A. Additional Material: Presentation Video

We have described our results in an easily accessible manner on our project page, where a brief **presentation video** is also available. The link to the **project page** is as follows: https://micv-yonsei.github.io/storm2025/.

B. Implementation Details

We discuss the hyperparameter settings and selection of object and attribute tokens (Section B.1 and Section B.2).

B.1. Hyperparameter Details

In this section, we detail the hyperparameter settings used in our implementation, ensuring alignment with previous models for a fair comparison. We adopt a scale factor of 20 and a scale range of (1.0, 0.5), consistent with prior work, to maintain uniformity in updating the denoised latent z_t across denoising steps. Similarly, the Gaussian smoothing parameters, such as the standard deviation (σ) of 0.5 and the kernel size of 3, are set identically to those in other models. Optimization is applied only for the first 25 timesteps, as in prior work, to prevent quality degradation in the generated image. This ensures that modifications primarily enhance spatial awareness during the critical early denoising stages while preserving overall image fidelity in later stages. For timesteps $t \in 5, 10, 15, 20$, additional iterations are performed during optimization if the specific target values of 0.05, 0.01, 0.005, and 0.001 are not achieved. The optimization process ensures that the model converges toward these precise thresholds, with a maximum of 30 iterations allowed for each timestep.

B.2. Selection of Object and Attribute Tokens

We utilize a part-of-speech (POS) tagger to extract nouns (object tokens) and adjectives (attribute tokens) from the given prompt. Additionally, users have the flexibility to manually specify tokens of interest, a method consistent with approaches employed in previous studies [2, 4, 9, 10], allowing for further customization and refinement based on specific requirements. These tokens are then analyzed through attention maps to ensure the model focuses more effectively on the identified tokens. For tokens conveying positional information (*e.g.,* "on the left," "next to," "above"), the model leverages the extracted spatial context from the text prompt to guide its operations. The explicitly stated positional information dynamically adjusts the attention maps of both object and attribute tokens, ensuring that spatial relationships in the prompt are accurately reflected in the generated output.

C. Method Details

In this section, we discuss details of reference point and target distribution (Section C.1 and Section C.2, respectively), details of ST Cost (Section C.3), Sinkhorn algorithm-based Transport Plan (Section C.4), and algorithm of our method (Section C.5).

C.1. Reference Centroid Positioning

The reference point is defined as the centroid of the attention map for the relative object, serving as an anchor for spatial adjustments. This centroid guides the placement of the target distribution, ensuring that the source distribution aligns accurately with the desired spatial relationship. The centroid coordinates along the horizontal and vertical dimensions are computed as follows:

$$j_{\text{A}} = \frac{\sum_{j=0}^{n-1} j \cdot \left(\sum_{i=0}^{n-1} A_{ij}\right)}{\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} A_{ij}},\ i_{\text{A}} = \frac{\sum_{i=0}^{n-1} i \cdot \left(\sum_{j=0}^{n-1} A_{ij}\right)}{\sum_{i=0}^{n-1}\sum_{j=0}^{n-1} A_{ij}}. \tag{1}$$

Here, A_{ij} represents the attention value at position (i, j) in the attention map A. The computed values, j_A and i_A, correspond to the centroid positions along the horizontal and vertical axes, respectively, obtained as weighted averages over each dimension.

C.2. Target Distribution

The target distribution is an arbitrary distribution representing the desired position of the source distribution relative to the reference point. To model this, we adopt a circular Gaussian distribution, providing a probabilistic representation of the object's spatial presence. This formulation allows the distribution to adapt dynamically based on the specified spatial constraints at each timestep. The reference point determines the centroid of the Gaussian distribution (see Section C.1 for details on reference point computation). To compute this centroid, we consider the relative spatial relationship (left, right, above, below) with respect to the reference point $(j_{\text{ref}}, i_{\text{ref}})$. The centroid is defined as follows, where N represents the size of one dimension of the image: $c^{\leftarrow} = (\frac{0+j_{\text{ref}}}{2}, \frac{N}{2}), c^{\rightarrow} = (\frac{N+j_{\text{ref}}}{2}, \frac{N}{2}), c^{\uparrow} = (\frac{N}{2}, \frac{0+i_{\text{ref}}}{2}), c^{\downarrow} = (\frac{N}{2}, \frac{N+i_{\text{ref}}}{2})$. Each arrow corresponds to a specific spatial direction. After computing the centroid of the target Gaussian distribution, the final target distribution is defined as:

$$D(x, y) = \exp\Big(-\frac{(x-c_x)^2 + (y-c_y)^2}{2\sigma^2}\Big), \tag{2}$$

where c_x and c_y denote the centroid coordinates along the horizontal (x) and vertical (y) dimensions, respectively.

C.3. Details of ST Cost

Details of ω. In Equation 2 of the main paper, ω represents the progressive adaptive weights in the Spatial Transport (ST) cost function. It controls the trade-off between

aligning the source distribution in the desired direction and penalizing movement in restricted directions. ω is defined as:

$$\omega(t) = 1 + (\omega_{max} - 1)(1 - e^{-kt}), \tag{3}$$

where t represents the timestep and ω_{max} is the maximum weight value, and k is a constant controlling the rate of increase. In our experimental setting, we set ω_{max} to 100. A higher ω emphasizes precise alignment in the desired direction, potentially sacrificing flexibility, whereas a lower ω may allow more flexibility but reduce positional accuracy. It gradually increases in later steps to enforce precise positioning and ensure smooth movement that maintains image quality. Starting with a lower ω for flexibility, it gradually increases in later steps to enforce precise positioning and ensure smooth movement that maintains image quality. Further details on the selection of ω values are provided in the ablation studies (Section E.1.1).

Details of Cost. Since C_{ij} (refer to Equation 3 in the main paper) is evaluated only along a specific dimension, we extend it to the other axis by integrating $\mathbf{1}_N$ along that dimension. This extension is formulated as $\mathbf{C}^{\text{st}} = C_{\text{flat}} \otimes \mathbf{1}_N$, where C_{flat} represents the flattened version of C_{ij} (refer Equation 3 in the main paper) and N is the number of patches ($n \times n$). To ensure an optimal transport plan that accounts for positional relationships, we also incorporate the standard OT cost matrix, following traditional OT formulations to quantify the spatial cost of transporting mass between distributions. To construct this matrix, we compute the p-norm distance between all possible pairs of points in the source and target distributions. First, the source and target distributions are represented as 2D grids of dimensions H (height) and W (width). Each grid is then converted into a list of patch coordinates, (i, j) for the source and (k, l) for the target, capturing all possible spatial locations. For each pair of coordinates (i, j) and (k, l), we compute the p-norm distance defined as:

$$d_p((i, j), (k, l)) = (|i - k|^p + |j - l|^p)^{1/p} \tag{4}$$

The cost matrix $\mathbf{C}^{\text{dist}}$ is constructed by taking the p-th power of these distances, resulting in :

$$\mathbf{C}^{\text{dist}}_{uv} = (|i_u - k_v|^p + |j_u - l_v|^p), \tag{5}$$

where $\mathbf{C}^{\text{dist}}_{uv}$ is the cost of transporting mass from the u-th coordinates in the source to the v-th coordinate in the target. Finally, the overall cost matrix is defined as:

$$\mathbf{C} = \lambda \mathbf{C}^{\text{dist}} + (1 - \lambda)\mathbf{C}^{\text{st}}, \tag{6}$$

where $\lambda = 0.01$ is chosen to minimize the influence of uncertainty in the target distribution.

Optimization for Objects. We simplify by removing position-based terms, focusing solely on ensuring that objects do not overlap. This is expressed as $\mathbf{C}^{\text{st}} = A_{\text{flat}} \otimes \mathbf{1}_N$, while $\mathbf{C}^{\text{dist}}$ operates in the same manner as described above.

