Supplementary Material

In this supplementary material, we provide detailed descriptions of the algorithm and implementation, additional qualitative comparisons, experimental results, a detailed user study setup, and limitations with discussion.

Table of Contents

A. DETAILS OF SCRIBBLE DIFFUSION

Fig. S1 shows images inferred from the scribble prompt with different timesteps. As discussed in the P2 weighting [25], we extend the scribble prompt at certain timesteps related to content generation, effectively enhancing alignment between the scribble and the image.

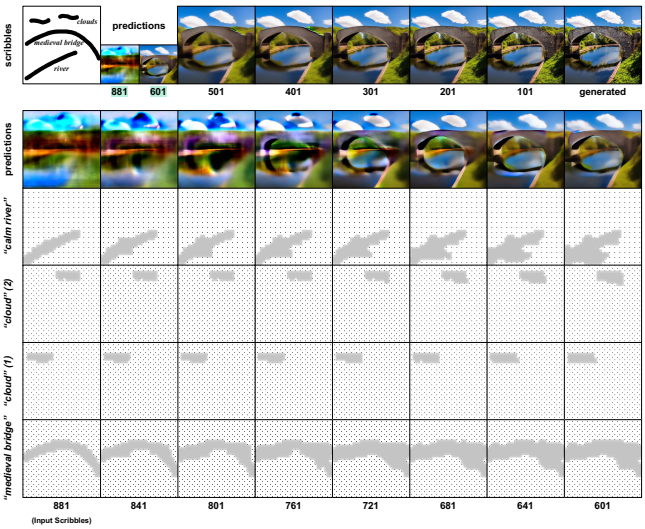


Fig. S1: **Scribble Propagation.** At specific timesteps, our method extends the input scribble, improving alignment with the generated image.

Different Propagation Methods. Naively applying techniques like Gaussian kernel or dilation to intentionally thicken scribbles is suboptimal or ineffective. Thickening the lines can distort the abstract shape that the user intended to express, as the expanded lines may blur or dilute the original form. This is particularly problematic for objects with fine details, as certain parts of the object should be expanded while others, such as thin features like an elephant's trunk, should remain unblurred to preserve accuracy. An example of this issue is illustrated in Fig. S2 (second row), where despite thickening the scribble by 16 times from the start, the resulting image lacks key features like sunglasses, leading to an unnatural outcome without proper scribble propagation.

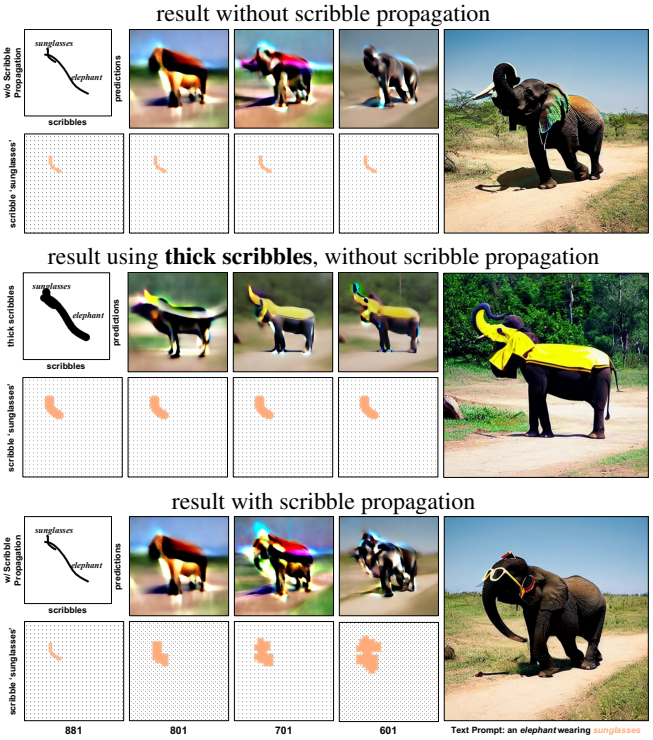


Fig. S2: **Additional Ablation of Scribble Propagation and the comparison with only using *thick scribbles*.** Without scribble propagation, the generated object ***"sunglasses"*** is not properly captured due to the thin nature of the input scribble, leading to incomplete and incorrect object generation. By applying scribble propagation, our method extends the input scribble over time, ensuring that finer details such as the ***"sunglasses"*** are captured and aligned with the text prompt. (Text Prompt: *an* ***elephant*** *wearing* ***sunglasses***)

