Spatial Transport Optimization by Repositioning Attention Map for Training-Free Text-to-Image Synthesis

Woojung Han Yeonkyung Lee Chanyoung Kim Kwanghyun Park Seong Jae Hwang*
Yonsei University
{dnwjddl, yeonkyung.lee, chanyoung, kwanghyun.park, seongjae}@yonsei.ac.kr

Abstract

Recent training-free diffusion models generate high-quality images but still misplace objects due to inherent difficulties in spatial guidance from text prompts. We introduce STORM (Spatial Transport Optimization by Repositioning Attention Map), a training-free method that ensures spatial coherence. STORM applies optimal-transport theory to reshape object-level attention maps in denoising, guided by a spatial-transport cost encoding prompt-specified locations. This intervention corrects mislocations and also reduces missing objects and attribute mismatches. Extensive experiments demonstrate that STORM consistently outperforms previous methods, establishing a new state-of-the-art for spatially faithful text-to-image synthesis without additional training.

1. Introduction

Diffusion-based text-to-image (T2I) models can synthesize high-quality images from textual prompts, and training-free variants further provide real-time adaptability, lower computational cost, and broad task generalization [5, 8, 12, 14, 15]. Yet these methods still face three persistent failure modes, *missing objects*, *mismatched attributes*, and, most critically, *mislocated objects*. While recent studies have reduced the first two problems [1, 11, 13], reliable spatial alignment remains largely unaddressed, as current models often overlook positional cues and place objects in unintended locations, ultimately diminishing output fidelity and limiting practical reliability across downstream applications.

Existing solutions for *mislocated objects* often rely on fixed spatial templates or predefined layouts [2, 3, 17], which offer some control but impose rigid constraints and require additional inputs. For instance, these templates may fix an object's position in the image (*e.g.,* far-left corner) instead of offering flexible guidance like placing one object "to the left" of another. Rather than depend on such restrictive templates, we aim to push the boundaries of T2I models by enabling precise alignment with textual guidance, essential for unlocking their full potential across applications.

*Corresponding author

Figure 1. Comparison of Spatial Awareness. { position* } in each prompt denotes the spatial relationship in each column. While Stable Diffusion (SD) shows limited spatial awareness by generating similar images regardless of spatial prompts, our model accurately reflects specified positions. (Same seed for all synthesis).

To this end, we propose **S**patial **T**ransport **O**ptimization by **R**epositioning Attention **M**ap, **STORM**, a training-free approach that adjusts relative object positions dynamically. While SD disregards spatial cues (see Fig. 1), STORM achieves various object positioning with the same model weights, ensuring spatial alignment. To implement STORM, we introduce the Spatial Transport Optimization (STO) framework, which interprets each text token's attention map as a distribution, and leverages Optimal Transport (OT) [16] to efficiently reposition distributions. While OT considers both distribution and distance to minimize transformation costs, we propose the Spatial Transport Cost (ST Cost) to align with our objectives and guide distribution positioning. This function updates and optimizes the latent vector through backward guidance, following methods from prior work [1, 2]. Furthermore, we found that missing objects can be resolved concurrently as mislocated objects. Extensive experiments show that our approach outperforms existing methods, setting a new spatial alignment benchmark.

2. Method

Since objects form in high-attention regions, we aim to reposition the attention map according to prompt positions, using relative objects as references. To achieve this, our approach involves three attention maps: (1) source distribution: the