C.4. Sinkhorn Algorithm-based Transport Plan

Once the cost function is established, the next step is to compute a transport plan $\mathbf{P}$ that minimizes this cost. To solve this OT problem efficiently, we employ the Sinkhorn algorithm, an iterative approach that introduces an entropic regularization term to the standard OT objectives. The regularization term ensures that $\mathbf{P}$ becomes more evenly distributed and computationally stable, particularly for high-dimensional attention maps. The cost matrix $\mathbf{C}$ obtained from the customized cost function (ST Cost), encodes the spatial alignment objectives between the source and target distributions. For instance, $\mathbf{C}_{uv}$ represents the alignment cost between source position u and target position v. The regularized OT problem is formulated as:

$$\min_{\mathbf{P} \geq 0} \quad \sum_{u=1}^{N} \sum_{v=1}^{N} \mathbf{C}_{uv} \mathbf{P}_{uv} - \lambda \sum_{u=1}^{N} \sum_{v=1}^{N} \mathbf{P}_{uv} (\log \mathbf{P}_{uv} - 1), \tag{7}$$

where λ controls the strength of the entropic regularization. The transport plan $\mathbf{P}$ is initialized as a uniform matrix and iteratively refined to satisfy the marginal constraints defined by the source (A) and target (B) distributions. To satisfy the row and column marginals, A and B, the Sinkhorn algorithm alternately updates the scaling vectors u and v as follows:

$$\mathbf{u} \leftarrow \frac{A}{\mathbf{P}\mathbf{v}}, \mathbf{v} \leftarrow \frac{B}{\mathbf{P}^\top \mathbf{u}}. \tag{8}$$

At each step, these updates ensure that the rows and columns of $\mathbf{P}$ sum to the respective marginal distributions. The algorithm iterates until the constraints are satisfied within a predefined tolerance. To ensure spatial consistency, the cost function is applied bi-directionally. If the position of A is adjusted using B as a reference, the reverse operation is also performed, adjusting B using A.

C.5. Algorithm

Algorithm 1 provides an overview of the denoising process using STORM, which includes the update and optimization process.

D. Evaluation Metrics and Datasets

This section provides a comprehensive overview of the three primary metrics [3, 6, 7], along with the details of the user studies presented in the main paper. Additionally, it includes descriptions of the datasets used for calculating each metric.

Algorithm 1 A Denoising Step using STORM

Input:
- A text prompt $\mathcal{P}$
- Attention map keys $\mathcal{K} = \{\text{source}, \text{reference}\}$
- Current timestep t
- Iterations for refinement $\{t_1, \ldots, t_k\}$
- Thresholds $\{T_1, \ldots, T_k\}$
- Trained Stable Diffusion model SD

Output:
- A noised latent z_{t-1} for the next timestep

$\_, A_t \leftarrow \text{SD}(z_t, \mathcal{P}, t)$ ▷ Obtain attention map A_t from SD.
$A_t \leftarrow \text{Softmax}(A_t - \langle\text{sot}\rangle)$ ▷ Apply softmax to exclude special tokens.
$\mathcal{A} \leftarrow \{\}$ ▷ Initialize attention map dictionary.
$\mathcal{C} \leftarrow \{\}$ ▷ Initialize centroid dictionary.
for $k \in \mathcal{K}$ **do** ▷ Process all specified attention keys.
  $\mathcal{A}[k] \leftarrow A_t[:, :, k]$ ▷ Extract attention map for key k.
  $\mathcal{C}[k] \leftarrow \text{ComputeCentroid}(\mathcal{A}[k])$ ▷ Compute centroid for key k.
end for
$A_t^{\text{source}} \leftarrow \mathcal{A}[\text{source}]$
$c^{\text{source}} \leftarrow \mathcal{C}[\text{source}]$
$A_t^{\text{ref}} \leftarrow \mathcal{A}[\text{reference}]$
$c^{\text{ref}} \leftarrow \mathcal{C}[\text{reference}]$
$D_{\text{target}} \leftarrow \text{Gaussian}(c_{\text{ref}})$ ▷ Target Distribution based on Gaussian or Spatial Prior aligned with c^{ref}.
$\mathcal{C} \leftarrow \text{ST Cost}(A_t^{\text{source}}, c^{\text{ref}}, D_{\text{target}})$ ▷ Compute cost matrix using centroids.
$\mathcal{T} \leftarrow \text{Sinkhorn}(A_t^{\text{source}}, D_{\text{target}}, \mathcal{C})$ ▷ Compute transport plan.
$\mathcal{L} \leftarrow \sum \mathcal{T} \cdot \mathcal{C}$ ▷ Calculate st loss.
$z_t' \leftarrow z_t - \alpha_t \cdot \nabla_{z_t} \mathcal{L}$ ▷ Update latent z_t using gradient.
if $t \in \{t_1, \ldots, t_k\}$ **then** ▷ Check if iterative refinement is needed.
  if $\mathcal{L} > 1 - T_t$ **then** ▷ Compare loss against threshold T_t.
    $z_t \leftarrow z_t'$
    Go to Step 1
  end if
end if
$z_{t-1}, \_ \leftarrow \text{SD}(z_t', \mathcal{P}, t)$ ▷ Obtain updated latent z_{t-1}.
Return z_{t-1}

D.1. VISOR metric

The VISOR (Verifying Spatial Object Relationships) evaluates the spatial reasoning capabilities of T2I models by assessing how accurately they generate images that reflect the spatial relationships described in text prompts. The key components of VISOR are defined as follows:

Object Accuracy (OA) Object Accuracy (OA) measures whether both objects specified in the text prompt are present in the generated image. It is computed as $OA(x, A, B) = \mathbb{1}_{h(x)}(\exists A \cap \exists B)$, 1 if both A and B are detected in image x, otherwise 0. OA measures object presence using OWL-ViT [11], a pre-trained open-vocabulary object detector.

VISOR$_{\text{uncond}}$ The VISOR$_{\text{uncond}}$ evaluates spatial correctness by determining whether the generated spatial relationship aligns with the ground truth relationship specified in the text prompt. It assesses both the presence of the objects in the generated image and whether their spatial arrangement accurately reflects the prompt description.

$$\text{VISOR}_{\text{uncond}}(x, A, B, R) = \begin{cases} 1, & \text{if } (R_{\text{gen}} = R) \wedge (\exists A \cap \exists B), \\ 0, & \text{otherwise,} \end{cases} \tag{9}$$

where R_{gen} indicates the spatial relationship detected between objects in the generated image, and R denotes the ground truth relationship specified in the text prompt. The term $\exists A \cap \exists B$ indicates that both objects A and B are detected in the image. This metric provides a holistic evaluation of the model's ability to both generate objects and accurately position them according to the specified spatial relationships. Unlike metrics that strictly require the detection of both objects, VISOR$_{\text{uncond}}$ captures a more comprehensive view of the model's real-world performance. For example, given the prompt "A cat to the left of a dog", if the generated image contains both a cat and a dog with a cat positioned correctly to the left of a dog, then: VISOR$_{\text{uncond}} = 1$. Otherwise, VISOR$_{\text{uncond}}$ = 0.

VISOR$_{\text{cond}}$ This metric evaluates spatial correctness only when both objects are correctly generated in the image. The spatial relationship is determined using centroid-based rules (*e.g.,* $x_A < x_B$ implies A is to the left of B). For example, given the prompt "A cat to the left of a dog," only generated images that contain both a cat and a dog are considered for evaluation. If the cat is correctly positioned to the left of the dog, then VISOR$_{\text{cond}} = 1$. Otherwise, 0.

VISOR$_n$ VISOR$_n$ measures the probability of generating a least n spatially correct images for a given text prompt when multiple images are generated. If reflects a model's practical utility for users who select from multiple outputs, capturing its consistency in producing spatially accurate generations.

SR2D Dataset The SR2D (Spatial Relationships in 2D) dataset is specifically curated to evaluate spatial reasoning in T2I models. It contains 25,280 text prompts describing spatial relationships (*e.g.,* left, right, above, below) between pairs of objects. The objects are drawn from 80 categories based on the MS-COCO dataset. Prompts are generated using predefined templates (*e.g.,* "A [object A] to the left of a [object B]") to ensure linguistic clarity and consistency. Spatial relationships are uniformly represented

across all object pairs, providing a standardized evaluation framework. For each prompt, multiple images are generated and assessed using VISOR and related metrics, offering insights into model performance on spatial reasoning tasks.