B. IMPLEMENTATION DETAILS

Our method is implemented on the GLIGEN [27] baseline. GLIGEN allows the use of bounding boxes as grounding inputs, so we first generate bounding boxes that encompass the scribbles, adding 5% padding to both the width and height of each box. These bounding boxes are then used as ground-

ing inputs for GLIGEN. In our implementation, several hyperparameters were chosen to balance the effectiveness and efficiency of the proposed method. For the scribble propagation, we set the merging threshold τ to 0.001 to effectively merge anchors near the boundary of a scribble without over-expanding into irrelevant regions. The number of top-k tokens for token selection was fixed at 20, providing a sufficient range for propagating the scribble to neighboring areas. The scribble propagation starts at timestep $k_1 = 5$ and ends at timestep $k_2 = 15$ within the reverse diffusion process, ensuring that the model has ample time to incorporate the scribble information early in the denoising steps while maintaining computational efficiency.

For self-attention map aggregation, we utilized multiple resolutions, specifically [8, 16, 32, 64], to capture attention from various scales and downsampled the aggregated self-attention maps to a resolution of 64. This multi-resolution approach allowed us to better capture fine-grained spatial information while maintaining computational feasibility.

The moment alignment process was guided by two terms: λ_1, which controls the contribution of the centroid moment loss, and λ_2, which regulates the central moment loss. We empirically set both λ_1 and λ_2 to 0.6, which provided a good balance between aligning the position and the orientation of the generated object with the scribble prompt.

Additionally, to ensure balanced optimization, the loss terms were weighted with a ratio of 5:3 for the cross-attention focal loss ($\mathcal{L}_{\texttt{focal}}$) and the moment loss ($\mathcal{L}_{\texttt{moment}}$), respectively. This weighting reflects the relative importance of ensuring precise alignment between the generated image and the scribble in terms of both spatial placement and orientation. Furthermore, we set β in Eq. (5) as 2.0. Finally, the anchor grid size was set to 16×16 with each anchor representing a 2×2 token cluster, which provided sufficient granularity for the scribble propagation process without causing unnecessary computational overhead.

C. EXPERIMENTAL SETUP DETAILS

BoxDiff [3] primarily uses bounding box guidance but also includes scribble constraints in certain cases. DenseDiffusion [4], on the other hand, leverages region masks for image synthesis. For a fair comparison, we run BoxDiff experiments using the GLIGEN pipeline, while DenseDiffusion experiments are conducted using Stable Diffusion v1.5, as it directly modifies the attention layers in Stable Diffusion. We fine-tune ControlNet using scribble inputs from the PASCAL-Scribble training set for 100 epochs.

D. OVERALL ALGORITHM

The overall workflow of our method, ScribbleDiff, involves iterative guidance during the reverse diffusion process using two main components: **Cross-Attention Control with Moment Alignment** and **Scribble Propagation**.

At each timestep in the reverse diffusion process, the latent code is adjusted based on the focal loss and moment alignment, ensuring that the generated object reflects both the spatial alignment and orientation of the scribble input. The scribble propagation process occurs within a specified interval of timesteps (k_1 to k_2) and involves iteratively expanding the scribble regions. Notably, the merging of scribble regions is guided by a distance metric similar to Dijkstra's algorithm, where anchors near the boundary of a scribble are evaluated based on Kullback-Leibler divergence. The algorithm selects the k closest anchors, gradually extending the scribble regions. This approach is akin to a shortest-path search, where regions with the smallest divergence are progressively included in the scribble. For further details on the algorithm, see Algorithm 1.

E. MORE QUALITATIVE RESULTS

Additional qualitative comparison results are provided alongside Fig. 3. The additional experimental results Fig. S3 show that the proposed model demonstrates better alignment with scribbles.

