as multi-view consistency. In this work, we introduce Progressive Semantic Illusions, a novel vector sketching task where a single sketch undergoes a dramatic semantic transformation through the sequential addition of strokes. We present Stroke of Surprise, a generative framework that optimizes vector strokes to satisfy distinct semantic interpretations at different drawing stages. The core challenge lies in the "dual-constraint": initial prefix strokes must form a coherent object (e.g., a duck) while simultaneously serving as the structural foundation for a second concept (e.g., a sheep) upon adding delta strokes. To address this, we propose a sequence-aware joint optimization framework driven by a dual-branch Score Distillation Sampling (SDS) mechanism. Unlike sequential approaches that freeze the initial state, our method dynamically adjusts prefix strokes to discover a "common structural subspace" valid for both targets. Furthermore, we introduce a novel Overlay Loss that enforces spatial complementarity, ensuring structural integration rather than occlusion. Extensive experiments demonstrate that our method significantly outperforms state-of-the-art baselines in recognizability and illusion strength, successfully expanding visual anagrams from the spatial to the temporal dimension. Project page: https://stroke-of-surprise.github.io/*

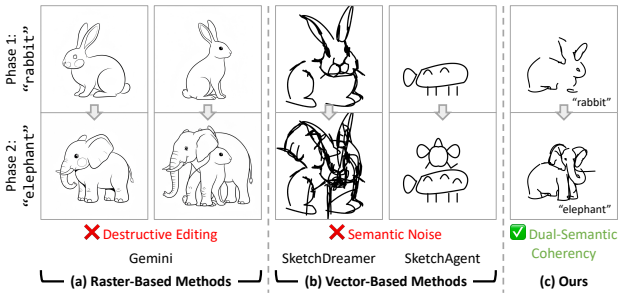



Figure 2. **Challenges in progressive illusion sketching.** (a) Raster-based methods (e.g., Nano Banana Pro) rely on **destructive editing**, modifying the initial structure to fit the final target and thus violating the progressive constraint. (b) Vector-based baselines (e.g., SketchDreamer [93] or SketchAgent [110]) employ a greedy strategy, where specific Phase 1 details become **semantic noise** or clutter in Phase 2. (c) Ours achieves **dual-semantic coherency** by jointly optimizing for a common structural subspace, ensuring the initial strokes are valid building blocks for both interpretations (e.g., "`rabbit`" → "`elephant`").

## 1. Introduction

Visual illusions traditionally exploit spatial ambiguities, requiring viewers to change viewpoints to perceive hidden meanings (e.g., "Visual Anagrams" [36]). In this work, we introduce a new dimension to this artistic interplay: *time*. We propose **Progressive Semantic Illusions**, a novel vector sketching task where the drawing process itself drives semantic transformation. As illustrated in Fig. 1, our method generates a coherent initial sketch (e.g., "a pig") that is subsequently re-contextualized by additional strokes into a distinct concept (e.g., "an angel"). This "Stroke of Surprise" subverts expectations, achieving a perceptual shift through sequential stroke accumulation rather than spatial manipulation.

Sketch generation has evolved from category-specific RNNs [39] to open-vocabulary models leveraging CLIP [94] and diffusion priors [97]. Methods like CLIPasso [108] and VectorFusion [52] utilize differentiable rasterization for high-fidelity sketching, while sequential approaches like SketchAgent [110] and SketchDreamer [93] mimic step-by-step human drawing. Regarding illusions, Visual Anagrams [36] and ShadowDraw [79] explore multi-view effects via diffusion. However, these prior works focus on static pixel representations or spatial rearrangements, leaving the challenge of *temporal* semantic transformation in vector graphics unexplored.

Generating progressive illusions presents a unique "Dual-Constraint": early strokes must depict object "A" while simultaneously functioning as the structural foundation for object "B" (Fig. 2). Existing methods fail to address this additive nature. Raster-based models (e.g., Gemini) rely on *destructive editing*, overwriting initial pixels and violating the progressive constraint. Conversely, sequential vector models (e.g., SketchAgent) employ a *greedy strategy*, optimizing strokes solely for "A". This renders the fixed prefix as *semantic noise* when extending to "B", resulting in clutter. Crucially, these baselines fail to find a **"Common Subspace"**, which is a shared geometric configuration valid for both semantic interpretations.

To overcome these limitations, we propose **Stroke of Surprise**, a sequence-aware joint optimization framework designed to discover this common structural subspace (Fig. 3). Unlike sequential approaches, we optimize parameters for both the prefix (Object "A") and full phase (Object "B") simultaneously using a dual-branch Score Distillation Sampling (SDS) mechanism. This guidance ensures prefix strokes are recognizable as the initial concept yet "primed" for re-interpretation. Furthermore, we introduce a geometric *Overlay Loss* to enforce spatial complementarity and prevent occlusion. This enables delta strokes to structurally integrate with and re-contextualize the prefix. For example, it can transform pig ears into angel wings, creating a seamless illusion.

Our main contributions are summarized as follows:

- **Task:** We introduce **Progressive Semantic Illusion**, extending visual illusions from the spatial to the *temporal* dimension. This task requires a single vector sketch to reveal distinct semantic interpretations through progressive stroke accumulation.
- **Method:** We propose a **sequence-aware joint optimization framework** that optimizes shared stroke parameters under simultaneous semantic constraints and a novel Overlay Loss, ensuring the prefix strokes form a robust structural foundation for the final illusion.
- **Insight & Scalability**: We demonstrate that joint optimization identifies a "**Common Subspace**" that resolves conflicts between early semantic clarity and later structural integration. Our method scales to multi-phase illusions ("A" → "B" → "C") and significantly outperforms baselines in recognizability and coherence.

## 2. Related Work

**Sketch Generation and Sequential Modeling.** Vector sketch synthesis evolved from *category-specific* to *open-vocabulary* generation, progressing from edge detection [16, 118] and sketch datasets [29, 39, 54, 99] through RNNs [39], Transformers [17, 70, 96], GANs [34, 71], and autoregressive models [120, 130]. CLIP enabled text-driven synthesis [32, 108, 109], while diffusion-based Score Distillation [90] was adapted for SVG [52, 93, 106, 125, 127, 136, 138], with recent feed-forward [4, 22] and LLM-based methods [89, 121, 128]. Stroke ordering encodes semantics, as shown in foundational rendering [41, 42] and sequential modeling via attention [37], VAEs [39], embeddings [1], completion [71, 104], transformers [11], Bézier curves [25], temporality [55], and diffusion [114]. Stroke semantics are

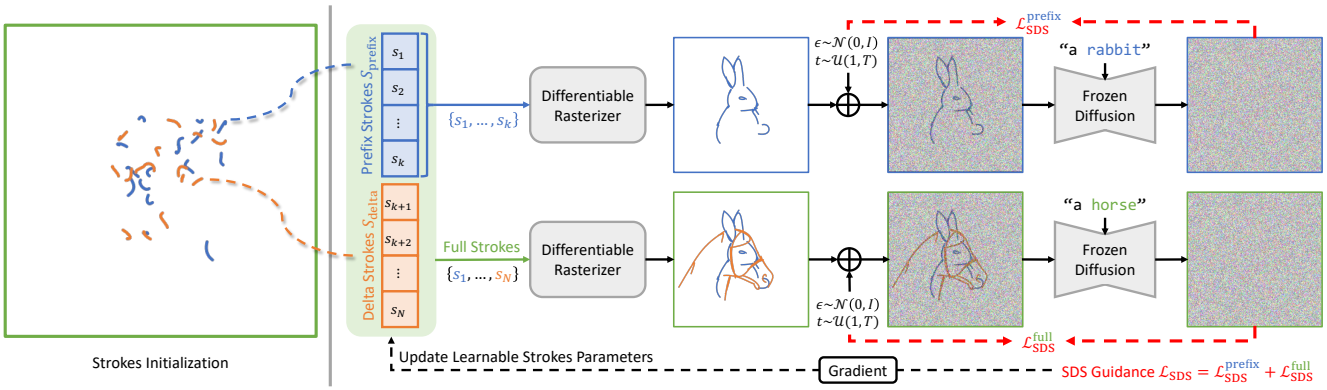



Figure 3. **Pipeline overview.** Our method optimizes a set of learnable stroke parameters, which are divided into prefix strokes $S_{\text{prefix}}$ and delta strokes $S_{\text{delta}}$. The optimization process involves two parallel branches. In the top branch, only the prefix strokes are rendered by a differentiable rasterizer to create a partial sketch (e.g., a rabbit). This sketch is then guided by a pre-trained, frozen text-to-image diffusion model using a prompt corresponding to the prefix ("`a rabbit`"), resulting in the prefix SDS loss $\mathcal{L}_{\text{SDS}}^{\text{prefix}}$. In the bottom branch, the full set of strokes is rendered to create the complete sketch (e.g., a horse). This is guided by the same diffusion model using a prompt for the full object ("`a horse`"), resulting in the full SDS loss $\mathcal{L}_{\text{SDS}}^{\text{full}}$. The total SDS guidance loss is the sum of these two terms $\mathcal{L}_{\text{SDS}} = \mathcal{L}_{\text{SDS}}^{\text{prefix}} + \mathcal{L}_{\text{SDS}}^{\text{full}}$. Gradients from this total loss are backpropagated to update all learnable stroke parameters.

explored through RL [49], optimal transport [139], feed-forward [73], canvas-aware [47], style [101], recognizability [10], and hierarchical methods [135]. Human-AI collaboration includes co-creative [26], turn-taking [86], creativity [57], synchronous [64], and LLM systems [50, 110]. All prior methods target a *single* semantic goal. We introduce *dual-constraint* optimization where prefix strokes forming Object A are *re-contextualized* into Object B.