D.2. T2I-CompBench

We evaluate spatial relationships and attribute binding through T2I-CompBench Framework [7], which provides a comprehensive evaluation of T2I synthesis performance.

Spatial Alignment Spatial relationships serve as a key sub-category for evaluating T2I synthesis. The benchmark defines spatial relationships between objects using terms such as left, right, top, bottom, next to, near, and on the side of. For "left", "right", "top," and "bottom", spatial relationships are evaluated by comparing the relative positions of the centers of bounding boxes for two objects in the generated image. Specifically, an object A is considered to be on the left of object B if: $x_1 < x_2, |x_1 - x_2| > |y_1 - y_2|$,and mIoU < 0.1, where and (x_1, y_1) and (x_2, y_2) represent the center coordinates of objects A and B, respectively. For "near to", "near", and "on the side of," these relationships are determined based on the distances between bounding box centers of two objects relative to a predefined threshold. To detect objects and determine their spatial positions, UniDet [14], a pre-trained object detection model, is utilized.

Attribute Binding Attribute binding in T2I-CompBench evaluates whether attributes such as color, shape, and texture are correctly associated with the corresponding objects in the generated images.

- Texture Binding: Assesses the model's ability to associate texture descriptors (*e.g.,* "fluffy," "metallic") with the correct objects. Prompts such as "A rubber ball and a plastic bottle" test texture-related attribute binding. Texture descriptors are generated from predefined attributes, including "wooden", "glass", and "fabric".
- Color Binding: Evaluates whether colors are correctly assigned to the objects mentioned in the prompt. For example, the prompt "A blue backpack and a red bench" tests whether the correct colors are applied to the respective objects. Color confusion is a common issue when multiple objects and attributes coexist within a prompt.
- Shape Binding: Focuses on correctly binding shape descriptors (*e.g.,* "rectangular", "circular") to objects. Prompts such as "An oval sink and a rectangular mirror" evaluate shape-related accuracy. Shape descriptors include common geometric terms such as "cubic," "pyramidal," and "circular".

The evaluation utilizes the BLIP-VQA [8] model for a fine-grained assessment of object attribute alignment. BLIP-VQA takes a generated image as input and answers questions about object-attribute pairs (*e.g.,* "A green bench?", "A red car?"). The model assigns probabilities to each answer ("Yes" or "No"), which are used to compute an overall attribute-binding score. The final score is calculated as the product of the probabilities for all attribute-related questions: score $= P(\text{'A green bench?'}) \times P(\text{'A red car?'})$. T2I-CompBench systematically evaluates the model's capability to handle both spatial relationships and attribute binding by providing structured text prompts and analyzing whether the generated images meet the specified constraints.

Dataset T2I-CompBench is a benchmark consisting of 6,000 text prompts generated using predefined templates and ChatGPT [1]. Each sub-category (*e.g.,* Color, Shape, Texture) includes 1,000 prompts, with 700 used for training and 300 for testing.

D.3. TIFA

The TIFA (Text-to-Image Faithfulness Evaluation) metric [6] is designed to measure the alignment between generated images and their corresponding input text prompts. Unlike traditional metrics such as the FID score, which primarily evaluates the visual quality of images, TIFA emphasizes semantic consistency, assessing whether the content of an image faithfully represents the objects, attributes, and relationships described in the text. TIFA operates by generating targeted questions based on the input text, leveraging large language models such as LLaMA2 [13] to identify key objects, attributes, and spatial relationships. These questions are designed to ensure alignment between the image and the prompt. For example, given the text prompt "A red apple to the left of a green mug," the model generates queries such as, "What color is the apple?" or "What object is on the left of the mug?". The generated questions are then directed at the output image using a Visual Question Answering (VQA) system. TIFA typically employs advanced VQA models, such as Owl-ViT [11] or BLIP (Bootstrapped Language-Image Pretraining) [8], to extract objects and attributes from the image and provide answers to the posed questions. At this stage, the system maps detected objects and their attributes in the image to the corresponding text-based questions, ensuring semantically relevant responses. Finally, TIFA evaluates the degree of alignment between the generated answers and the expected responses inferred from the input text. High semantic accuracy results in higher scores, while inconsistencies lead to lower scores. Through this process, TIFA quantitatively measures the semantic fidelity between text and image, offering a robust assessment of how well a model adheres to textual descriptions during image generation.

Dataset. The dataset used to compute this metric is based on the same datasets utilized by preceding models [2, 4, 9, 10]. The text prompts fall into three categories: (1) "a [animal A] and a [animal B]", (2) "a [animal] and a [color][object],", and (3) "a [colorA][objectA] and a [colorB][objectB]". These prompts are constructed using 12 animals, 12 objects, and 11 colors. Each prompt incorporates a subject-color combination, with colors randomly assigned to each subject. This process results in 66 combinations for animal-animal and object-object pairs, along with 144 animal-object pairs. Each prompt is then used to generate 64 images with 64 random seeds, ensuring a diverse evaluation of model performance.

D.4. User Studies

We conducted a user study to evaluate our STORM model based on its ability to generate images that align with detailed text prompts. We created 10 custom prompts, each describing specific objects, attributes, and spatial relationships. Using different random seeds, we generated corresponding images with various T2I training-free models. These images were evaluated by 30 participants, who rated them on a scale from 1 (lowest) to 5 (highest) across four criteria: (1) object accuracy, (2) attribute matching, (3) spatial correctness, and (4) overall fidelity. The total score for each model within a given criterion was obtained by summing the scores across all participants. To compare performance, we calculated the percentage score of the i-th model as the ratio of its total score of all models evaluated within that criterion, using the following formula:

$$\text{Percentage Score } (\%) = \left(\frac{S_i}{\sum_{j=1}^{N} S_j} \right) \times 100 \qquad (10)$$

where S_i represents the total score of the i-th model within a given criterion, and $\sum_{j=1}^{N} S_j$ is the sum of the total scores for all models within that criterion, and N is the total number of models. For example, in the spatial correctness criterion, the total scores were 460 for SD, 514 for Attend&Excite, 507 for Divide&Bind, 512 for INITNO, s for CONFORM, and 1399 for STORM. The percentage score for STORM in this criterion was calculated as:

$$\frac{1399}{460 + 514 + 507 + 512 + 503 + 1399} \times 100 \approx 35.92\% \qquad (11)$$

This process was repeated for each criterion and model, with results rounded to the third decimal place.

E. Additional Experiments

E.1. Additional Ablation Study

We construct an additional ablation study on varying ω values (Section E.1.1) and the effects of applying STO for shorter durations (Section E.1.2).

Table 1. Ablation study on ω values, comparing fixed settings ($\omega = 1, 50, 100$) with our dynamically adjusted $\omega(t)$, evaluated on OA(%) and VISOR metrics. It shows lower ω performs poorly, while moderate and high fixed ω improve alignment. Our dynamic $\omega(t)$ achieves the best performance by balancing flexibility in early timesteps and precision in later timesteps.

Values of ω	OA (%)	VISOR					
		uncond	cond	1	2	3	4
1	39.05	33.57	85.98	66.45	41.40	20.09	6.52
50	60.73	55.65	91.64	83.81	69.09	47.82	21.91
100	58.67	54.67	93.18	83.62	67.74	45.89	21.63
Ours	**61.01**	**57.58**	**94.39**	**85.93**	**69.71**	**49.01**	**25.70**

E.1.1. Ablation Studies for ω

This ablation study examines the impact of using fixed values of ω compared to the dynamic adjustment employed in our approach. In our STORM model, ω is dynamically updated across timesteps, as defined in Eq. (3). This function enables ω to gradually increase throughout the diffusion process, maintaining a balance between spatial flexibility in the early steps and precise alignment in later steps. Table 1 presents an evaluation of VISOR [3] under different values of ω. The first row, which corresponds to a low ω, demonstrates poor performance in both OA(%) and VISOR metrics, particularly in VISOR_4. The moderate $\omega = 50$ and $\omega = 100$ show the improvement in OA and VISOR metrics compared to low ω. However, these fixed values of ω fail to capture the optimal balance across timesteps, as seen in the lower scores for VISOR_4 when compared to our dynamically adjusted $\omega(t)$. Specifically, a high $\omega = 100$ overpenalizes deviations, reducing the flexibility needed in early timesteps, while moderate $\omega = 50$ does not provide sufficient precision in later timesteps. In contrast, our dynamically adjusted $\omega(t)$ achieves the best performance across all metrics. By gradually increasing ω throughout the diffusion process, our model effectively balances early-stage flexibility with late-stage spatial accuracy. This dynamic adjustment leads to superior results, by the significant improvements in VISOR_4 (25.70%) and $\text{VISOR}_{\text{uncond}}$ (57.58%).