In Fig. S4, we compare ScribbleDiff with fine-tuned ControlNet and other diffusion models on the PASCAL-Scribble dataset. ScribbleDiff surpasses both training-free and fine-tuned methods in reflecting scribble prompts without requiring additional training. Unlike ControlNet, which lacks explicit learning of scribble direction, ScribbleDiff leverages moment alignment to better capture the intended prompts, achieving superior alignment and directional consistency.

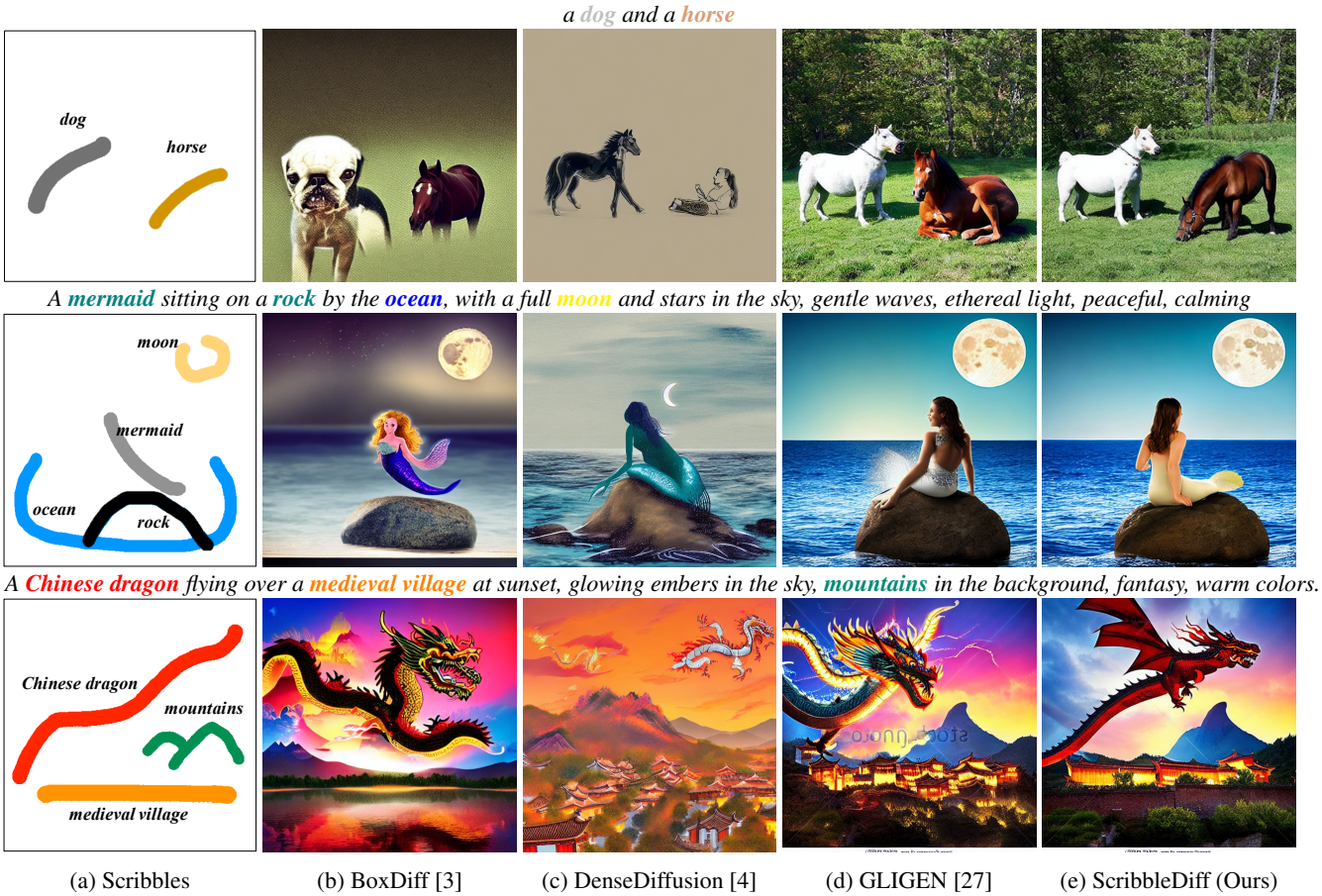
Fig. S5 presents additional examples generated by ScribbleDiff. The scribbles serve as a structural guide, providing the layout that the images should follow.