**Sketch Perception and Visual Illusions.** Gestalt principles [111, 117], Recognition-by-Components [13, 14], illusory contours [56], and cognitive studies [18, 30] explain how sparse strokes evoke recognition. Computational approaches surpassed human benchmarks [29, 134] via RL [84, 85], primitives [2], implicit [7], dynamic [61], grouping [66], saliency [12], graphs [132], transformers [113], explainability [92], and neuroscience [103]. Visual illusions encode multiple meanings via *spatial* transformations: frequency [87], shadows [83], wire art [46], camouflage [24], evolution [81], and morphing [3, 8, 100]. Diffusion expanded this through view averaging [36], frequency decomposition [35], multi-task [131], phase transfer [33], 3D [31], fabrication [15], ambigrams [137], anamorphic [19, 28], neural shadows [112], cross-modal [23], sculpture [91, 119], viewpoint-dependent [59], biases [88], and predictive coding [116]. These rely on *spatial* transformations. We introduce *temporal* concealment where semantics emerge through stroke accumulation.

**Differentiable Rendering and Score Distillation.** Bézier foundations [9, 27] enabled differentiable rasterization [67], improved by splatting [74], ranking [44], layers [80], implicit [95, 107], latent diffusion [126], interpolation [77], simplification [102], typography [51], discrete stylization [48], and transformation [123]. Score Distillation [45, 90] was adapted for vectors [52, 125] and editing [40], with improvements via coarse-to-fine [69], variational [115], noise-free [58], interval [68], DDIM [78], bridge [82], collaborative [60], likelihood score matching [20], and posterior [62] formulations. Multi-concept methods use composition [72], cross-attention [63], neurons [76], attention [21], subject-driven [98], decomposition [5], fusion [38], training-free [124], and guidance [53]. Multi-view leverages correspondence [105], joint modeling [75], hybrid [6], satellite-to-ground synthesis [65], and Pareto [133]. These focus on *spatial* composition. Our dual-branch SDS addresses *temporal* revelation: prefix strokes receive gradients from both targets, discovering configurations valid for two interpretations.

## 3. Method

Progressive illusions require prefix strokes to depict an initial object while forming the structural basis for a final one. We propose a joint optimization framework via multi-branch Score Distillation Sampling to discover a common structural subspace valid for both interpretations. Prefix strokes receive simultaneous gradients to satisfy dual roles, while an overlay loss enforces spatial separation, ensuring structural integration rather than occlusion.

### 3.1. Progressive Semantic illusion in Vector Form

We partition a set of learnable Bézier strokes $S$ into disjoint subsets: prefix $S_{\text{prefix}} = \{s_1, \ldots, s_k\}$ and delta $S_{\text{delta}} = \{s_{k+1}, \ldots, s_N\}$. The progressive illusion requires $S_{\text{prefix}}$ to

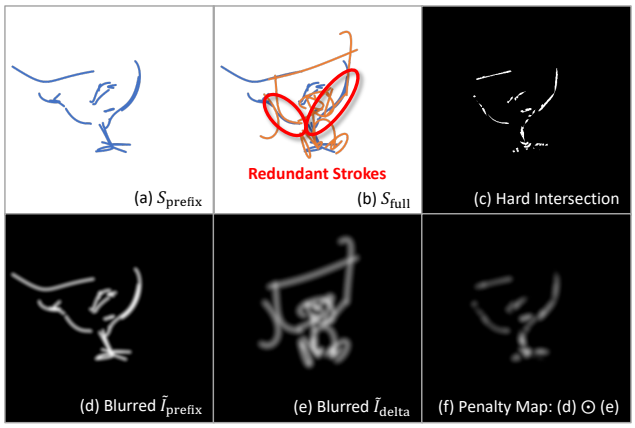



Figure 4. **Motivation and formulation of the overlay loss. (*Top*) Motivation**: Without constraints, redundant strokes (b) occlude the prefix. Hard intersection (c) allows strokes to be placed arbitrarily close, causing crowding. **(*Bottom*) Formulation**: We compute a **soft overlay loss** (f) from blurred maps (d, e). The blur expands the penalty region to create a **spatial buffer**, forcing new strokes to **maintain sufficient distance** from the prefix to ensure visual clarity and separation.

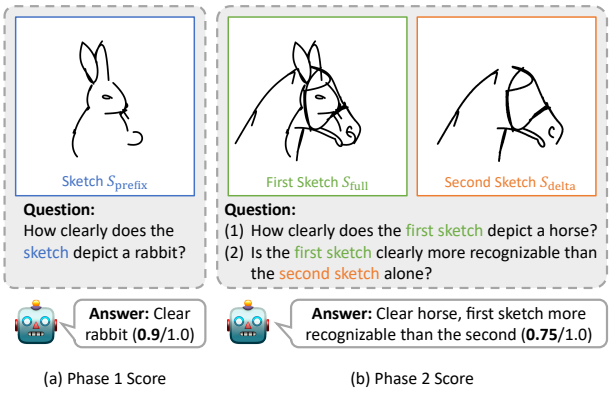



Figure 5. **VLM-based evaluation and ranking pipeline.** We employ GPT-4o to assess the quality of illusion sketches. **(a) For Phase 1**, the model evaluates the recognizability of the prefix sketch ($S_{\text{prefix}}$). **(b) For Phase 2**, the model evaluates the full sketch ($S_{\text{full}}$) while simultaneously comparing it against the delta strokes ($S_{\text{delta}}$). This comparison ensures that the prefix strokes provide **essential structural scaffolding** for the second concept, rather than being merely overwritten. High scores are awarded only when $S_{\text{full}}$ is significantly more recognizable than $S_{\text{delta}}$ alone.

depict the initial concept $p_1$, while the full sketch $S_{\text{full}} = S$ depicts the target $p_2$, achieved by delta strokes recontextualizing the prefix. We optimize stroke parameters $\theta$ such that the rasterized outputs $\mathcal{R}(S_{\text{prefix}}; \theta)$ and $\mathcal{R}(S_{\text{full}}; \theta)$ align with $p_1$ and $p_2$, respectively. The core challenge lies in discovering configurations where prefix strokes meaningfully serve both semantic interpretations.

### 3.2. Joint Optimization Pipeline

We employ a dual-branch strategy to simultaneously refine both stroke subsets (Fig. 3). Unlike sequential methods, our pipeline coordinates semantic objectives via parallel guidance on shared learnable parameters $\theta$. We initialize $N$ strokes near the canvas center, partitioning them into $S_{\text{prefix}}$ (first $k$) and $S_{\text{delta}}$ (remaining). At each iteration, the prefix branch renders $I_{\text{prefix}} = \mathcal{R}(S_{\text{prefix}}; \theta)$. We apply the gradient of the Score Distillation Sampling (SDS) loss conditioned on $p_1$:

$$\nabla_\theta \mathcal{L}_{\text{SDS}}^{\text{prefix}} = \left[ w(t) \left( \epsilon_\phi(z_t, t, p_1) - \epsilon \right) \frac{\partial z_t}{\partial \theta} \right], \tag{1}$$

where $z_t$ is the noised latent, $\epsilon_\phi$ the noise predictor, and $w(t)$ a weighting function.

Simultaneously, the full branch renders $I_{\text{full}} = \mathcal{R}(S_{\text{full}}; \theta)$ conditioned on $p_2$, yielding $\nabla_\theta \mathcal{L}_{\text{SDS}}^{\text{full}}$. We combine these gradients as

$$\nabla_\theta \mathcal{L}_{\text{SDS}} = \nabla_\theta \mathcal{L}_{\text{SDS}}^{\text{prefix}} + \nabla_\theta \mathcal{L}_{\text{SDS}}^{\text{full}}. \tag{2}$$

This ensures prefix strokes receive simultaneous gradients from both targets, satisfying dual roles, while delta strokes optimize to complement them. To prevent delta strokes from merely occluding the prefix, which is a common issue with pure semantic guidance, we introduce an *overlay loss* that penalizes spatial overlap to enforce structural integration.

### 3.3. Overlay Loss for Spatial Coordination

Semantic guidance alone fails to prevent spatial redundancy, often causing delta strokes to clutter prefix strokes (Fig. 4(b)). We introduce an *overlay loss* to enforce spatial complementarity. We render stroke subsets separately and apply Gaussian blur $G_\sigma$ to create soft spatial buffers ($\tilde{I}_{\text{prefix}}, \tilde{I}_{\text{delta}}$), as shown in Fig. 4(d,e). We then compute the normalized overlap:

$$\mathcal{L}_{\text{overlay}} = \frac{2\langle \tilde{I}_{\text{prefix}}, \tilde{I}_{\text{delta}} \rangle}{\|\tilde{I}_{\text{prefix}}\|_1 + \|\tilde{I}_{\text{delta}}\|_1}, \tag{3}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product over pixel space.

This constraint promotes structural integration and smoother semantic transitions, ensuring prefix strokes serve as essential components rather than being obscured. The final objective is:

$$\mathcal{L} = \mathcal{L}_{\text{SDS}} + \lambda_{\text{overlay}} \mathcal{L}_{\text{overlay}}, \tag{4}$$

where $\lambda_{\text{overlay}}$ weights the penalty. Gradients are backpropagated via differentiable rasterization.

### 3.4. Filtering and Ranking

To ensure quality, our systematic pipeline selects the best candidates using VLM assessment and quantitative metrics.

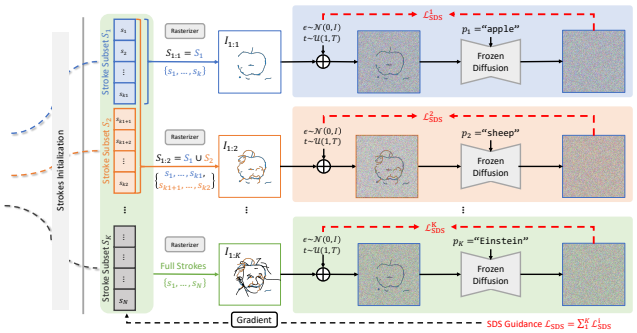



Figure 6. **Multi-phase pipeline.** We scale to $K$ phases (e.g., Apple→Sheep→Einstein) using cumulative stroke subsets $(S_1, \ldots, S_K)$. Parallel branches optimize each cumulative sketch $I_{1:i}$ against prompt $p_i$. **Joint optimization** ensures early strokes receive gradients from all subsequent losses ($\sum \mathcal{L}^i_{\text{SDS}}$), creating a structure primed for the entire evolutionary sequence.