E.1.2. Ablation Studies for applying STO through Timestep

In Table 2, we present an ablation study evaluating the effect of applying STO over different timestep ranges. In the main paper, STO was applied during the later stages of generation, specifically in the ranges 19–24, 13–24, 7–24, and 1–24, focusing on its impact when image details are refined (see Fig. 3 for more results). Here, we shift our attention to earlier timesteps, applying STO in the ranges 1-6 (Exp.#1),

Table 2. Ablation study on the impact of applying STO at different timesteps. Exp.#A0 represents the baseline results from SD without STO. From Exp.#A1 to Exp.#A4, STO is progressively applied over increasing timestep ranges: 1–6, 1–12, 1–18, and 1–24.

#Exp.	OA (%)	VISOR					
		uncond	cond	1	2	3	4
0 (SD)	29.86	18.81	62.98	46.60	20.11	6.89	1.64
1	49.00	43.45	88.67	75.92	53.53	31.70	12.71
2	56.17	51.33	91.37	81.62	63.56	41.35	18.92
3	59.05	54.30	91.96	82.73	66.29	45.64	22.74
4 (Ours)	**61.01**	**57.58**	**94.39**	**85.93**	**69.71**	**49.01**	**25.70**

1-12 (Exp.#2), 1-18 (Exp.#3), and 1-24 (Exp.#4). This allows us to examine its effectiveness during the early stages of generation, where the model primarily establishes the structural layout and ensures broader spatial consistency. As shown in Table 2, optimizing over a longer timestep range yields better results than optimizing over a smaller range. However, when analyzing the overall VISOR scores, we observe that they are significantly higher than those presented in Table 4 of the main paper. This suggests that applying STO during the early timesteps is particularly beneficial as it enables better spatial adjustments in the initial stages of generation, ultimately leading to improved overall performance.

E.2. Additional Qualitative Results

E.2.1. Synergy with Stronger Text Encoder

Powerful text encoders have been proposed to address spatial alignment in T2I synthesis, with ELLA [5] being a prominent example. We leverage STORM's training-free characteristic to integrate it with ELLA [5], and we experimentally validate the effectiveness of this combination on the VISOR benchmark. As shown in Table 3, adding STORM (training-free) to ELLA achieves noticeably better results. Fig. 2 further provides a visual demonstration of this improvement, illustrating how our training-free approach can complement advanced text-encoder-based methods for spatial alignment.

Table 3. Quantitative results on the VISOR benchmark by combining STORM with ELLA.

model	OA	VISOR (cond)	VISOR (uncond)
SD 1.5	28.49	62.94	17.93
SD 1.5 + ELLA (fixed)	52.7	67.31	35.48
SD 1.5 + ELLA (flexible)	54.33	67.51	36.68
SD 1.5 + **STORM**	**62.03**	**90.82**	**56.33**

E.3. Additional Visualization

We provide additional visualizations to further support our results. These include comparisons between Stable Diffusion (SD) and our method on the SR2D dataset (Section E.2.1), qualitative results for both SD 1.4 and SD 2.1 (Section E.2.2), visualizations of attention maps across denoising timesteps (Section E.2.3), additional ablation visualizations illustrating the effect of applying STO during the denoising process (Section E.2.4), and positional variations observed within the same seed (Section E.2.5).

Note. Full-page figures are placed at the bottom of the document.

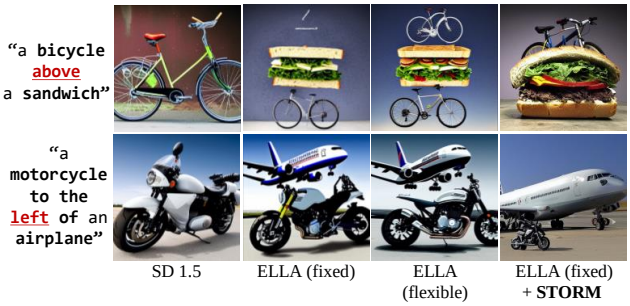

Figure 2. Comparison of results from ELLA and ELLA + STORM using SD 1.5.

E.3.1. Comparison between SD and Ours

As illustrated in Fig. 7, we provide additional visualization on stable diffusion [12] and ours using SR2D Dataset [3]. Our model, STORM, demonstrates a remarkable ability to accurately position objects in the desired locations.

E.3.2. Additional Qualitative Results

In Fig. 8 and Fig. 9, we present the qualitative comparisons between Stable Diffusion 1.4 [12] and other state-of-the-art methods [2, 4, 9, 10]. Our model excels in accurately matching attributes while ensuring that all objects are distinctly generated without overlaps. Moreover, unlike other methods that often struggle with positional accuracy, our approach consistently maintains precise spatial arrangements, demonstrating superior performance. To further validate our findings, we conducted the same qualitative comparisons across all methods using Stable Diffusion 2.1 [12]. As shown in Fig. 10 and Fig. 11, our method continues to exhibit strong spatial understanding, regardless of the model version. Additionally, it effectively mitigates object overlap issues, a common weakness in Stable Diffusion, further highlighting its robustness in generating well-structured outputs.

E.3.3. Additional Visualization for attention map

We provide extended visualization of attention map progression throughout the denoising process in Fig. 12 and

Fig. 13. The figure on the far left shows the attention map at the initial stages of the denoising process, while the subsequent figures to the right illustrate the attention maps as the denoising progresses through later steps. As shown in the figures, although both models start with the same noise distribution, they gradually exhibit different patterns. Notably, our model exhibits a clear tendency to focus on regions requiring refinement, ensuring a precise distribution in those areas. In contrast, Stable Diffusion often displays scattered attention distributions, with some cases showing complete dissipation of attention in certain regions.

E.3.4. Applying STO During the Denoising Process

In Fig. 3, we provide additional visualizations for the ablation study, demonstrating the impact of applying STO during the denoising process.

E.3.5. Additional Visualization of Positional Variations

To further validate the effectiveness of our method in understanding and reflecting spatial prompts, we provide additional examples demonstrating the model's spatial awareness across various object combinations within the spatial relationships. As shown in Fig. 4, SD exhibits limited spatial awareness, often generating nearly identical images regardless of the given spatial prompts. In contrast, our model effectively captures and preserves the specified spatial relationships, demonstrating a superior understanding of spatial constraints.

E.4. Experiments on Complex, Diverse, and 3D Positional Prompts

To further evaluate the robustness and versatility of our model, we conducted additional experiments using (a) complex prompts and (b) diverse positional prompts, as shown in Fig. 5. Complex prompts involve three or more spatial relationships, requiring the model to interpret and generate objects accurately while maintaining overlapping or hierarchical spatial constraints. These experiments demonstrate that our method effectively captures spatial alignment across a wide range of challenging and diverse scenarios. Additional results include (a) Complex Prompts, where prompts contain three or more spatial relationships, and (b) Diverse Positional Prompts, which extend beyond left, right, above, and below to include diagonal spatial relationships. Our model successfully captures these intricate spatial constraints, consistently outperforming Stable Diffusion. We also generate images using 3D positional prompts, as shown in Fig. 14. Although our approach is fundamentally designed for 2D spatial reasoning, resulting in slightly fewer natural outcomes compared to 2D scenarios, it significantly outperforms other models that entirely disregard positional cues, demonstrating significantly better generation quality.