F. ADDITIONAL ABLATION STUDIES

$\mathcal{L}_{\texttt{moment}}$	Scribble Prop.	mIoU (↑)	Scribble Ratio (↑)
✗	✗	0.391	0.697
✓	✗	0.406	0.715
✗	✓	0.396	0.697
✓	✓	**0.410**	**0.717**

Table S1: **Ablation study on our proposed components.** With all components activated, our approach achieves the highest mIoU and Scribble Ratio score. This result indicates that each element plays a vital role in enhancing the quality of the final output.

We conduct an ablation study on the PASCAL Scribble dataset to evaluate the effectiveness of our components: mo-



(a) Scribbles (b) BoxDiff [3] (c) DenseDiffusion [4] (d) GLIGEN [27] (e) ScribbleDiff (Ours)

Fig. S3: **Additional qualitative comparison of Text-to-Image generation methods using scribble prompts.** ScribbleDiff yields outcomes that better reflect the scribble inputs, especially concerning the accuracy of object orientations and abstract shape representation.

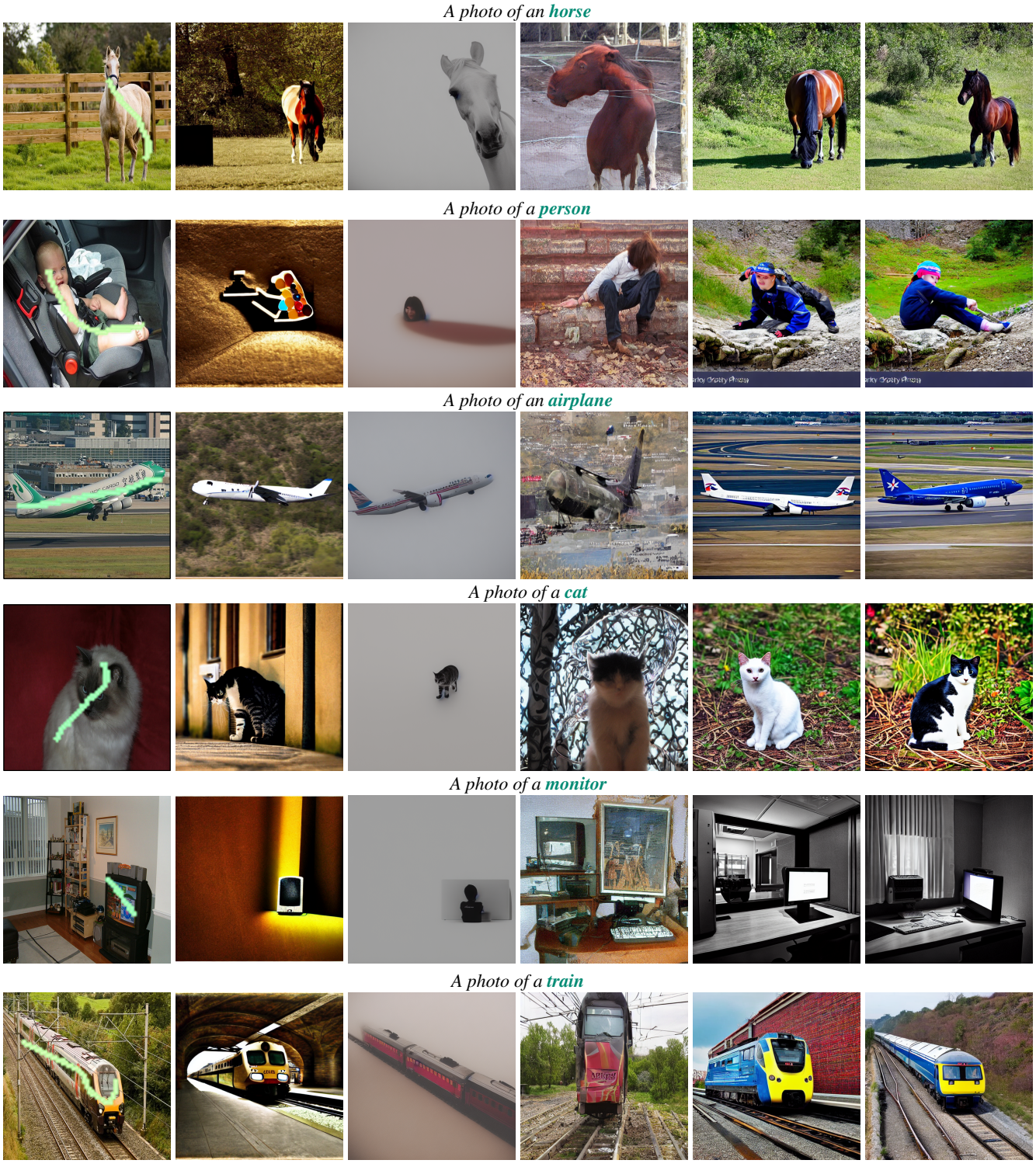
ment loss $\mathcal{L}_{\texttt{moment}}$ and scribble propagation. Table S1 shows the performance of different configurations in terms of mIoU and Scribble Ratio. As shown in Table S1, the increase of $\mathcal{L}_{\texttt{moment}}$ improves both the mIoU and scribble ratio. Moreover, the proposed scribble propagation also contributes to further improvements in mIoU. Comprehensively, employing scribble propagation and $\mathcal{L}_{\texttt{moment}}$ achieves a 0.02 point improvement in the mIoU and 0.02 gain in the scribble ratio.