**VLM-based Quality Assessment.** We employ GPT-4o to evaluate four dimensions (Fig. 5). *Phase recognizability* and *Single-object integrity* ensure semantic accuracy and coherence. *Illusion quality* validates the prefix's structural contribution by confirming $S_{\text{full}}$ is significantly more recognizable than $S_{\text{delta}}$ alone. *Sketch quality* penalizes visual clutter. Each phase receives individual scores across these dimensions, and candidates failing minimum thresholds are filtered.

**Ranking Strategies.** GPT-based ranking (Fig. 18) favors semantic accuracy: $\mathcal{R}_{\text{GPT}} = \text{Score}_{\text{Phase 1}} \cdot \text{Score}_{\text{Phase 2}}$. Metric-based ranking (Fig. 17) emphasizes perceptual contrast [79] by penalizing independent delta stroke quality:

$$S_{\text{CLIP}} = (\text{CLIP}_{\text{p1}} \cdot \text{CLIP}_{\text{p2}})/\text{CLIP}^2_{\text{delta}}, \tag{5}$$

$$S_{\text{IR}} = \Phi(\text{IR}_{\text{p1}})^2 + \Phi(\text{IR}_{\text{p2}})^2 - \Phi(\text{IR}_{\text{delta}})^2, \tag{6}$$

$$S_{\text{HPS}} = \text{HPS}^2_{\text{p1}} + \text{HPS}^2_{\text{p2}} - \text{HPS}^2_{\text{delta}}, \tag{7}$$

where $\Phi(\cdot)$ is the standard Gaussian CDF. The final score $\mathcal{R} = S_{\text{CLIP}} \cdot S_{\text{IR}} \cdot S_{\text{HPS}}$ ensures the prefix provides substantial structural contribution.

### 3.5. Extension to Multi-Phase Illusions

Our framework naturally scales to $K$-phase illusions $(A_1, \ldots, A_K)$ by partitioning strokes into disjoint subsets $S_1, \ldots, S_K$. Each cumulative prefix $S_{1:i} = \bigcup_{j=1}^{i} S_j$ renders concept $A_i$. We employ parallel branches (Fig. 6) to jointly optimize all parameters, rendering $I_{1:i}$ conditioned on prompt $p_i$. This ensures early strokes (e.g., $S_1$) receive gradients from all subsequent branches, coordinating cumulative interpretations. We extend the overlay loss to penalize overlap between $S_{1:i}$ and the next subset $S_{i+1}$:

$$\mathcal{L} = \sum_{i=1}^{K} \mathcal{L}^i_{\text{SDS}} + \sum_{i=1}^{K-1} \lambda^i_{\text{overlay}} \mathcal{L}^i_{\text{overlay}}. \tag{8}$$

## 4. Experiments

### 4.1. Experimental Setup

**Baseline.** We adapt state-of-the-art methods: Nano Banana Pro (raster), SketchAgent [110], and SketchDreamer [93] (vector). We design two protocols: (1) **Text-to-illusion**: Baselines generate sketches sequentially (prefix from $p_1$, then full from $p_2$). For the raster-based Nano Banana Pro, we enforce the progressive constraint by overlaying the prefix onto the final output, whereas vector baselines natively support stroke addition. (2) **Ours-to-illusion**: We provide our optimized prefix sketches as input to evaluate whether baselines can complete the transformation given an ideal structural foundation.

**Data.** Our evaluation dataset comprises 64 common objects spanning diverse categories. We randomly sample pairs to form $(p_1, p_2)$ combinations, run multiple optimization iterations per pair, then apply filtering and ranking to select top-k results for evaluation.

**Implementation Details** We implement our framework using Stable Diffusion v1.5 for Score Distillation Sampling guidance on an NVIDIA RTX 4090 GPU. We optimize stroke parameters $\theta$ for 2,000 iterations using Adam optimizer with guidance scale 100 and overlay loss weight $\lambda_{\text{overlay}} = 0.1$. Generation requires approximately 13 minutes for two-phase and 15 minutes for three-phase illusions.

**Metrics.** For quantitative evaluation, we employ both standard and specialized metrics to assess illusion quality. We use CLIP score computed as the minimum across all phases to measure semantic alignment. Beyond standard metrics, we define two illusion-specific measures. Structural concealment evaluates whether prefix strokes contribute substantively to the full sketch rather than being occluded by delta strokes. For any metric $M \in \{\text{CLIP}, \text{ImageReward}, \text{HPS}\}$ [43, 122, 129], we compute: $C^M_{\text{struct}} = M_{\text{full}} - M_{\text{delta}}$. Higher scores indicate prefix strokes retain significant structural roles. Semantic concealment measures whether non-current phase semantics are effectively hidden. Following [36], we compute:

$$C_{\text{semantic}} = \text{tr}(\text{softmax}(S/\tau)), \tag{9}$$

where $S$ is the CLIP image-text similarity matrix and $\tau$ is temperature. Higher scores indicate clear phase-specific semantics.

We further conduct two user studies with 143 participants for additional quantitative validation. The first compares our

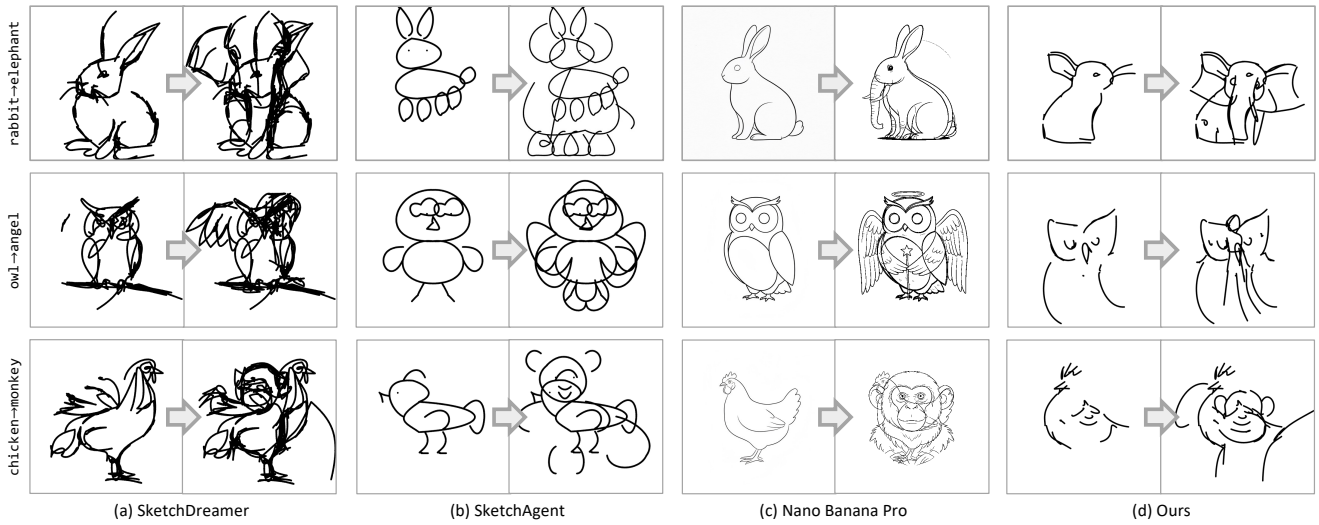



Figure 7. **Qualitative comparisons.** We compare against SketchDreamer [93], SketchAgent [110], and Nano Banana Pro. (a) SketchDreamer produces noisy strokes, causing severe visual clutter. (b) SketchAgent yields overly abstract results with low recognizability. (c) Nano Banana Pro relies on **destructive editing** (e.g., overwriting the pig structure to draw an angel), failing the progressive constraint despite high image quality. (d) Ours generates clean, structurally consistent sketches where prefix strokes are creatively repurposed (e.g., rabbit whiskers becoming elephant ear). We provide additional video progressive illusion results and visualization of the optimization process in an interactive HTML in the supplementary material.

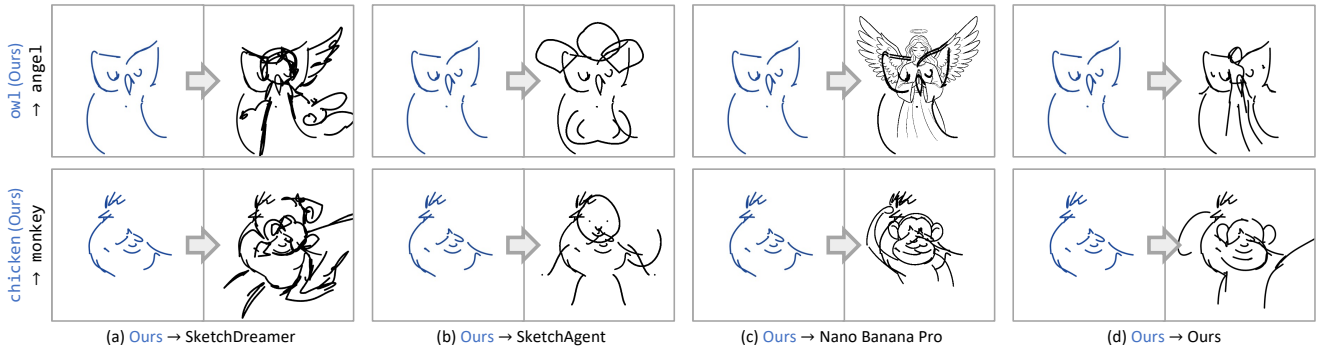



Figure 8. **Phase 2 extension with fixed prefix (ours).** We evaluate how methods extend a fixed Phase 1 sketch generated by our method. Interestingly, baselines produce better Phase 2 results here than in Fig. 7 (where they generate Phase 1 themselves). This indicates that our Phase 1 strokes inherently embed structural cues for the second concept, validating that our joint optimization successfully finds a versatile common subspace. However, comparing (a-c) with (d), our method still achieves the highest success rate and structural consistency, as $S_{\text{delta}}$ is jointly optimized with the prefix rather than sequentially appended.

top-1 result against baselines across five prompt pairs. The second assesses our ranking pipeline by asking participants to select satisfactory results from our top-4 outputs across four prompt pairs, evaluating both technical performance and practical user satisfaction.