F. Discussion

F.1. Failure Cases

Despite achieving remarkable performance in spatial alignment, our method faces challenges with extremely rare object-attribute combinations and positional prompts requiring three-dimensional spatial reasoning. These difficulties arise from the training-free nature of our approach, which inherently suffers from data biases and lacks exposure to such uncommon scenarios. For instance, as shown in the second row of Fig. 6, the model effectively generates common objects like a "yellow bus". However, it struggles with rarer combinations, such as a "blue strawberry," resulting in either failed generations or outputs with significantly lower image quality. While our method is capable of producing a "blue strawberry," it exhibits a noticeable degradation in overall image fidelity. Similarly, as seen in the first row of Fig. 6, placing a "zebra" on the top of an "umbrella" leads to an unnatural and awkward composition, highlighting the difficulty of generating plausible outputs for spatially improbable scenarios. Furthermore, our model is currently designed to reason within a 2D space, effectively capturing relationships such as left, right, above, and below. However, since 3D positional cues are not explicitly considered, the generations for 3D prompts can sometimes appear less natural or accurate compared to their 2D counterparts (see Fig. 14). Despite these limitations, our method consistently outperforms others that do not account for spatial relationships, delivering superior results overall.

F.2. Future Works

Our proposed STORM framework effectively mitigates spatial misalignment in training-free T2I synthesis, paving the way for several promising research directions. One key avenue for future work is extending STORM to support multimodal inputs, such as integrating audio or video cues within text-based prompts. This would enhance the model's adaptability across diverse creative applications. Additionally, optimizing the computational efficiency of STO could enable real-time applications, including interactive art and game design. Another promising direction involves developing methods for dynamically optimizing image generation based on immediate user input, allowing greater flexibility and responsiveness. While this study primarily focuses on relative positioning (e.g., left, right, above, below), future research could explore more complex spatial relationships, such as 3D spatial reasoning and multi-object interactions. Although our model already demonstrates strong performance in these areas, we believe there is substantial potential for further advancements that could push the boundaries of spatially aware text-to-image generation.

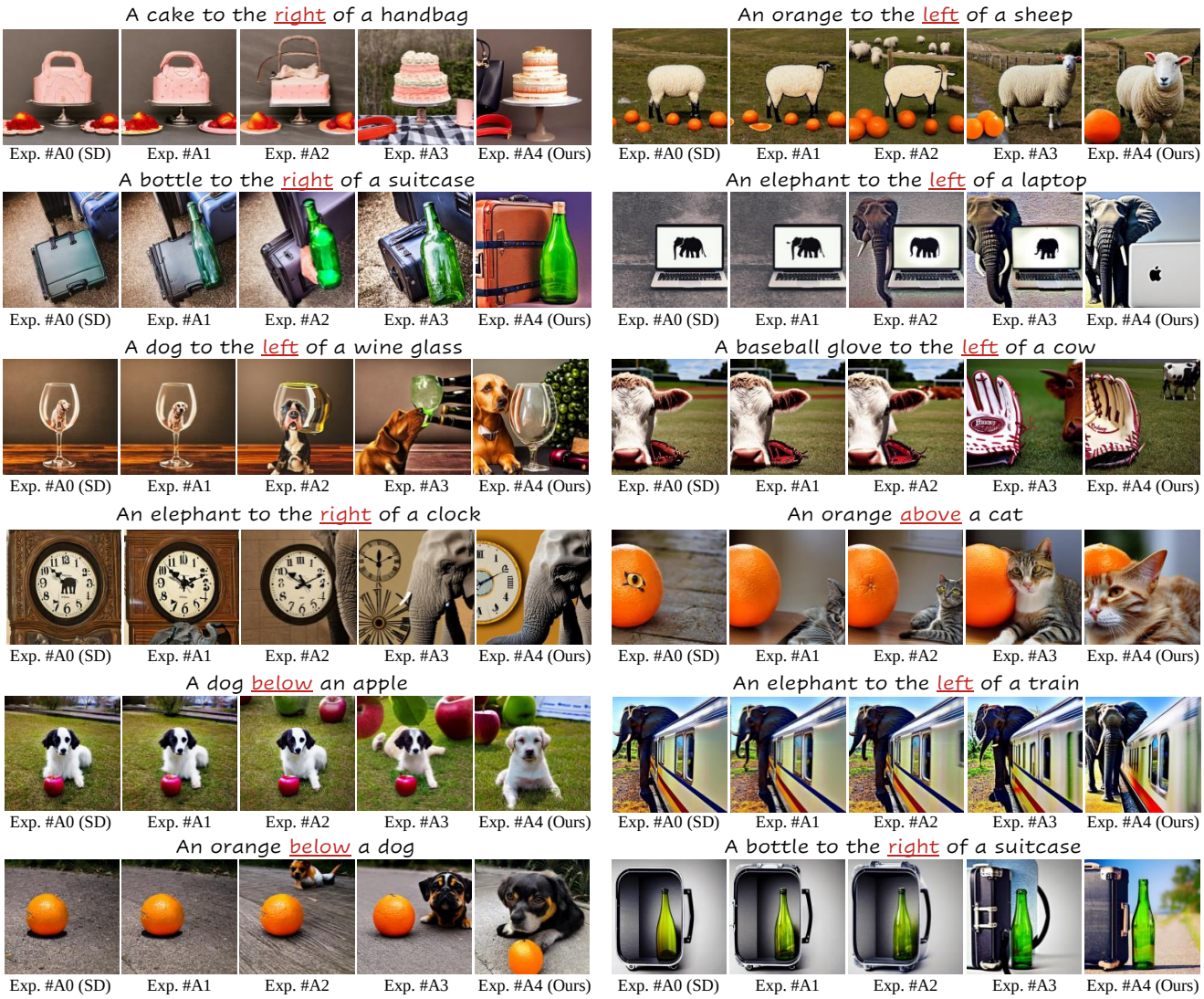

Figure 3. Additional comparison of results when applying STO at different timesteps. Experiments are organized as follows: no STO (Exp.#A0), STO applied from timesteps 19-24 (Exp.#A1), 13-24 (Exp.#A2), 7-24 (Exp.#A3), and 1-24 (Exp.#A4). As seen in the images, earlier STO application improves object positioning and reduces overlap, resulting in more accurately positioned objects.

Figure 4. Additional Comparison of Spatial Awareness. { position* } in each prompt denotes the spatial relationship in each column (*e.g.,* "to the left of", "to the right of", "above," and "below").

(a) Complex Prompts

(b) Diverse Positional Prompt

Figure 5. Additional results on (a) Complex Prompts: prompts with three or more spatial relationships, and (b) Diverse Positional Prompts: including not only left, right, above, and below but also diagonal spatial relationships. Our model successfully captures these complex and diverse spatial constraints, outperforming Stable Diffusion.

Figure 6. Limitations. Difficulty handling rare combinations of objects and attributes.