Fig. S6 demonstrates that, without propagation, scribbles remain narrow and constrained (e.g., timestep 901), leading to incomplete object representation. With scribble propagation, scribbles expand and improve object coverage by timestep 701. Furthermore, as demonstrated in Fig. S2, omitting scribble propagation results in significant issues during generation, particularly when handling thin and sparse scribbles. For example, without scribble propagation, the thin scribble representing *"sunglasses"* is ignored, and no sunglasses are generated. By contrast, when applying scribble propagation, our method iteratively extends the scribble during the denoising process, ensuring that smaller, detailed elements—such as the sunglasses—are accurately generated and aligned with the input prompt. This effect is particularly beneficial when handling thin scribbles, as they are more prone to being overlooked during generation.

We also show the impact of the scales λ_1 and λ_2 while fixing other parameters in Fig. S8. Both λ_1 and λ_2 are hyperparameters used to weigh the centroid and central moment losses. We observe that as the λ_1 and λ_2 scales increase, the image becomes more closely aligned with the thin scribble input. This is particularly noticeable in the *bamboo raft*, whose shape adapts to better reflect the thin scribble structure. In addition, the orientation of the *cute panda* moves from facing forward to the left by increasing λ_1 and λ_2

G. USER STUDY DETAILS

User study focused on evaluating image quality and alignment to determine the human-preferred approach. Human evaluators were presented with a prompt and an input scribble and were asked to select the best result from four different models: BoxDiff, DenseDiffusion, GLIGEN, and our proposed method. The images were randomly ordered and labeled A



(a) Scribbles [11] (b) BoxDiff [3] (c) DenseDiffusion [4] (d) ControlNet [7] (e) GLIGEN [27] (f) ScribbleDiff (Ours)

Fig. S4: **Additional qualitative results on the PASCAL-Scribble dataset [11].** Comparison of various text-to-image generation methods, including the ControlNet fine-tuned on the training dataset. ScribbleDiff demonstrates superior alignment with the input scribbles, particularly in handling abstract shapes and object orientations. As shown in (f), ScribbleDiff provides the closest representation of the original image (a) Scribbles, effectively capturing both the orientation and the standing posture.