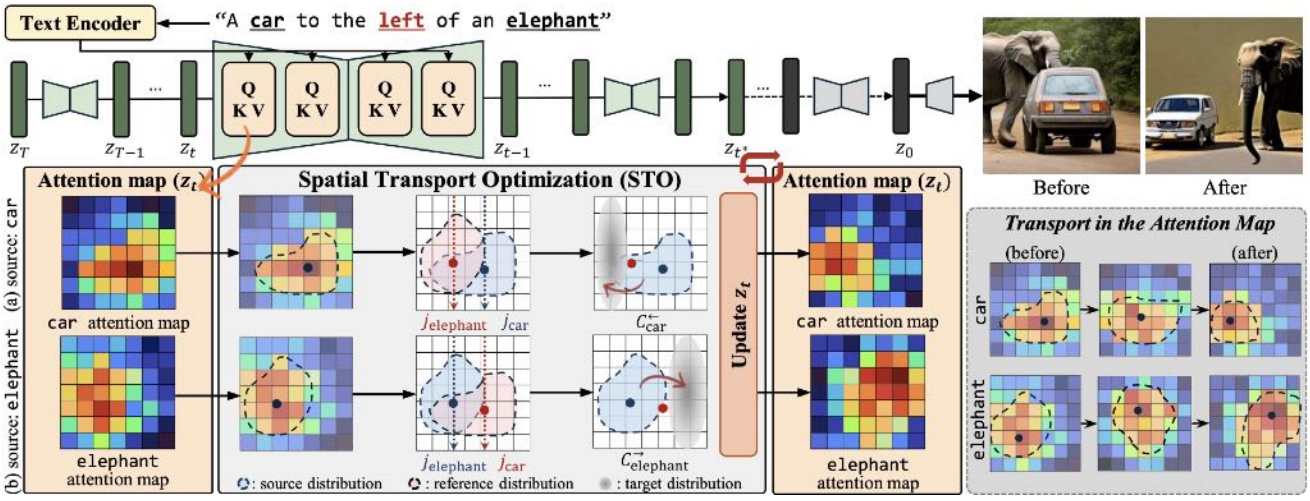

Figure 2. Overview pipeline of **STORM**. Our method leverages Optimal Transport in a training-free manner, allowing the model to accurately reflect relative object positions at each step without additional inputs. Given the prompt "*A car to the left of an elephant*", our method dynamically adjusts the attention maps to induce the specified spatial relationship. The process starts with initial attention maps for the "*car*" and "*elephant*" at time step z_t. Using the centroids of these attention maps, the Spatial Transport Optimization (STO) computes the losses to correct positional relationships (*e.g.,* ensuring the car is to the left of the elephant). The updated attention map is then used to refine the latent representation z_t, leading to a final image that adheres to the desired spatial arrangement. The comparison of attention maps (before and after STO) shows improved alignment, effectively placing the car to the left of the elephant as instructed in the prompt.

attention map of the object to be moved; (2) reference distribution: the attention map of the relative object; and (3) target distribution: arbitrary distribution which is influenced by the reference centroid $(j_{\text{ref}}, i_{\text{ref}})$; it guides the source distribution toward itself exerting minimal direct influence.

2.1. Spatial Transport Optimization Framework

Our framework has two components: a cost function and a transport plan. The cost function steers the source distribution toward the desired direction, enabling the attention map to shift with spatial cues. Extending standard distance-based costs, our Spatial Transport Cost (ST Cost) adds objectives specific to spatial alignment. The transport plan, implemented with the Sinkhorn algorithm, moves the distribution optimally. Starting from an initial attention map, STO iteratively updates the latent vector via ST Cost–driven losses, gradually aligning the distribution to the target.

2.1.1. Construction of Target Distributions

At each timestep, we introduce an auxiliary target distribution that steers the source relative to the reference. Depending on the text-specified direction, only the relevant coordinate of the reference centroid $(j_{\text{ref}}, i_{\text{ref}})$ is used: j_{ref} for "left" or "right" and i_{ref} for "above" or "below," while the other axis remains unconstrained. In Fig. 2a, for example, the `car` (source) is guided to the left of the `elephant` (reference); the same rule applies to any source object.

2.1.2. Spatial Transport Cost Function

In the STO framework, we aim to find an optimal transport plan that minimizes the cost of transferring the attention map. The cost function measures the expense of moving distribution mass by computing distances between points, where a higher value indicates less desirable locations. While our overall cost function merges the standard OT loss with an additional term, here we focus on our newly introduced ST Cost, which is built on two core principles:

(I) Positional Cost: We developed a cost function to guide the distribution to its intended position relative to the reference point. In simple terms, when the source distribution is intended to be on the left side of the reference point, it has a low cost when positioned on the left and a high cost when positioned on the right. To quantify this directional positioning, we define a set of positional δ values, representing the distances between the current point and the reference point in each of the four directions. Specifically, the δ values are defined as follows: $\delta^{\leftarrow}_{ij} = j_{\text{ref}} - j, \delta^{\rightarrow}_{ij} = j - j_{\text{ref}}, \delta^{\uparrow}_{ij} = i_{\text{ref}} - i, \delta^{\downarrow}_{ij} = i - i_{\text{ref}}$, with j and i representing the current position values, and the arrows indicate the desired and restricted direction for positioning the source attention map relative to the reference point. When constructing the overall cost function that incorporates all directional δ values, each direction is categorized as either desired or restricted. The desired direction (δ^{des}) corresponds to the intended movement of the source distribution, while the restricted direction (δ^{res}) represents the opposite or undesired direction. For instance, if the source distribution should move to the left, $\delta^{\leftarrow}_{ij}$ is marked as the desired direction (δ^{des}), while $\delta^{\rightarrow}_{ij}$ is treated as the restricted direction (δ^{res}). The combined cost function, accounting for both desired and restricted positions, can be expressed as:

$$\Delta_{ij}(\delta^{\text{des}}_{ij}, \delta^{\text{res}}_{ij}) = \frac{1}{\omega(\delta^{\text{des}}_{ij} + \epsilon)} \mathbb{1}_{+}[\delta^{\text{des}}_{ij}] + \omega(\delta^{\text{res}}_{ij} + \epsilon)\mathbb{1}_{+}[\delta^{\text{res}}_{ij}], \quad (1)$$

where $\omega(\cdot) > 1$ is a progressive adaptive weight that controls the alignment importance across timesteps, and ϵ is