## 4.2. Results and Analysis

As shown in Tab. 1(a,c), our method substantially outperforms baselines in CLIP and concealment scores, achieving 100% coverage compared to Nano Banana Pro's 34.9%. Fig. 7 and Fig. 16 highlights baseline limitations: visual clutter (SketchDreamer), oversimplification (SketchAgent), and destructive editing (Nano Banana Pro). Tab. 1(b,c) and Fig. 8 show that in the fixed-prefix setting, baselines achieve improved recognizability, suggesting our prefix strokes embed implicit structural cues ("common subspace"). However, they remain substantially inferior to our method across all metrics, confirming that joint optimization of the complete stroke sequence is essential for seamless integration.

**User Studies.** Our user studies strongly reinforce these findings. In comparisons against baselines, participants se-

Table 1. **Quantitative comparison.** (a) Vector baselines lack quality; Nano Banana fails coverage (∼35%). (b) Extending our Phase 1 helps, but still lags behind (c), validating joint optimization. (c) Ours achieves top metrics with 100% coverage.

| | | Phase 1 | CLIP ↑ | Concealment (structural) | | | $C_{\text{semantic}}$ | Coverage |
|---|---|---|---|---|---|---|---|---|
| | Method | Source | Avg min | CLIP ↑ | IR ↑ | HPS ↑ | CLIP ↑ | (%) ↑ |
| (a) | SketchDreamer | - | 24.803 | -0.393 | 0.338 | 0.011 | 0.887 | **100.0%** |
| | SketchAgent | - | 24.393 | -2.544 | 0.095 | 0.000 | 0.752 | **100.0%** |
| | Nano Banana Pro | - | 26.821 | -2.774 | -0.663 | -0.019 | 0.875 | 34.9% |
| (b) | SketchDreamer | Ours | 28.148 | 0.060 | 0.302 | 0.011 | 0.961 | **100.0%** |
| | SketchAgent | Ours | 24.019 | -2.778 | 0.080 | 0.003 | 0.762 | **100.0%** |
| | Nano Banana Pro | Ours | 28.903 | -1.065 | -0.426 | -0.014 | 0.958 | 35.2% |
| (c) | Ours (GPT-ranking) | – | 29.873 | 1.668 | 0.839 | 0.023 | **0.983** | **100.0%** |
| | Ours (Metric-ranking) | – | **30.044** | **3.282** | **1.237** | **0.029** | 0.980 | **100.0%** |

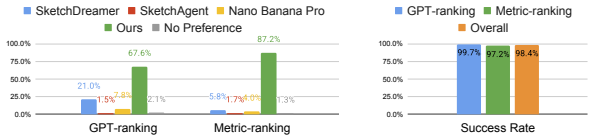



Figure 9. **User study. (Left) Preference**: Participants overwhelmingly favor our method (**green**) over baselines across both ranking strategies. **(Right) Reliability**: A high success rate (¿97%) confirms that our pipeline consistently yields valid illusions, ensuring robustness against the inherent stochasticity of the generation process.

lected our method in 67.7% of GPT-ranking and 87.1% of Metric-ranking cases (Fig. 9(a)). Our ranking pipeline demonstrates strong reliability with over 98% overall satisfaction rates (Fig. 9(b)), thoroughly validating our framework's effectiveness.

### 4.3. Ablation Studies

**Optimization Strategy** We evaluate our joint optimization approach against a sequential alternative that first optimizes prefix strokes for the initial concept, then fixes these and optimizes only delta strokes. As shown in Fig. 10(a), this sequential approach produces rigid structures where features conflict with the final object, causing failed transitions. In contrast, our joint optimization (Fig. 10(b)) updates both stroke sets simultaneously, discovering a common structural subspace where prefix strokes represent the initial concept while integrating naturally into the final representation. This enables improved visual consistency and smooth transitions, confirming that joint optimization is essential for high-quality progressive illusion sketches.

**Stroke Initialization.** Since our objective function is highly non-convex, initialization is critical for convergence. Fig. 11 shows that spatial concentration is paramount; scattered initialization fails to capture essential semantic features. In contrast, both centered and shifted gathered configurations succeed, indicating that local stroke density outweighs absolute position. We therefore adopt centered gathered initialization to balance density with spatial coverage, avoiding

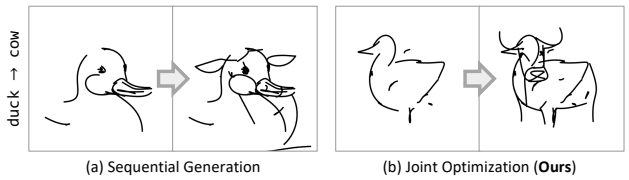



Figure 10. **Ablation on optimization strategy. (a) Sequential generation** yields a rigid Phase 1, creating structural conflicts (e.g., the duck's beak) that fail Phase 2 repurposing. **(b) Joint optimization (Ours)** identifies a **common structural subspace**, yielding a versatile Phase 1 where features serve both interpretations (e.g., the beak doubles as the cow's ear).

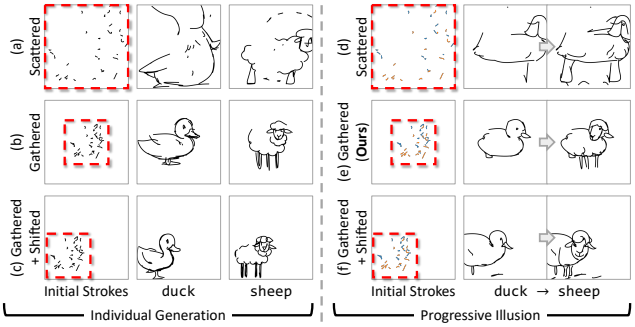



Figure 11. **Ablation on stroke initialization. (a, d) Scattered** fails to aggregate strokes, resulting in disconnected artifacts. **(c, f) Shifted** yields valid sketches, proving that **spatial concentration** is critical for convergence, though it risks boundary cropping. **(b, e) Centered (Ours)** offers the optimal balance, ensuring structural integrity without clipping.

potential boundary clipping.

**Overlay Loss.** We validate the necessity of $\mathcal{L}_{\text{overlay}}$. As shown in Fig. 12(a), without this component, semantic guidance alone fails to prevent spatial redundancy, resulting in delta strokes that clutter the prefix. $\mathcal{L}_{\text{overlay}}$ addresses this by penalizing overlap, enforcing spatial complementarity, and significantly reducing intersection artifacts (Fig. 12(b)). Crucially, this promotes structural coherence: prefix strokes are forced to integrate naturally into the subsequent concept rather than being obscured, confirming that geometric constraints are essential for generating clean progressive illusions.

**Stroke Count.** Optimal stroke budget depends on concept complexity (Fig. 13). Simple transformations (e.g., rabbit-to-horse) succeed with minimal strokes (8–16), whereas complex subjects like Einstein require 32–64 strokes to capture essential details; insufficient budgets compromise recognizability. We therefore adopt a default of 16 prefix strokes and 32 total strokes, robustly balancing structural simplicity with

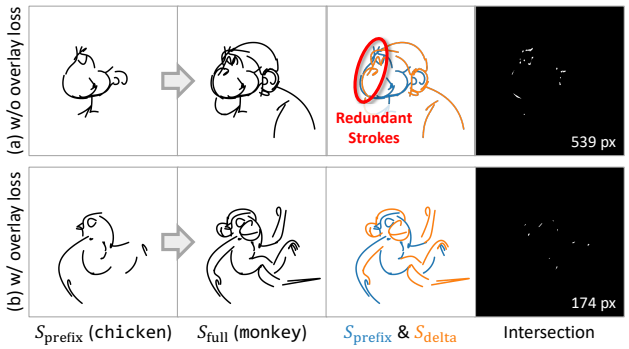



Figure 12. **Ablation of overlay loss ($\mathcal{L}_{\text{overlay}}$).** (a) Without $\mathcal{L}_{\text{overlay}}$, the model generates redundant strokes atop existing ones to satisfy the semantic target, resulting in visual clutter (red circle) and high intersection artifacts. (b) With $\mathcal{L}_{\text{overlay}}$, the generated strokes ($S_{\text{delta}}$) become spatially complementary to the prefix ($S_{\text{prefix}}$), avoiding collisions to produce a clean, coherent line drawing.

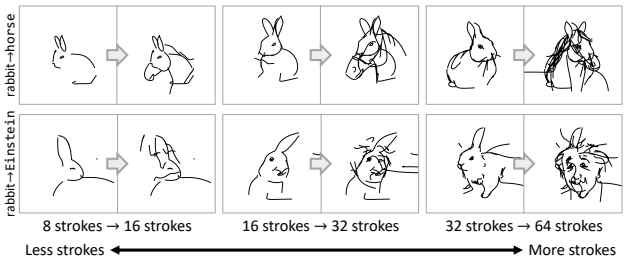



Figure 13. **Analysis of stroke count.** (*Top*) Simple concepts (`horse`) form recognizable silhouettes with minimal strokes (8→16). (*Bottom*) While complex concepts (`Einstein`) require a larger budget (32→64) to capture essential details. Fewer strokes result in abstraction. Our default (16→32) balances structural simplicity and semantic fidelity.

semantic fidelity.

### 4.4. Applications

We demonstrate versatility beyond standard two-phase scenarios. Fig. 14 confirms robustness across diverse concept pairs, ranging from structurally similar to semantically distant. Fig. 15 extends this to three-phase illusions (e.g., apple-to-rabbit-to-pig), showcasing effective multi-target coordination. Furthermore, our framework generalizes to alternative representations, including B-spline curves (Fig. 19), vector graphics (Fig. 20), and colored sketches (Fig. 21), validating the broad applicability of our joint optimization principle.

## 5. Conclusion

We present Stroke of Surprise, the first framework for progressive semantic illusions in vector sketching. By shifting from spatial to temporal dimensions, we enable real-time semantic re-contextualization. Our joint optimization strategy demonstrates that prefix strokes must be ”primed” for future semantics. Greedy baselines do not have this ability. Meanwhile, the Overlay Loss ensures structural integration without obfuscation. Evaluations confirm our results are both semantically accurate and perceptually surprising.

**Limitations.** Our method inherits limitations from pre-trained diffusion priors; weak SDS guidance for complex structures (e.g., “scissors”) causes optimization failure. We provide visual examples in the supplementary material.