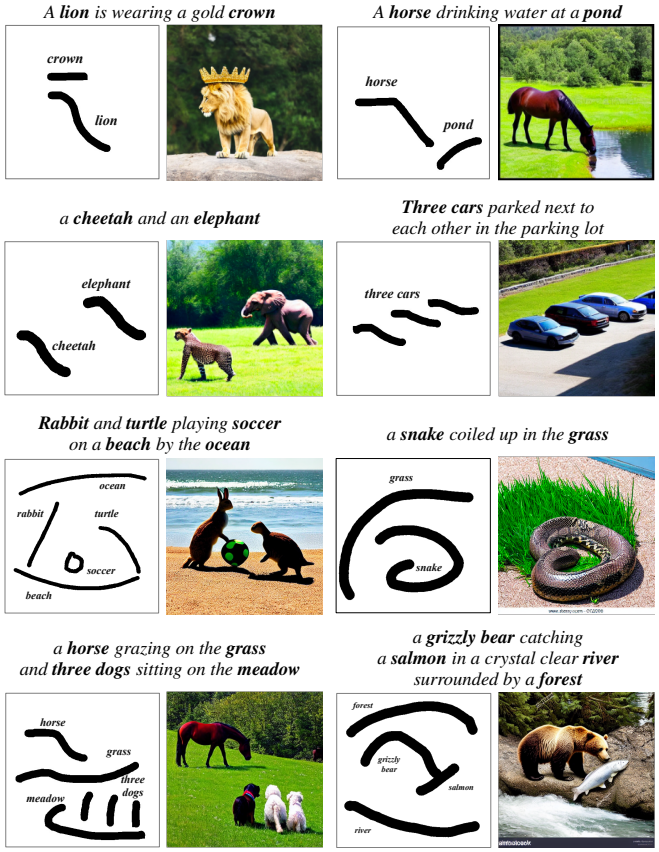


Fig. S5: **Examples of Text-to-Image generation using scribble prompts by ScribbleDiff.** Each row contains two pairs of scribbles and their generated images, with the corresponding prompt placed above each pair. The layout ensures alignment and clarity for each example.

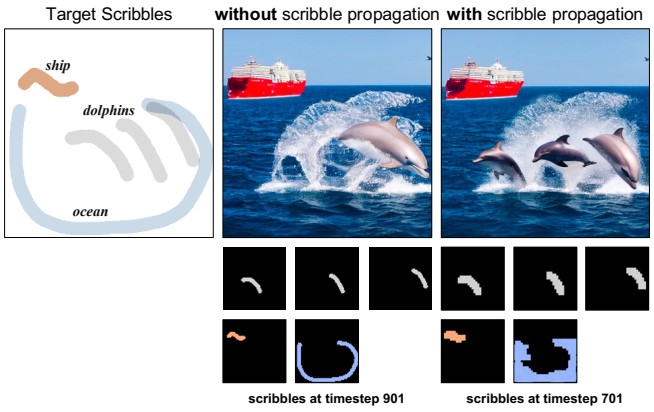


Fig. S6: **Effect of scribble propagation.** With scribble propagation in Stable Diffusion, the scribble expands significantly by timestep, improving object shape and enhancing visual coherence.

through D. Each participant was tasked with completing a total of 30 evaluation questions, as there were three distinct questions associated with each set of 10 samples. An example of the survey is shown in Fig. S9.

Below we include the full questions used for our user study.

- Choose the image that **best reflects the input scribble** (***e.g.***, orientation, abstract shape, and overall spatial alignment of the object with the scribble.)
- Choose the image that best represents the content of the text prompt, considering all key elements described in the text. (***e.g.***, **no key elements in bold are missing** and the generated object is coherent and complete.)
- Choose the image that best balances **reflecting the input scribble** and **accurately representing the content of the text prompt**. (The best image considering both Set 1 and Set 2 criteria.)

The first question aims to assess the generated image's alignment with the input scribble. This measure is crucial for determining how well the model adheres to user-provided visual guides, such as scribbles, which are necessary for customization or specific design constraints. This question evaluates aspects such as orientation, abstract shape, and spatial arrangement.

The second question evaluates how effectively the generated images capture the essence of the text prompt, ensuring that all critical elements highlighted in the prompt are correctly depicted in the generated images. This question is asked to measure the model's capacity not to neglect any necessary key objects, leading to complete representations.