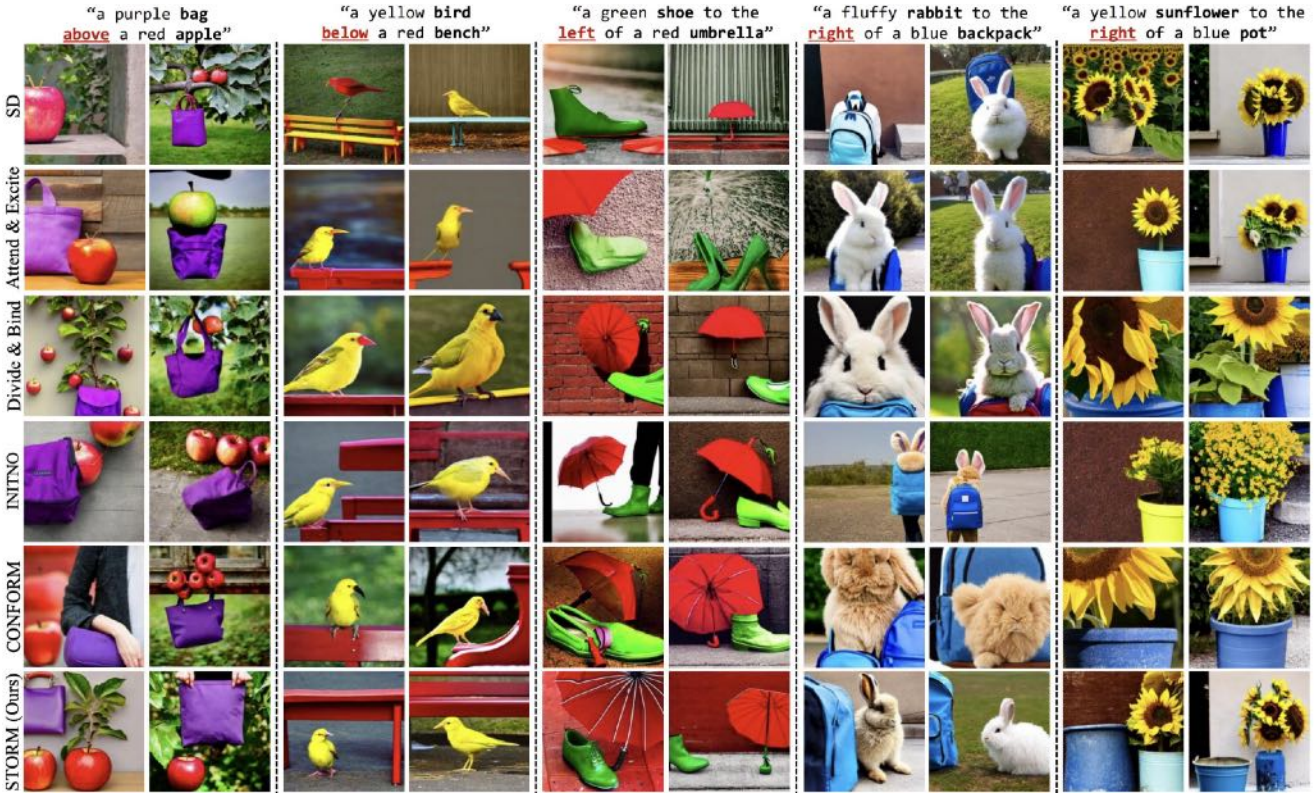

Figure 3. Qualitative comparison across the custom prompt, which involves attribute and positional information in text, evaluating previous state-of-the-art training-free T2I methods, Attend&Excite [1], Divide&Bind [11], INITNO [7], CONFORM [13], and ours.

a stabilizing factor. $\mathbb{1}_+$ is an indicator function that equals 1 when the input is positive and 0 otherwise, allowing selective application of computation based on positivity. As the source distribution aligns more closely with the desired direction, the Δ value approaches zero, indicating minimal cost for correctly aligned positioning.