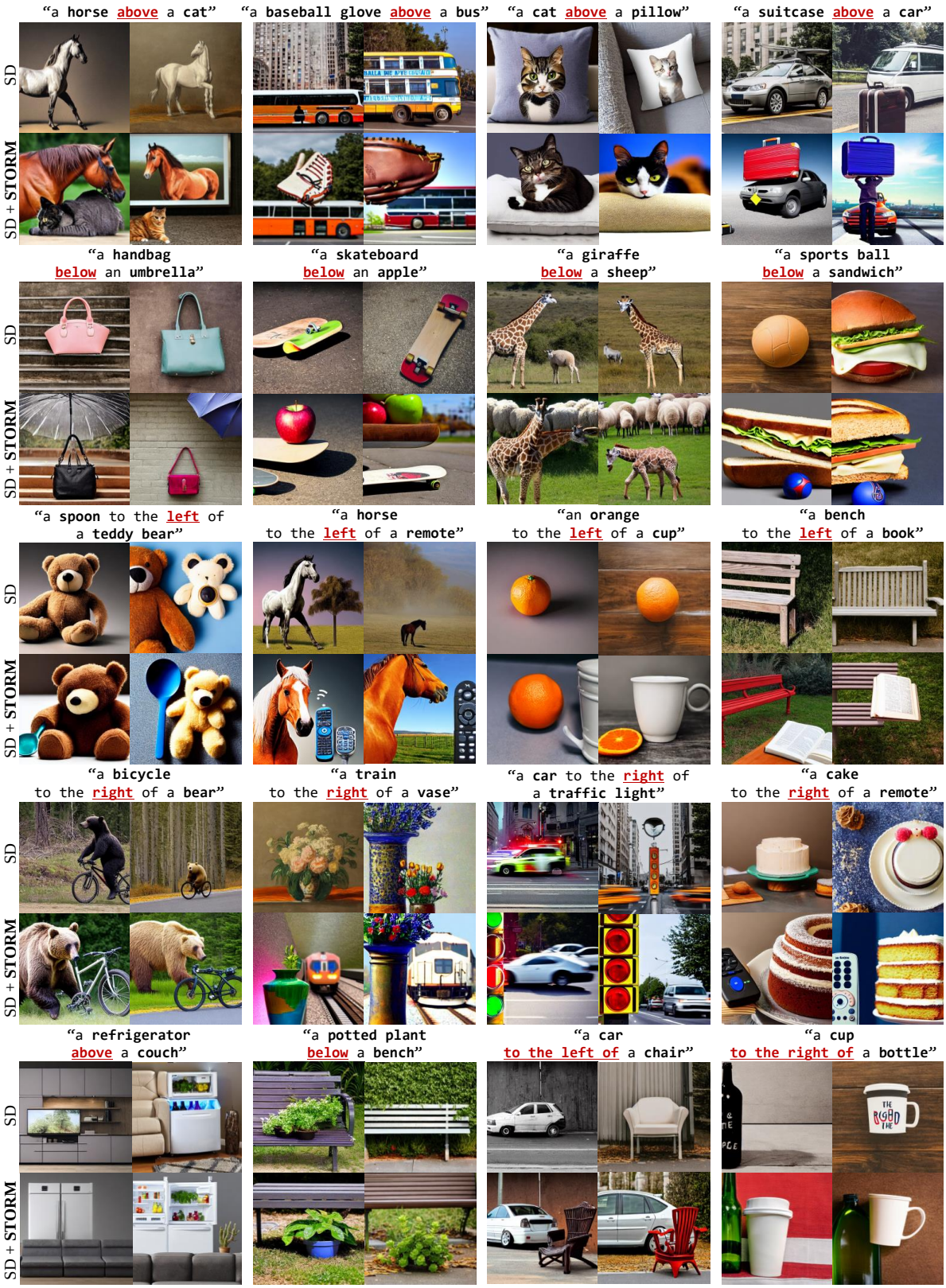

Figure 7. Comparison between Stable Diffusion 1.5 and ours on SR2D Dataset [3].

Figure 8. Additional qualitative results using *Stable Diffusion 1.4* in comparison with state-of-the-art methods.

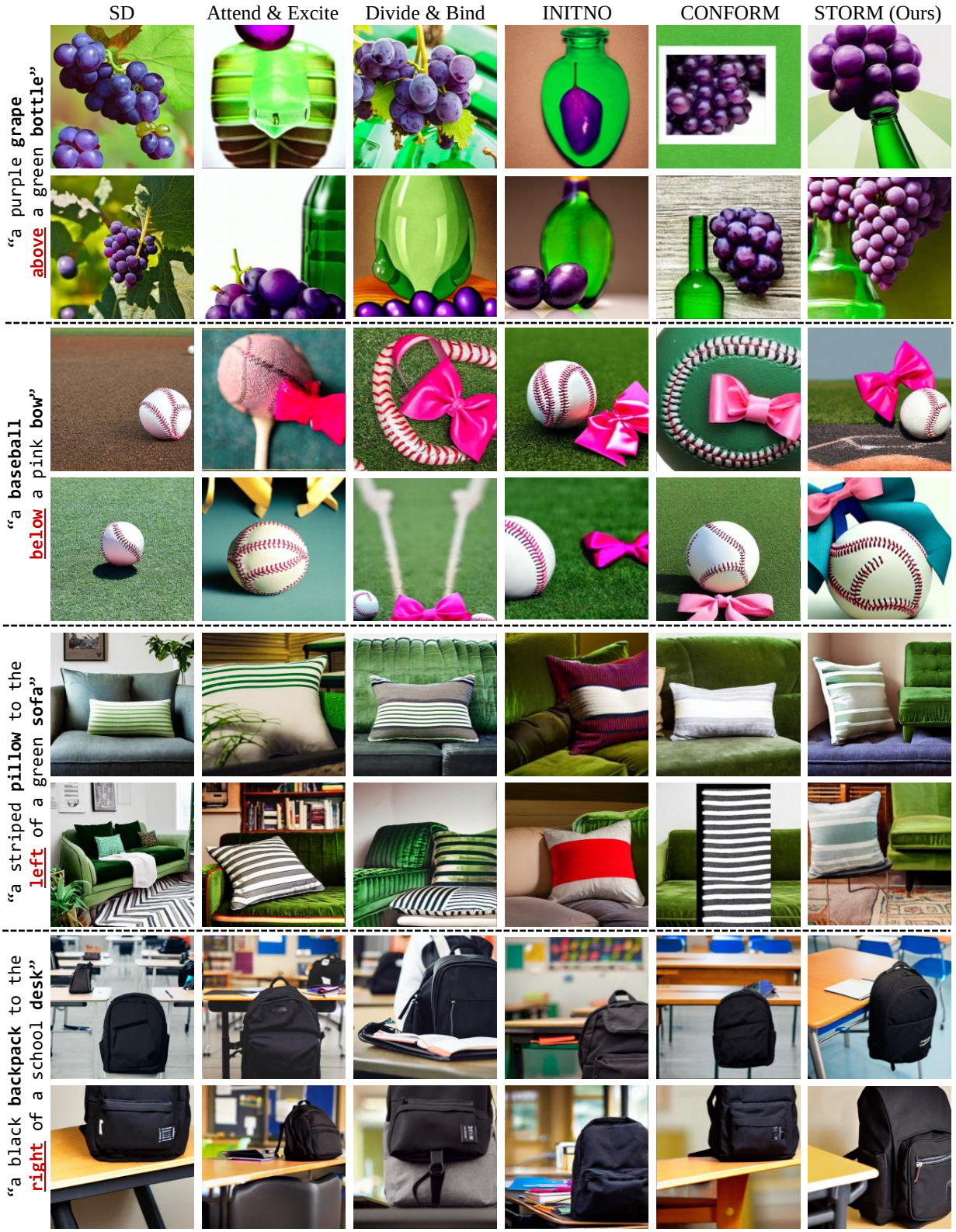

Figure 9. Additional qualitative results using *Stable Diffusion 1.4* in comparison with state-of-the-art methods.