The last question seeks to determine the optimal balanced assessment, which combines the criteria asked in the two previous questions. This is particularly relevant to scenarios where both textual and visual cues must be considered to generate contextually appropriate and visually coherent outputs.

H. LIMITATION & DISCUSSION

This study focuses on improving the incorporation of scribbles as a form of guidance in text-to-image (T2I) generation models, rather than enhancing the overall T2I performance. Future research can explore methods to boost the performance of T2I models directly while maintaining improvements in scribble-based guidance.

In addition to the Bezier Scribbles [9] used in this study, future work could investigate developing models that are robust across various types of sketches, such as Axial Scribbles and Boundary Scribbles. These models should effectively handle different forms of sketch input to improve flexibility in practical applications.

Algorithm 1 Scribble-Guided Diffusion

Input: A diffusion model ϵ_θ with parameters θ, a latent code z_T on timestep T, a scribble $s \in \{0,1\}^{H \times W}$, and a scribble region $\mathcal{S}$ corresponding to scribble s.
Hyperparameters: Timestep interval for scribble propagation $[k_1, k_2]$, weights for moment losses λ_1 and λ_2, and aggregation weights ω_i for each resolution level i.
Output: z_0.

for $t = T, T-1, \ldots, 1$ **do**
  Calculate $\hat{z}_{t-1}$ by Eq. (2)

  # Cross Attention and Moment Loss (Section 3.2)
  # Calculate cross attention loss
  Calculate $\mathcal{L}_{\texttt{focal}}$ by Eq. (5) using $\forall c \in \mathcal{C}(s)$
  Calculate $\mathcal{L}_{\texttt{centroid}}$ by Eq. (6) using $\forall c \in \mathcal{C}(s)$
  Calculate $\mathcal{L}_{\texttt{central}}$ by Eq. (7) using $\forall c \in \mathcal{C}(s)$
  $\mathcal{L}_{\texttt{moment}} = \lambda_1 \mathcal{L}_{\texttt{centroid}} + \lambda_2 \mathcal{L}_{\texttt{central}}$
  $\mathcal{L}_{\texttt{cross}} = \mathcal{L}_{\texttt{focal}} + \mathcal{L}_{\texttt{moment}}$
  # Shift latent code
  $z_{t-1} \leftarrow \hat{z}_{t-1} - \nabla_{z_t} \mathcal{L}_{\texttt{cross}}$

  # Scribble Propagation (Section 3.3)
  if not $(k_1 \le t \le k_2)$ **then**
    continue
  end if
  # Aggregate self-attention maps (as DiffSeg [26])
  for $i, (H, W)$ in `res` **do**
    $\delta \leftarrow H^{\texttt{agg}} / H$
    $\mathcal{A}^{\texttt{new}} \leftarrow$ **Resize** $\left(\mathcal{A}^{H \times W}_{\texttt{self}}, H^{\texttt{agg}} \times W^{\texttt{agg}}\right)$
    for each patch (h, w) in $\mathcal{A}^{\texttt{agg}}$ **do**
      $\mathcal{A}^{\texttt{agg}}[h, w] \mathrel{+}= \omega_i \cdot \mathcal{A}^{\texttt{new}}[h//\delta, w//\delta]$
    end for
  end for
  $\delta_{\texttt{anc}} \leftarrow H^{\texttt{agg}} / H^{\texttt{anchor}}$
  # Region-avg pool aggregated self-attention maps
  $\mathcal{A}^{\texttt{anc}} \leftarrow$ **AvgPool** $(\mathcal{A}^{\texttt{agg}}, \delta_{\texttt{anc}} \times \delta_{\texttt{anc}})$
  for each object o **do**
    $\mathcal{A}^{\texttt{scr}}[o] \leftarrow \frac{1}{\mathcal{S}_o} \sum_{(i,j) \in \mathcal{S}_o} \mathcal{A}^{\texttt{anc}}[i, j]$
  end for
  `MergeNeighbors`$(s, \mathcal{S}, \mathcal{B}^s)$
end for

Algorithm 2 `MergeNeighbors()`

Input: a scribble s, a scribble region $\mathcal{S}$ of s, boundary anchors $\mathcal{B}^s$ of a scribble s.