## References


[1] Emre Aksan, Thomas Deselaers, Andrea Tagliasacchi, and Otmar Hilliges. Cose: Compositional stroke embeddings. *Advances in Neural Information Processing Systems*, 33: 10041–10052, 2020. 2

[2] Stephan Alaniz, Massimiliano Mancini, Anjan Dutta, Diego Marcos, and Zeynep Akata. Abstracting sketches through simple primitives. In *European Conference on Computer Vision*, pages 396–412. Springer, 2022. 3

[3] Marc Alexa, Daniel Cohen-Or, and David Levin. As-rigid-as-possible shape interpolation. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 165–172. 2023. 3

[4] Ellie Arar, Yarden Frenkel, Daniel Cohen-Or, Ariel Shamir, and Yael Vinker. Swiftsketch: A diffusion model for image-to-vector sketch generation. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, pages 1–12, 2025. 2

[5] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–12, 2023. 3

[6] Sherwin Bahmani, Ivan Skorokhodov, Victor Rong, Gordon Wetzstein, Leonidas Guibas, Peter Wonka, Sergey Tulyakov, Jeong Joon Park, Andrea Tagliasacchi, and David B Lindell. 4d-fy: Text-to-4d generation using hybrid score distillation sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7996–8006, 2024. 3

[7] Hmrishav Bandyopadhyay, Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Aneeshan Sain, Tao Xiang, Timothy Hospedales, and Yi-Zhe Song. Sketchinr: A first look into sketches as implicit neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12565–12574, 2024. 3

[8] Thaddeus Beier and Shawn Neely. Feature-based image metamorphosis. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 529–536. 2023. 3

[9] Pierre E Bézier. How renault uses numerical control for car body design and tooling. Technical report, SAE Technical Paper, 1968. 3

[10] Ayan Kumar Bhunia, Ayan Das, Umar Riaz Muhammad, Yongxin Yang, Timothy M Hospedales, Tao Xiang, Yulia Gryaditskaya, and Yi-Zhe Song. Pixelor: A competitive

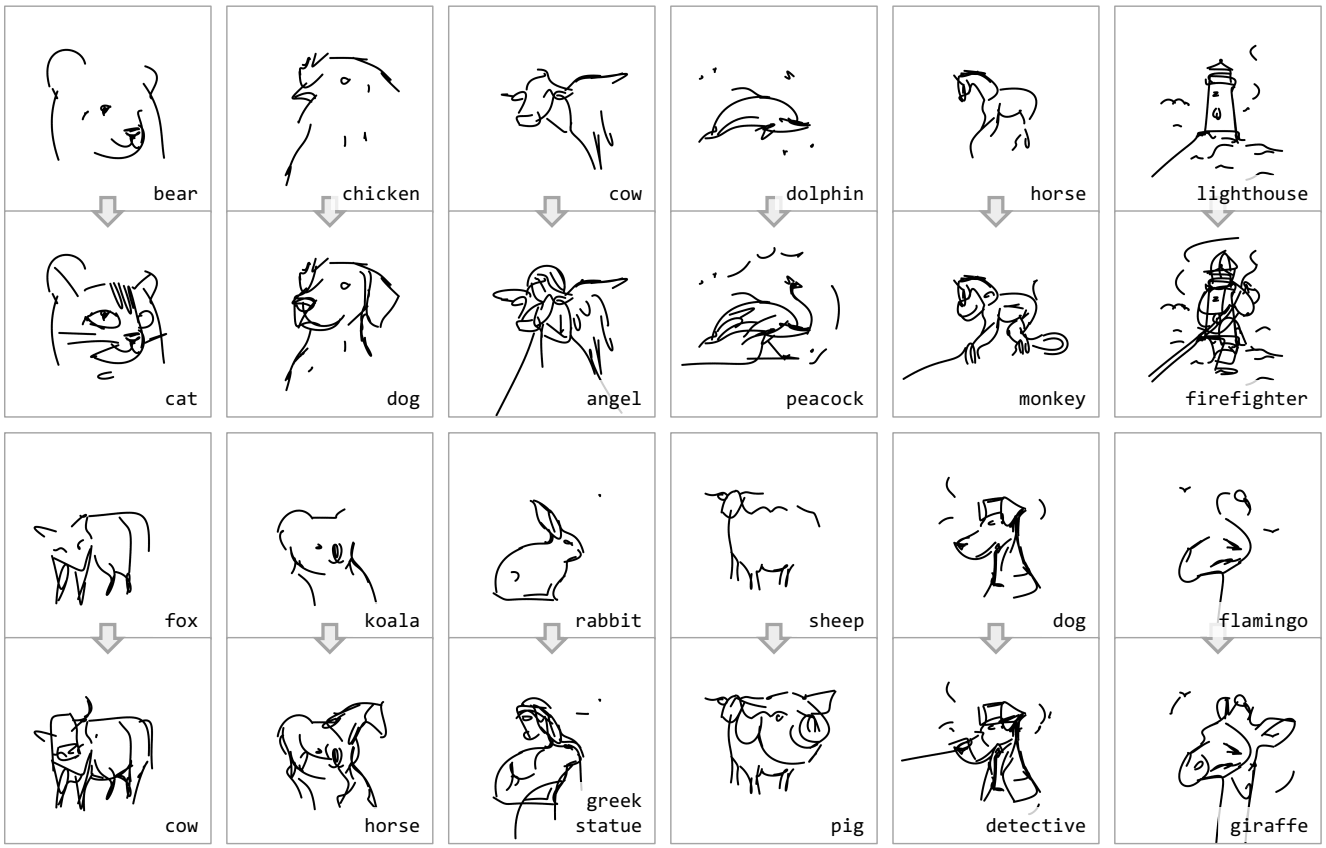



Figure 14. **Additional 2-phase progressive illusion results produced by our method.**

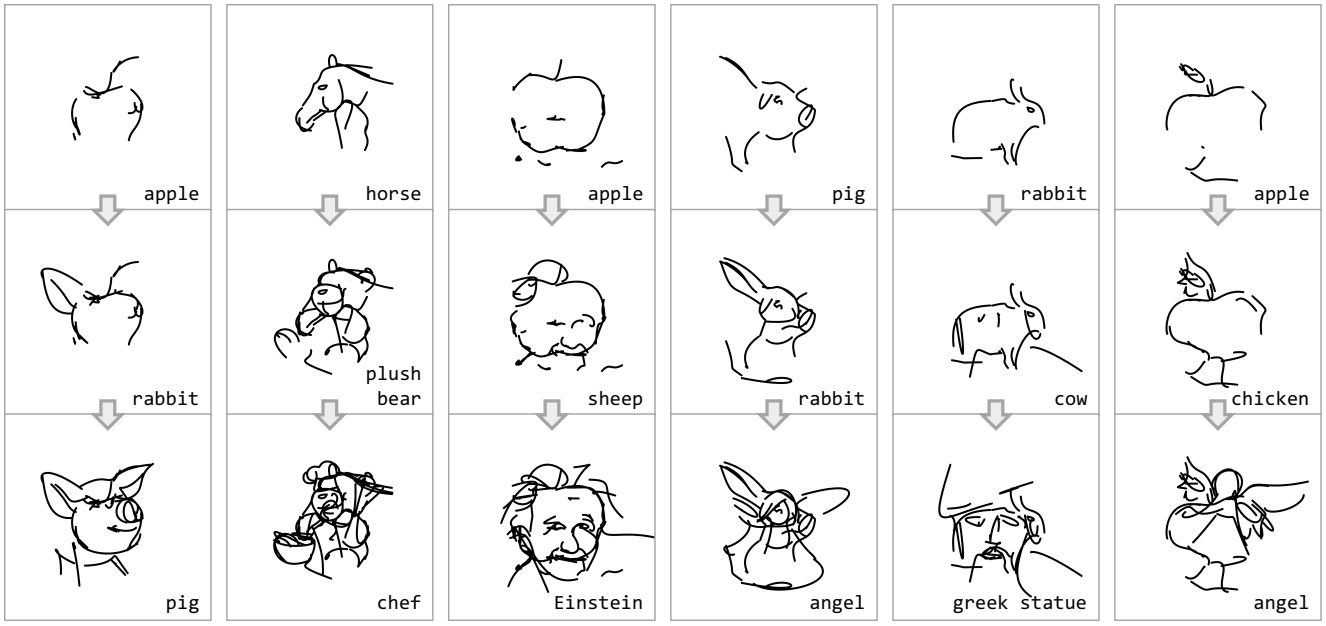



Figure 15. **Additional 3-phase progressive illusion results produced by our method.**