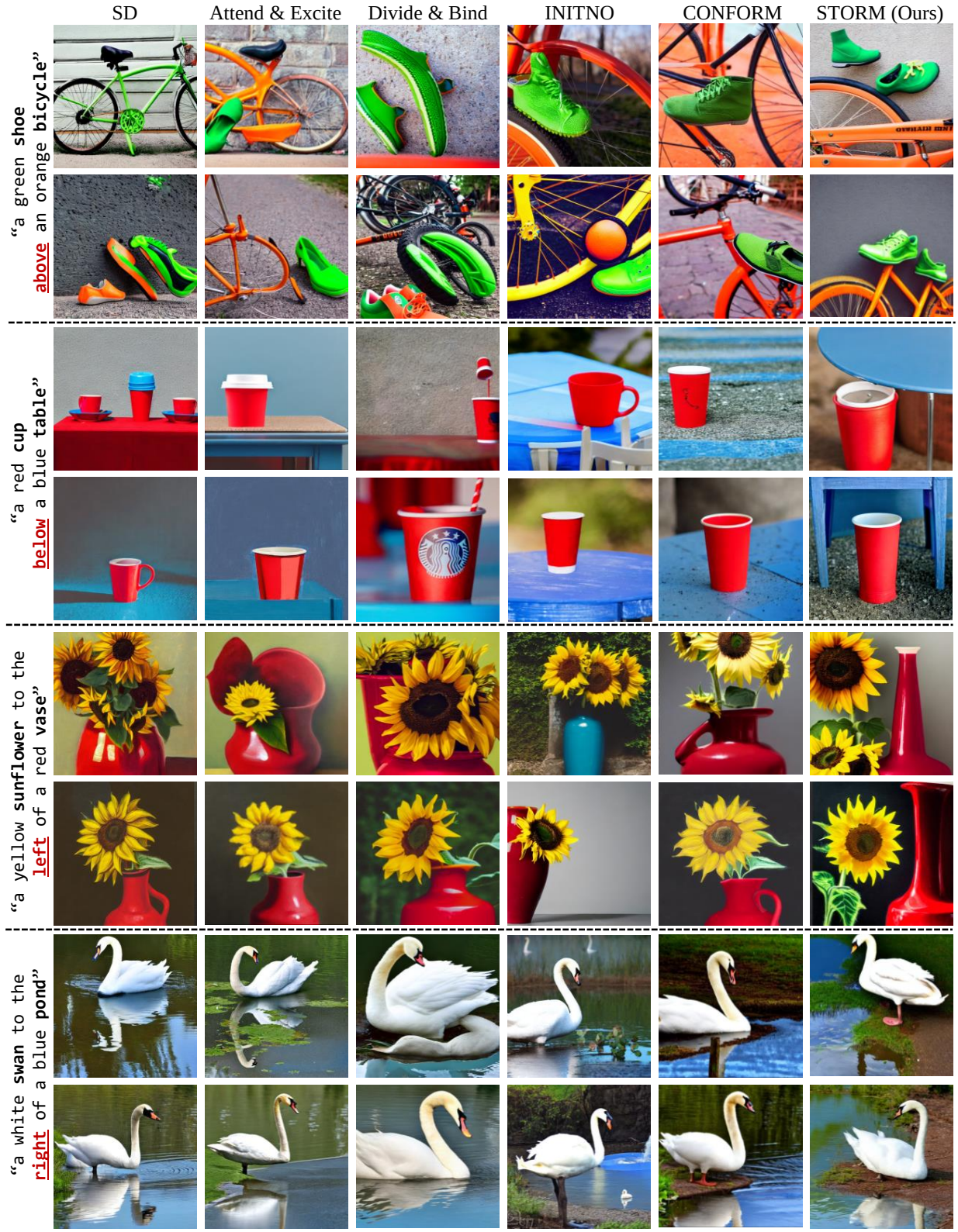

Figure 10. Additional qualitative results using *Stable Diffusion 2.1* in comparison with state-of-the-art methods.

Figure 11. Additional qualitative results using *Stable Diffusion 2.1* in comparison with state-of-the-art methods.

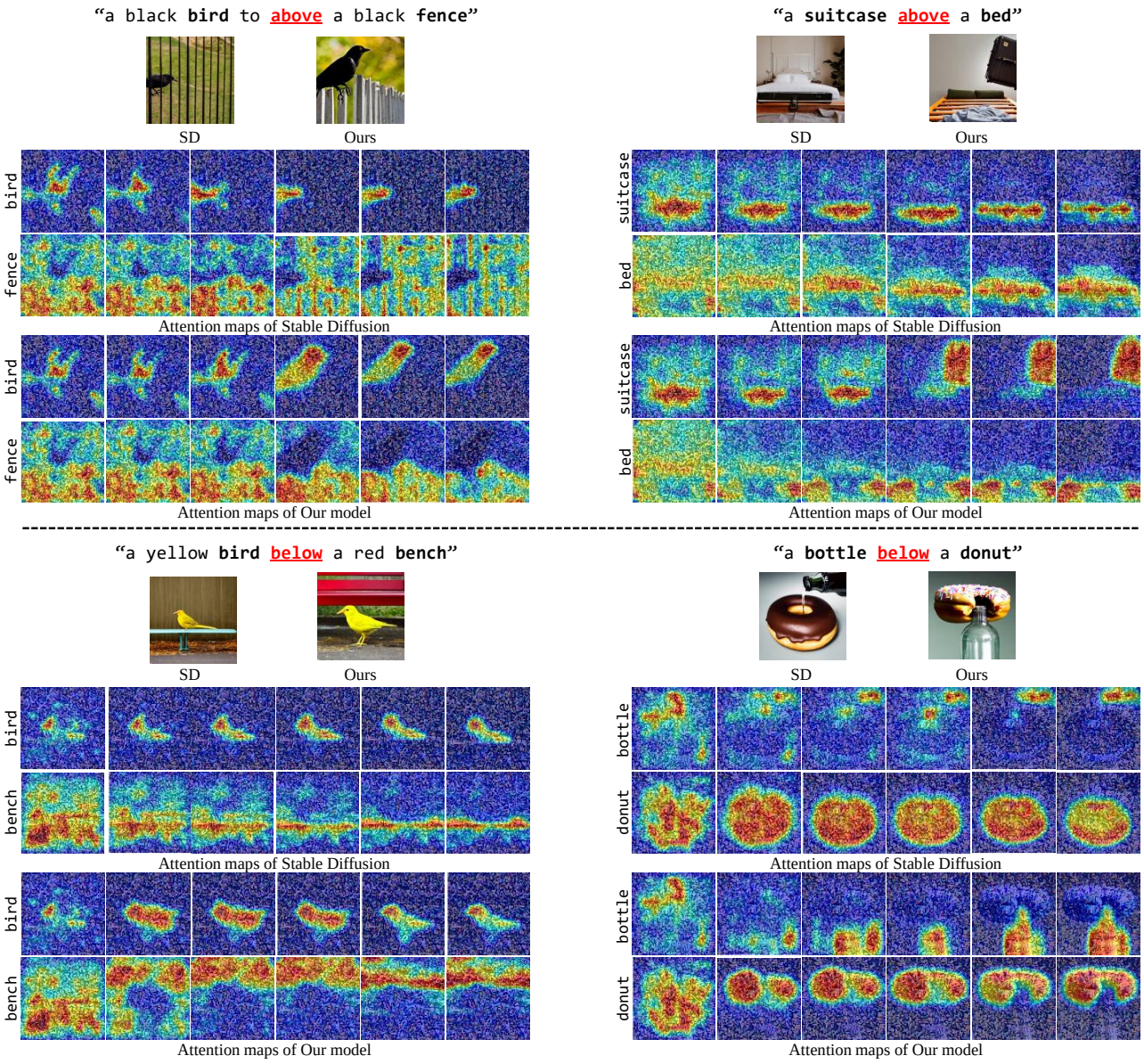

Figure 12. Additional visualizations of the attention map across different denoising timesteps. The leftmost figure represents the visualization of the attention map at the very early denoising steps, and as we move to the right, the figures show the attention maps after progressively more denoising steps.