Initialize $\textbf{dist}_{\texttt{nbr}}$ and $\textbf{obj}_{\texttt{nbr}}$ to ∞ and 0 respectively
for each object o and edge (i, j) in $\mathcal{B}^s$ **do**
  Find neighbors $\mathcal{N}(i, j)$
  for each neighbor $(n_i, n_j) \in \mathcal{N}(i, j)$ **do**
    if neighbor is visited **then**
      continue
    end if
    Calculate distance d using Eq. (9)
    # Select candidates
    # which distances are lower than threshold
    if $d < \tau_{\texttt{dist}}$ **then**
      $\textbf{dist}_{\texttt{nbr}}[n_i, n_j] \leftarrow d$
      $\textbf{obj}_{\texttt{nbr}}[n_i, n_j] \leftarrow o$
    end if
  end for
end for
# Select neighbors with K-highest similarities
$\textbf{indices}_{\texttt{nbr}} \leftarrow \textbf{TopK}(\textbf{dist}_{\texttt{nbr}}, k)$
# Integrate selected neighbors into scribble
for (n_i, n_j) in $\textbf{indices}_{\texttt{nbr}}$ **do**
  $S[\textbf{obj}_{\texttt{nbr}}[\text{idx}] - 1, n_i, n_j] \leftarrow$ **True**
end for

(a) scribbles (b) w/o $\mathcal{L}_{\texttt{moment}}$ (c) w/ $\mathcal{L}_{\texttt{moment}}$

Fig. S7: **Moment Loss.** We show a visual comparison of our approach both with and without moment loss. Notably, in the images labeled (c), where moment loss is applied, the subjects are oriented toward the target direction. This observation clearly indicates that moment loss effectively contributes to the proper alignment of the object's orientation.

(a) $\lambda_1, \lambda_2 = 0.2$ (b) $\lambda_1, \lambda_2 = 0.6$ (c) $\lambda_1, \lambda_2 = 1.0$

Fig. S8: **Change in image as the scale λ_1 and λ_2 changes.** As the values of λ_1 and λ_2 increase, the generated image increasingly aligns with the scribble input. This is evident in the images from left to right, where the shape of the *bamboo raft* progressively conforms to the thin scribble, and the orientation of the *cute panda* shifts from facing forward to the left, as specified by the input scribble.

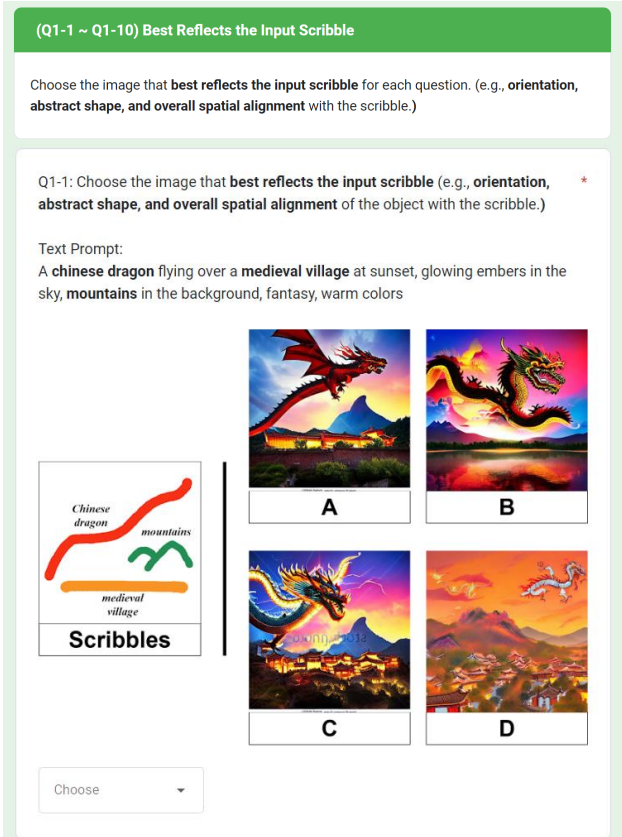


Fig. S9: **Screenshot of our user study.** Participants were asked to compare images from four methods, including our approach.