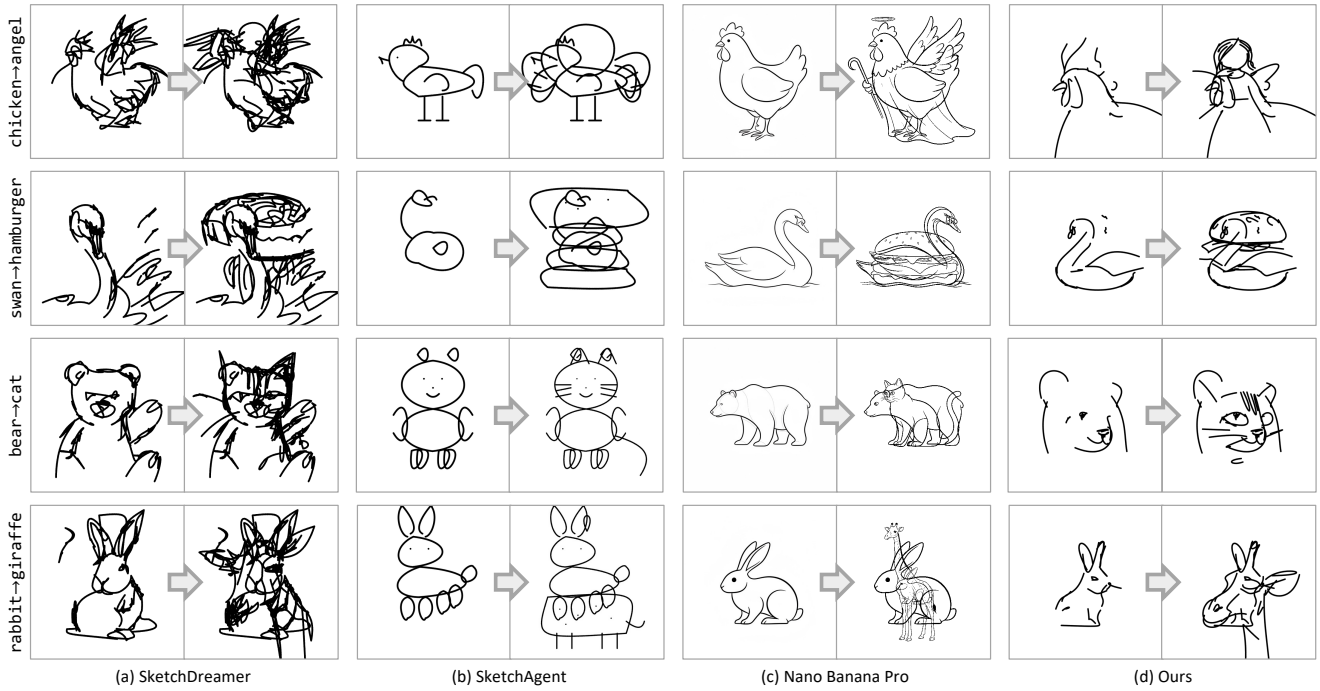



Figure 16. **Additional qualitative comparisons.**

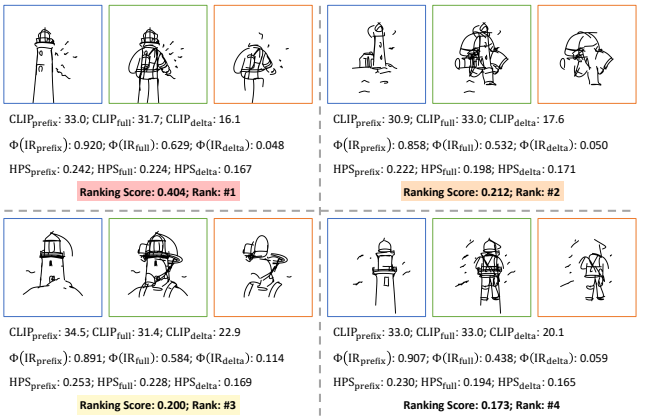



Figure 17. **Metric-based ranking.**

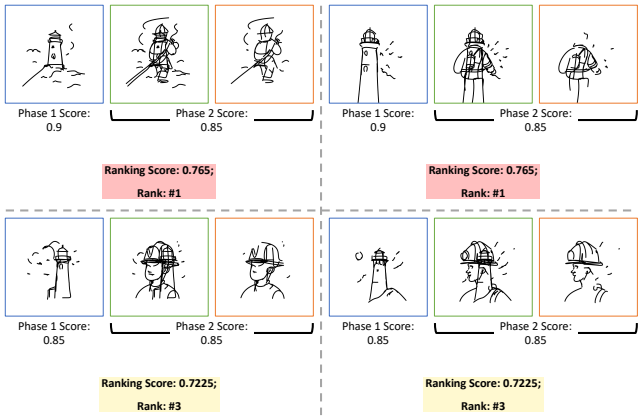



Figure 18. **GPT-based ranking.**

sketching ai agent. so you think you can sketch? *ACM Transactions on Graphics (TOG)*, 39(6):1–15, 2020. 3

[11] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Fahad Shahbaz Khan, Jorma Laaksonen, and Michael Felsberg. Doodleformer: Creative sketch drawing with transformers. In *European Conference on Computer Vision*, pages 338–355. Springer, 2022. 2

[12] Ayan Kumar Bhunia, Subhadeep Koley, Amandeep Kumar, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Sketch2saliency: Learning to detect salient objects from human drawings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2733–2743, 2023. 3

[13] Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2): 115, 1987. 3

[14] Irving Biederman and Ginny Ju. Surface versus edge-based determinants of visual recognition. *Cognitive psychology*, 20(1):38–64, 1988. 3

[15] Ryan Burgert, Xiang Li, Abe Leite, Kanchana Ranasinghe, and Michael Ryoo. Diffusion illusions: Hiding images in plain sight. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 3

[16] John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 2009. 2

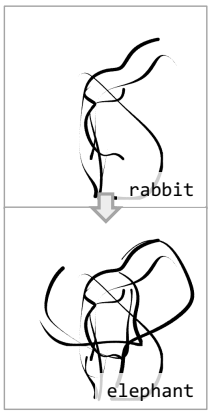
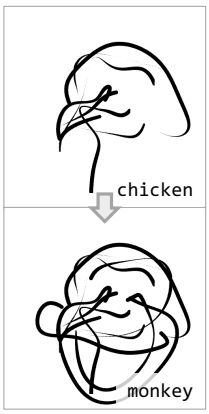
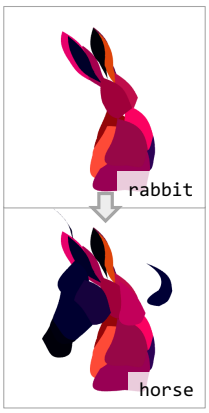



Figure 19. **Extension on variable-width B-spline.**



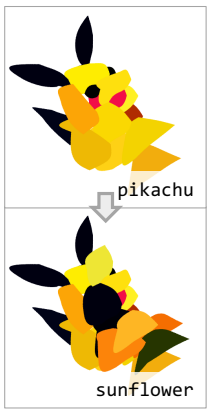
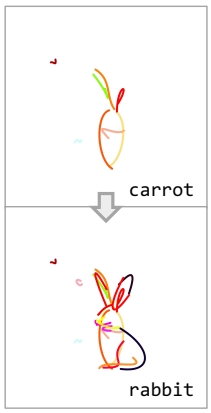




Figure 20. **Extension on vector graph.**



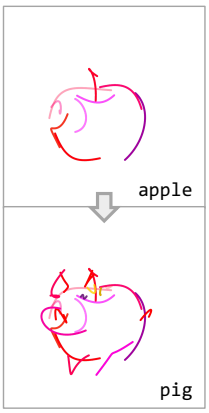




Figure 21. **Extension on colored strokes.**


[17] Alexandre Carlier, Martin Danelljan, Alexandre Alahi, and Radu Timofte. Deepsvg: A hierarchical generative network for vector graphics animation. *Advances in Neural Information Processing Systems*, 33:16351–16361, 2020. 2

[18] Patrick Cavanagh. The artist as neuroscientist. *Nature*, 434 (7031):301–307, 2005. 3

[19] Pascal Chang, Sergio Sancho, Jingwei Tang, Markus Gross, and Vinicius Azevedo. Lookingglass: Generative anamorphoses via laplacian pyramid warping. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24–33, 2025. 3

[20] Chen-Hao Chao, Wei-Fang Sun, Bo-Wun Cheng, Yi-Chen Lo, Chia-Che Chang, Yu-Lun Liu, Yu-Lin Chang, Chia-Ping Chen, and Chun-Yi Lee. Denoising likelihood score matching for conditional score-based data generation. *arXiv preprint arXiv:2203.14206*, 2022. 3

[21] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)*, 42(4):1–10, 2023. 3

[22] Zehao Chen and Rong Pan. Svgbuilder: Component-based colored svg generation with text-guided autoregressive transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2358–2366, 2025. 2

[23] Ziyang Chen, Daniel Geng, and Andrew Owens. Images that sound: Composing images and sounds on a single canvas. *Advances in Neural Information Processing Systems*, 37: 85045–85073, 2024. 3

[24] Hung-Kuo Chu, Wei-Hsin Hsu, Niloy J Mitra, Daniel Cohen-Or, Tien-Tsin Wong, and Tong-Yee Lee. Camouflage images. *ACM Trans. Graph.*, 29(4):51–1, 2010. 3

[25] Ayan Das, Yongxin Yang, Timothy Hospedales, Tao Xiang, and Yi-Zhe Song. Béziersketch: A generative model for scalable vector sketches. In *European conference on computer vision*, pages 632–647. Springer, 2020. 2

[26] Nicholas Davis, Chih-PIn Hsiao, Kunwar Yashraj Singh, Lisa Li, Sanat Moningi, and Brian Magerko. Drawing apprentice: An enactive co-creative agent for artistic collaboration. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*, pages 185–186, 2015. 3

[27] Paul De Casteljau. Outillages méthodes calcul. *Andr e Citro en Automobiles SA, Paris*, 4:25, 1959. 3

[28] Soumyaratna Debnath, Ashish Tiwari, Kaustubh Sadekar, and Shanmuganathan Raman. Rasp: Revisiting 3d anamorphic art for shadow-guided packing of irregular objects. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5849–5858, 2025. 3

[29] Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? *ACM Transactions on graphics (TOG)*, 31 (4):1–10, 2012. 2, 3

[30] Judith E Fan, Wilma A Bainbridge, Rebecca Chamberlain, and Jeffrey D Wammes. Drawing as a versatile cognitive tool. *Nature Reviews Psychology*, 2(9):556–568, 2023. 3

[31] Yue Feng, Vaibhav Sanjay, Spencer Lutz, Badour AlBahar, Songwei Ge, and Jia-Bin Huang. Illusion3d: 3d multiview illusion with 2d diffusion priors. *arXiv preprint arXiv:2412.09625*, 2024. 3

[32] Kevin Frans, Lisa Soros, and Olaf Witkowski. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. *Advances in Neural Information Processing Systems*, 35:5207–5218, 2022. 2

[33] Xiang Gao, Shuai Yang, and Jiaying Liu. Ptdiffusion: Free lunch for generating optical illusion hidden pictures with phase-transferred diffusion model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18240–18249, 2025. 3

[34] Songwei Ge, Vedanuj Goswami, C Lawrence Zitnick, and Devi Parikh. Creative sketch generation. *arXiv preprint arXiv:2011.10039*, 2020. 2

[35] Daniel Geng, Inbum Park, and Andrew Owens. Factorized diffusion: Perceptual illusions by noise decomposition. In *European Conference on Computer Vision*, pages 366–384. Springer, 2024. 3