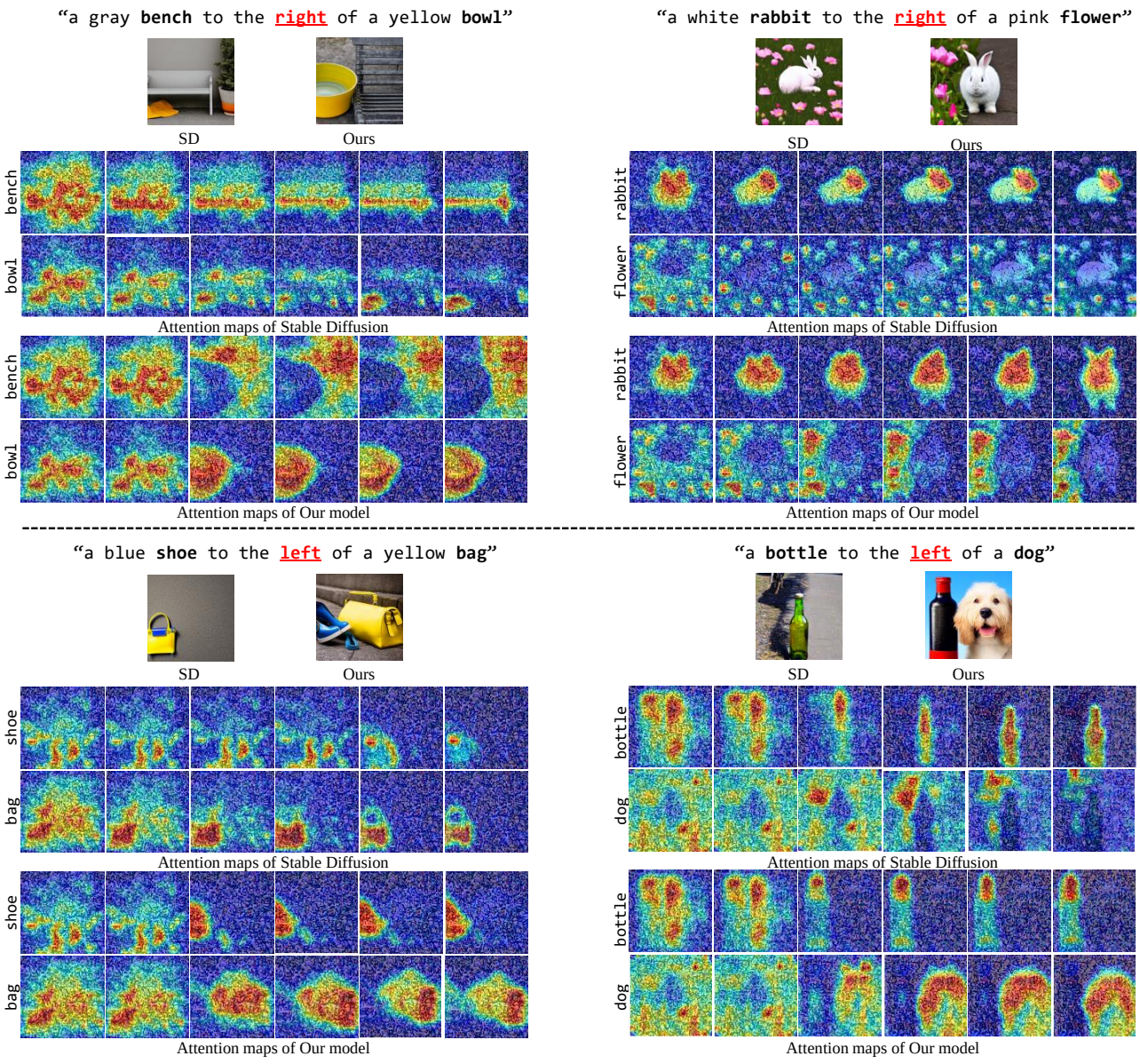

Figure 13. Additional visualizations of the attention map across different denoising timesteps. The leftmost figure represents the visualization of the attention map at the very early denoising steps, and as we move to the right, the figures show the attention maps after progressively more denoising steps.

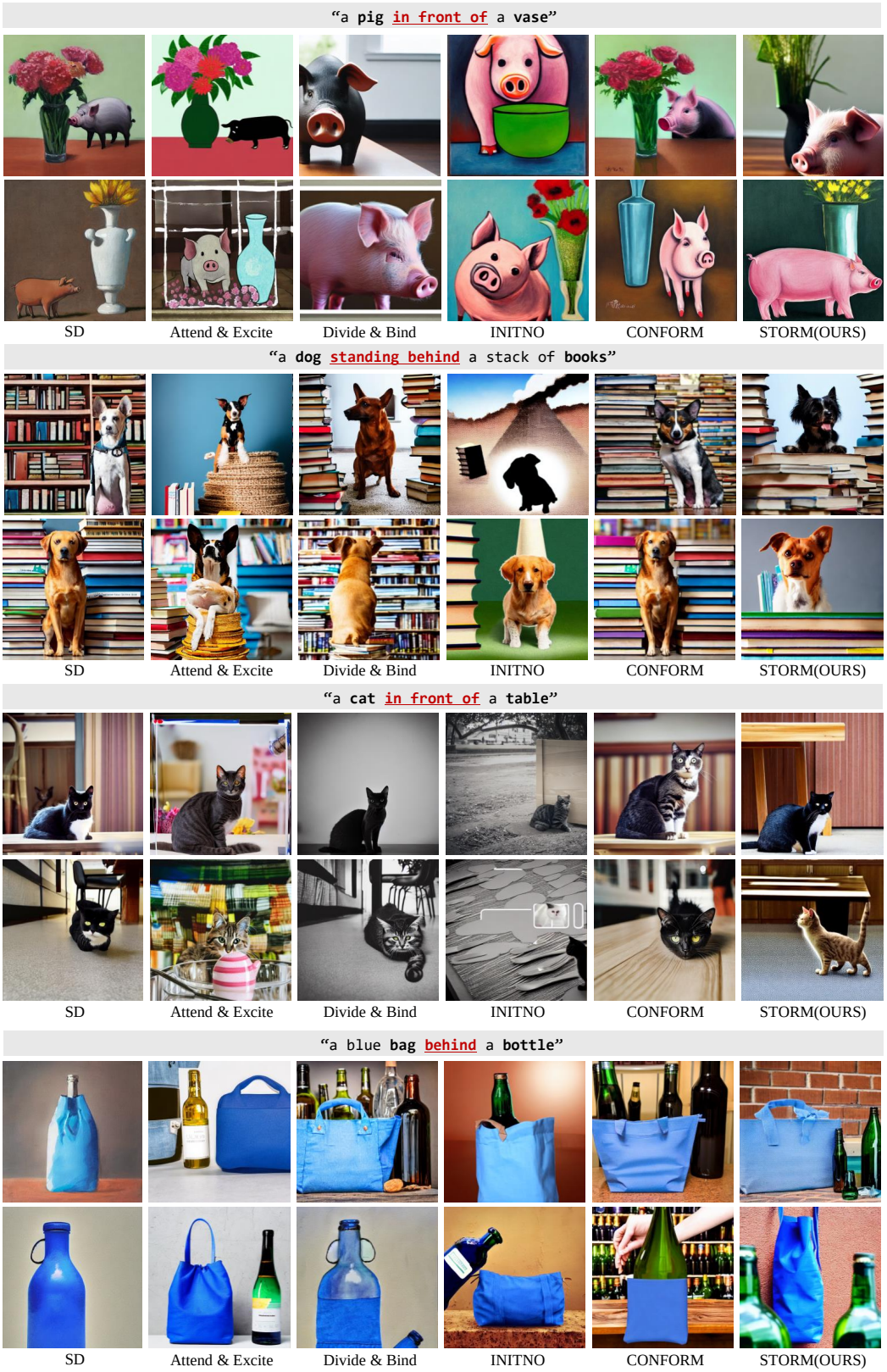

Figure 14. Additional results on 3D positional prompts.

References

[1] Tom B Brown. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 5

[2] 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. *SIGGRAPH*, 2023. 2, 6, 7

[3] Tejas Gokhale, Hamid Palangi, Besmira Nushi, Vibhav Vineet, Eric Horvitz, Ece Kamar, Chitta Baral, and Yezhou Yang. Benchmarking spatial relationships in text-to-image generation. *arXiv preprint arXiv:2212.10015*, 2022. 3, 6, 7, 12

[4] Xiefan Guo, Jinlin Liu, Miaomiao Cui, Jiankai Li, Hongyu Yang, and Di Huang. Initno: Boosting text-to-image diffusion models via initial noise optimization. In *CVPR*, 2024. 2, 6, 7

[5] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024. 7

[6] Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *ICCV*, 2023. 3, 5

[7] Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. *NeurIPS*, 2023. 3, 5

[8] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 5

[9] Jiachen Li, Jitesh Jain, and Humphrey Shi. Matting anything. In *CVPR*, 2024. 2, 6, 7

[10] Tuna Han Salih Meral, Enis Simsar, Federico Tombari, and Pinar Yanardag. Conform: Contrast is all you need for high-fidelity text-to-image diffusion models. In *CVPR*, 2024. 2, 6, 7

[11] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple open-vocabulary object detection. In *ECCV*, 2022. 4, 5

[12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 7

[13] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 5

[14] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Simple multi-dataset detection. In *CVPR*, 2022. 5