[36] Daniel Geng, Inbum Park, and Andrew Owens. Visual anagrams: Generating multi-view optical illusions with diffusion models. In *Proceedings of the IEEE/CVF Conference*

*on Computer Vision and Pattern Recognition*, pages 24154–24163, 2024. 2, 3, 5

[37] Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Rezende, and Daan Wierstra. Draw: A recurrent neural network for image generation. In *International conference on machine learning*, pages 1462–1471. PMLR, 2015. 2

[38] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, et al. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *Advances in Neural Information Processing Systems*, 36:15890–15902, 2023. 3

[39] David Ha and Douglas Eck. A neural representation of sketch drawings. *arXiv preprint arXiv:1704.03477*, 2017. 2

[40] Amir Hertz, Kfir Aberman, and Daniel Cohen-Or. Delta denoising score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2328–2337, 2023. 3

[41] Aaron Hertzmann. Painterly rendering with curved brush strokes of multiple sizes. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 453–460, 1998. 2

[42] Aaron Hertzmann. A survey of stroke-based rendering. Institute of Electrical and Electronics Engineers, 2003. 2

[43] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning, 2022. 5

[44] Or Hirschorn, Amir Jevnisek, and Shai Avidan. Optimize & reduce: a top-down approach for image vectorization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2148–2156, 2024. 3

[45] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3

[46] Kai-Wen Hsiao, Jia-Bin Huang, and Hung-Kuo Chu. Multi-view wire art. *ACM Trans. Graph.*, 37(6):242, 2018. 3

[47] Teng Hu, Ran Yi, Haokun Zhu, Liang Liu, Jinlong Peng, Yabiao Wang, Chengjie Wang, and Lizhuang Ma. Stroke-based neural painting and stylization with dynamically predicted painting region. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 7470–7480, 2023. 3

[48] Yi-Chuan Huang, Jiewen Chan, Hao-Jen Chien, and Yu-Lun Liu. Voxify3d: Pixel art meets volumetric rendering. *arXiv preprint arXiv:2512.07834*, 2025. 3

[49] Zhewei Huang, Wen Heng, and Shuchang Zhou. Learning to paint with model-based deep reinforcement learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8709–8718, 2019. 3

[50] Francisco Ibarrola, Tomas Lawton, and Kazjon Grace. A collaborative, interactive and context-aware drawing agent for co-creative design. *IEEE Transactions on Visualization and Computer Graphics*, 30(8):5525–5537, 2023. 3

[51] Shir Iluz, Yael Vinker, Amir Hertz, Daniel Berio, Daniel Cohen-Or, and Ariel Shamir. Word-as-image for semantic typography. *ACM Transactions on Graphics (TOG)*, 42(4): 1–11, 2023. 3

[52] Ajay Jain, Amber Xie, and Pieter Abbeel. Vectorfusion: Text-to-svg by abstracting pixel-based diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1911–1920, 2023. 2, 3

[53] Jiaxiu Jiang, Yabo Zhang, Kailai Feng, Xiaohe Wu, Wenbo Li, Renjing Pei, Fan Li, and Wangmeng Zuo. Mcˆ 2: Multi-concept guidance for customized multi-concept generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2802–2812, 2025. 3

[54] Jonas Jongejan, Henry Rowley, Takashi Kawashima, Jongmin Kim, and Nick Fox-Gieg. Quick, draw! the data. *dataset for online game Quick, Draw*, 2016. 2

[55] Marcelo Isaias de Moraes Junior and Moacir Antonelli Ponti. On the temporality for sketch representation learning. *arXiv preprint arXiv:2512.04007*, 2025. 2

[56] Gaetano Kanizsa, Paolo Legrenzi, and Paolo Bozzi. Organization in vision: Essays on gestalt perception. *(No Title)*, 1979. 3

[57] Pegah Karimi, Jeba Rezwana, Safat Siddiqui, Mary Lou Maher, and Nasrin Dehbozorgi. Creative sketching partner: an analysis of human-ai co-creativity. In *Proceedings of the 25th international conference on intelligent user interfaces*, pages 221–230, 2020. 3

[58] Oren Katzir, Or Patashnik, Daniel Cohen-Or, and Dani Lischinski. Noise-free score distillation. *arXiv preprint arXiv:2310.17590*, 2023. 3

[59] Bo-Hsu Ke, You-Zhe Xie, Yu-Lun Liu, and Wei-Chen Chiu. Stealthattack: Robust 3d gaussian splatting poisoning via density-guided illusions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 27400–27411, 2025. 3

[60] Subin Kim, Kyungmin Lee, June Suk Choi, Jongheon Jeong, Kihyuk Sohn, and Jinwoo Shin. Collaborative score distillation for consistent visual synthesis. *arXiv preprint arXiv:2307.04787*, 2023. 3

[61] Subhadeep Koley, Ayan Kumar Bhunia, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. How to handle sketch-abstraction in sketch-based image retrieval? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16859–16869, 2024. 3

[62] Juil Koo, Chanho Park, and Minhyuk Sung. Posterior distillation sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13352–13361, 2024. 3

[63] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1931–1941, 2023. 3

[64] Tomas Lawton, Francisco J Ibarrola, Dan Ventura, and Kazjon Grace. Drawing with reframer: Emergence and control in co-creative ai. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*, pages 264–277, 2023. 3

[65] Jie-Ying Lee, Yi-Ruei Liu, Shr-Ruei Tsai, Wei-Cheng Chang, Chung-Ho Wu, Jiewen Chan, Zhenjun Zhao,

Chieh Hubert Lin, and Yu-Lun Liu. Skyfall-gs: Synthesizing immersive 3d urban scenes from satellite imagery. *arXiv preprint arXiv:2510.15869*, 2025. 3

[66] Ke Li, Kaiyue Pang, Jifei Song, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Honggang Zhang. Universal sketch perceptual grouping. In *Proceedings of the european conference on computer vision (ECCV)*, pages 582–597, 2018. 3

[67] Tzu-Mao Li, Michal Lukáč, Michaël Gharbi, and Jonathan Ragan-Kelley. Differentiable vector graphics rasterization for editing and learning. *ACM Transactions on Graphics (TOG)*, 39(6):1–15, 2020. 3

[68] Yixun Liang, Xin Yang, Jiantao Lin, Haodong Li, Xiaogang Xu, and Yingcong Chen. Luciddreamer: Towards high-fidelity text-to-3d generation via interval score matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6517–6526, 2024. 3

[69] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 300–309, 2023. 3

[70] Hangyu Lin, Yanwei Fu, Xiangyang Xue, and Yu-Gang Jiang. Sketch-bert: Learning sketch bidirectional encoder representation from transformers by self-supervised learning of sketch gestalt. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6758–6767, 2020. 2

[71] Fang Liu, Xiaoming Deng, Yu-Kun Lai, Yong-Jin Liu, Cuixia Ma, and Hongan Wang. Sketchgan: Joint sketch completion and recognition with generative adversarial network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5830–5839, 2019. 2

[72] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European conference on computer vision*, pages 423–439. Springer, 2022. 3

[73] Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Ruifeng Deng, Xin Li, Errui Ding, and Hao Wang. Paint transformer: Feed forward neural painting with stroke prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6598–6607, 2021. 3

[74] Xi Liu, Chaoyi Zhou, Nanxuan Zhao, and Siyu Huang. B\'ezier splatting for fast and differentiable vector graphics rendering. *arXiv preprint arXiv:2503.16424*, 2025. 3

[75] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. *arXiv preprint arXiv:2309.03453*, 2023. 3

[76] Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones: Concept neurons in diffusion models for customized generation. *arXiv preprint arXiv:2303.05125*, 2023. 3

[77] Raphael Gontijo Lopes, David Ha, Douglas Eck, and Jonathon Shlens. A learned representation for scalable vector graphics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7930–7939, 2019. 3

[78] Artem Lukoianov, Haitz Sáez de Ocáriz Borde, Kristjan Greenewald, Vitor Guizilini, Timur Bagautdinov, Vincent Sitzmann, and Justin M Solomon. Score distillation via reparametrized ddim. *Advances in Neural Information Processing Systems*, 37:26011–26044, 2024. 3

[79] Rundong Luo, Noah Snavely, and Wei-Chiu Ma. Shadowdraw: From any object to shadow-drawing compositional art. *arXiv preprint arXiv:2512.05110*, 2025. 2, 5

[80] Xu Ma, Yuqian Zhou, Xingqian Xu, Bin Sun, Valerii Filev, Nikita Orlov, Yun Fu, and Humphrey Shi. Towards layer-wise image vectorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16314–16323, 2022. 3

[81] Penousal Machado, Adriano Vinhas, Joao Correia, and Aniko Ekárt. Evolving ambiguous images. *AI Matters*, 2(1):7–8, 2015. 3

[82] David McAllister, Songwei Ge, Jia-Bin Huang, David W Jacobs, Alexei A Efros, Aleksander Holynski, and Angjoo Kanazawa. Rethinking score distillation as a bridge between image distributions. *Advances in Neural Information Processing Systems*, 37:33779–33804, 2024. 3

[83] Niloy J Mitra and Mark Pauly. Shadow art. *ACM Trans. Graph.*, 28(5):156, 2009. 3

[84] Umar Riaz Muhammad, Yongxin Yang, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Learning deep sketch abstraction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8014–8023, 2018. 3

[85] Umar Riaz Muhammad, Yongxin Yang, Timothy M Hospedales, Tao Xiang, and Yi-Zhe Song. Goal-driven sequential data abstraction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 71–80, 2019. 3

[86] Changhoon Oh, Jungwoo Song, Jinhan Choi, Seonghyeon Kim, Sungwoo Lee, and Bongwon Suh. I lead, you help but only with enough details: Understanding user experience of co-creation with artificial intelligence. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pages 1–13, 2018. 3

[87] Aude Oliva, Antonio Torralba, and Philippe G Schyns. Hybrid images. *ACM Transactions on Graphics (TOG)*, 25(3): 527–532, 2006. 3

[88] Artemis Panagopoulou, Coby Melkin, and Chris Callison-Burch. Evaluating vision-language models on bistable images. *arXiv preprint arXiv:2405.19423*, 2024. 3

[89] Sagi Polaczek, Yuval Alaluf, Elad Richardson, Yael Vinker, and Daniel Cohen-Or. Neuralsvg: An implicit representation for text-to-vector generation. *arXiv preprint arXiv:2501.03992*, 2025. 2

[90] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 2, 3

[91] Louis Pratt, Andrew Johnston, and Nico Pietroni. Bending the light: Next generation anamorphic sculptures. *Computers & Graphics*, 114:210–218, 2023. 3

[92] Zhiyu Qu, Yulia Gryaditskaya, Ke Li, Kaiyue Pang, Tao Xiang, and Yi-Zhe Song. Sketchxai: A first look at explainabil-

ity for human sketches. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 23327–23337, 2023. 3

[93] Zhiyu Qu, Tao Xiang, and Yi-Zhe Song. Sketchdreamer: Interactive text-augmented creative sketch ideation. *arXiv preprint arXiv:2308.14191*, 2023. 2, 5, 6

[94] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[95] Pradyumna Reddy, Michael Gharbi, Michal Lukac, and Niloy J Mitra. Im2vec: Synthesizing vector graphics without vector supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7342–7351, 2021. 3

[96] Leo Sampaio Ferraz Ribeiro, Tu Bui, John Collomosse, and Moacir Ponti. Sketchformer: Transformer-based representation for sketched structure. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14153–14162, 2020. 2

[97] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[98] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 3

[99] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM Transactions on Graphics (TOG)*, 35(4):1–12, 2016. 2

[100] Steven M Seitz and Charles R Dyer. View morphing. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 21–30, 1996. 3

[101] Maria Shugrina, Chin-Ying Li, and Sanja Fidler. Neural brushstroke engine: learning a latent style space of interactive drawing tools. *ACM Transactions on Graphics (TOG)*, 41(6):1–18, 2022. 3

[102] Edgar Simo-Serra, Satoshi Iizuka, Kazuma Sasaki, and Hiroshi Ishikawa. Learning to simplify: fully convolutional networks for rough sketch cleanup. *ACM Transactions on Graphics (TOG)*, 35(4):1–11, 2016. 3

[103] Johannes JD Singer, Radoslaw M Cichy, and Martin N Hebart. The spatiotemporal neural dynamics of object recognition for natural images and line drawings. *Journal of Neuroscience*, 43(3):484–500, 2023. 3

[104] Guoyao Su, Yonggang Qi, Kaiyue Pang, Jie Yang, and Yi-Zhe Song. Sketchhealer a graph-to-sequence network for recreating partial human sketches. In *Proceedings of The 31st British Machine Vision Virtual Conference (BMVC 2020)*, pages 1–14. British Machine Vision Association, 2020. 2

[105] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. *Advances in Neural Information Processing Systems*, 36: 51202–51233, 2023. 3

[106] Vikas Thamizharasan, Difan Liu, Shantanu Agarwal, Matthew Fisher, Michaël Gharbi, Oliver Wang, Alec Jacobson, and Evangelos Kalogerakis. Vecfusion: Vector font generation with diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7943–7952, 2024. 2

[107] Vikas Thamizharasan, Difan Liu, Matthew Fisher, Nanxuan Zhao, Evangelos Kalogerakis, and Michal Lukac. Nivel: Neural implicit vector layers for text-to-vector generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4589–4597, 2024. 3

[108] Yael Vinker, Ehsan Pajouheshgar, Jessica Y Bo, Roman Christian Bachmann, Amit Haim Bermano, Daniel Cohen-Or, Amir Zamir, and Ariel Shamir. Clipasso: Semantically-aware object sketching. *ACM Transactions on Graphics (TOG)*, 41(4):1–11, 2022. 2

[109] Yael Vinker, Yuval Alaluf, Daniel Cohen-Or, and Ariel Shamir. Clipascene: Scene sketching with different types and levels of abstraction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4146–4156, 2023. 2

[110] Yael Vinker, Tamar Rott Shaham, Kristine Zheng, Alex Zhao, Judith E Fan, and Antonio Torralba. Sketchagent: Language-driven sequential sketch generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23355–23368, 2025. 2, 3, 5, 6

[111] Johan Wagemans, James H Elder, Michael Kubovy, Stephen E Palmer, Mary A Peterson, Manish Singh, and Rüdiger Von der Heydt. A century of gestalt psychology in visual perception: I. perceptual grouping and figure–ground organization. *Psychological bulletin*, 138(6):1172, 2012. 3

[112] Caoliwen Wang, Bailin Deng, and Juyong Zhang. Neural shadow art. *arXiv preprint arXiv:2411.19161*, 2024. 3

[113] Jiawei Wang and Changjian Li. Contextseg: Sketch semantic segmentation by querying the context with attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3679–3688, 2024. 3

[114] Qiang Wang, Haoge Deng, Yonggang Qi, Da Li, and Yi-Zhe Song. Sketchknitter: Vectorized sketch generation with diffusion models. In *The eleventh international conference on learning representations*, 2023. 2

[115] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *Advances in neural information processing systems*, 36:8406–8441, 2023. 3

[116] Veith Weilnhammer, Heiner Stuke, Guido Hesselmann, Philipp Sterzer, and Katharina Schmack. A predictive coding account of bistable perception-a model-based fmri study. *PLoS computational biology*, 13(5):e1005536, 2017. 3

[117] Max Wertheimer. Laws of organization in perceptual forms. first published as untersuchungen zur lehre von der gestalt ii. *Psychologische Forschung*, 4:301–350, 1923. 3

[118] Holger Winnemöller, Jan Eric Kyprianidis, and Sven C Olsen. Xdog: An extended difference-of-gaussians compendium including advanced image stylization. *Computers & Graphics*, 36(6):740–753, 2012. 2

[119] Kang Wu, Renjie Chen, Xiao-Ming Fu, and Ligang Liu. Computational mirror cup and saucer art. *ACM Transactions on Graphics (TOG)*, 41(5):1–15, 2022. 3

[120] Ronghuan Wu, Wanchao Su, Kede Ma, and Jing Liao. Iconshop: Text-guided vector icon synthesis with autoregressive transformers. *ACM Transactions on Graphics (TOG)*, 42(6): 1–14, 2023. 2

[121] Ronghuan Wu, Wanchao Su, and Jing Liao. Chat2svg: Vector graphics generation with large language models and image diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23690–23700, 2025. 2

[122] Xiaoshi Wu, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*, 2023. 5

[123] Yan Wu, Jeff Donahue, David Balduzzi, Karen Simonyan, and Timothy Lillicrap. Logan: Latent optimisation for generative adversarial networks. *arXiv preprint arXiv:1912.00953*, 2019. 3

[124] Guangxuan Xiao, Tianwei Yin, William T Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *International Journal of Computer Vision*, 133(3):1175–1194, 2025. 3

[125] Ximing Xing, Chuang Wang, Haitao Zhou, Jing Zhang, Qian Yu, and Dong Xu. Diffsketcher: Text guided vector sketch synthesis through latent diffusion models. *Advances in Neural Information Processing Systems*, 36:15869–15889, 2023. 2, 3

[126] Ximing Xing, Juncheng Hu, Jing Zhang, Dong Xu, and Qian Yu. Svgfusion: Scalable text-to-svg generation via vector space diffusion. *arXiv preprint arXiv:2412.10437*, 2024. 3

[127] Ximing Xing, Haitao Zhou, Chuang Wang, Jing Zhang, Dong Xu, and Qian Yu. Svgdreamer: Text guided svg generation with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4546–4555, 2024. 2

[128] Ximing Xing, Juncheng Hu, Guotao Liang, Jing Zhang, Dong Xu, and Qian Yu. Empowering llms to understand and generate complex vector graphics. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19487–19497, 2025. 2

[129] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: learning and evaluating human preferences for text-to-image generation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 15903–15935, 2023. 5

[130] Peng Xu, Timothy M Hospedales, Qiyue Yin, Yi-Zhe Song, Tao Xiang, and Liang Wang. Deep learning for free-hand sketch: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):285–312, 2022. 2

[131] Zhiyuan Xu, Yinhe Chen, Huan-ang Gao, Weiyan Zhao, Guiyu Zhang, and Hao Zhao. Diffusion-based visual anagram as multi-task learning. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 919–928. IEEE, 2025. 3

[132] Lumin Yang, Jiajie Zhuang, Hongbo Fu, Xiangzhi Wei, Kun Zhou, and Youyi Zheng. Sketchgnn: Semantic sketch segmentation with graph neural networks. *ACM Transactions on Graphics (TOG)*, 40(3):1–13, 2021. 3

[133] Yinghua Yao, Yuangang Pan, Jing Li, Ivor Tsang, and Xin Yao. Proud: Pareto-guided diffusion model for multi-objective generation. *Machine Learning*, 113(9):6511–6538, 2024. 3

[134] Qian Yu, Yongxin Yang, Feng Liu, Yi-Zhe Song, Tao Xiang, and Timothy M Hospedales. Sketch-a-net: A deep neural network that beats humans. *International journal of computer vision*, 122(3):411–425, 2017. 3

[135] Sicong Zang, Shuhui Gao, and Zhijun Fang. Generating sketches in a hierarchical auto-regressive process for flexible sketch drawing manipulation at stroke-level. *arXiv preprint arXiv:2511.07889*, 2025. 3

[136] Peiying Zhang, Nanxuan Zhao, and Jing Liao. Text-to-vector generation with neural path representation. *ACM Transactions on Graphics (TOG)*, 43(4):1–13, 2024. 2

[137] Boheng Zhao, Rana Hanocka, and Raymond A Yeh. Ambigen: Generating ambigrams from pre-trained diffusion model. *arXiv preprint arXiv:2312.02967*, 2023. 3

[138] Zhongyin Zhao, Ye Chen, Zhangli Hu, Xuanhong Chen, and Bingbing Ni. Vector graphics generation via mutually impulsed dual-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4420–4428, 2024. 2

[139] Zhengxia Zou, Tianyang Shi, Shuang Qiu, Yi Yuan, and Zhenwei Shi. Stylized neural painting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15689–15698, 2021. 3