(II) Non-overlap Cost: In this process, it is crucial to prevent the source distribution from overlapping with the reference distribution during its movement. This is one of the primary issues in addressing the missing object problem in SD, where avoiding overlap is essential for ensuring that each object is distinctly represented. To achieve this, we simply embed the reference distribution directly into the cost function, guiding objects to occupy separate locations. Specifically, with A_{ij} representing the attention weight of the reference distribution, we incorporated this distribution into the cost function, assigning high costs to regions occupied by the reference distribution and low costs elsewhere.

Combining this principle with our core ideas, our cost function can be defined as follows:

$$C_{ij} = A_{ij}\Delta_{ij}(\delta_{ij}^{\text{des}}, \delta_{ij}^{\text{res}}). \tag{2}$$

2.1.3. Sinkhorn Algorithm-based Transport Plan

After combining the standard OT approach with ST components, we compute the transport plan using the Sinkhorn algorithm [4], which applies entropic regularization for computational efficiency and stability, especially for high-dimensional attention maps. The Sinkhorn algorithm iteratively updates the transport plan **P** to meet the marginal constraints of source and target distributions, applying the cost function bidirectionally for both objects.

Table 1. Performance comparison between different models on VISOR (%) and Object Accuracy (OA) (%) metrics, based on Stable Diffusion 1.4 and Stable Diffusion 2.1 [15]. **Bold** marks best, underline marks second best.

Model	OA (%)	$\text{VISOR}_{\text{uncond}}$ (%)	$\text{VISOR}_{\text{cond}}$ (%)
Stable Diffusion 1.4 based			
SD 1.4	29.86	18.81	62.98
SD 1.4 + CDM	23.27	14.99	64.41
GLIDE	3.36	1.98	59.06
GLIDE + CDM	10.17	6.43	63.21
Structure Diffusion	28.65	17.87	62.36
Attend-and-Excite	42.07	25.75	61.21
Divide-and-Bind†	46.03	31.62	68.70
INITNO†	60.40	35.18	58.24
CONFORM†	60.73	38.48	62.33
Ours (SD 1.4)	**61.01**	**57.58**	**94.39**
Stable Diffusion 2.1 based			
SD 2.1	47.83	30.25	63.24
SPRIGHT	60.68	42.23	71.24
Ours (SD 2.1)	**62.55**	**59.35**	**94.88**

2.2. Optimization and Update Process

We update the latent vector using STO-based losses to guide the spatial alignment of the attention maps in a differentiable

Table 2. Comparison of methods on T2I-CompBench, calculating attribute binding and spatial relationships.

Method	Attribute Binding (Color ↑)	Attribute Binding (Shape ↑)	Attribute Binding (Texture ↑)	Object Relationship (Spatial ↑)
Stable Diffusion 1.4 Based				
Stable Diffusion-v1.4 [15]	0.3765	0.3576	0.4156	0.1246
Ours (SD 1.4)	**0.6458**	**0.5983**	**0.7539**	**0.1613**
Stable Diffusion 2.1 Based				
Stable Diffusion-v2.1 [15]	0.5065	0.4221	0.4922	0.1342
Composable Diffusion [12]	0.4063	0.3299	0.3645	0.0800
Structured Diffusion [5]	0.4990	0.4218	0.4900	0.1386
Attend-and-Excite [1]	0.6400	0.4517	0.5963	0.1455
Ours (SD 2.1)	**0.6777**	**0.6226**	**0.7884**	**0.1981**

Table 3. User study evaluating model performance on object synthesis, attribute matching, spatial correctness, and overall fidelity.

Method	Object Accuracy (%)	Attribute Matching (%)	Spatial Correctness (%)	Overall Fidelity (%)
Stable Diffusion [15]	14.68	14.31	11.81	15.02
Attend-and-Excite [1]	16.34	16.50	13.20	16.04
Divide-and-Bind [11]	16.39	15.69	13.01	16.52
INITNO [7]	15.98	16.03	13.15	15.63
CONFORM [13]	15.64	16.79	12.91	14.14
Ours	**20.97**	**20.68**	**35.92**	**22.65**

manner. At each timestep, we compute the ST loss $\mathcal{L}$ and update the latent vector z_t as $z_t' \leftarrow z_t - \alpha_t \cdot \nabla z_t \mathcal{L}$ where α_t is the step size. After the update, a forward pass computes z_{t-1} for the next denoising step.

3. Experiments

Implementation Details. We follow established protocols [1, 11, 13]: Stable Diffusion [15] v1.4 and v2.1, 50 denoising steps, guidance scale 7.5. To smooth the cross-attention map, a Gaussian filter with a kernel size of 3 and a standard deviation of 0.5 was applied. Optimization updates are applied at timesteps 5, 10, 15, 20, and halted at 25.

3.1. Evaluation Results

We evaluated our method against existing training-free T2I models using both qualitative and quantitative measures, focusing on two key aspects: Spatial Accuracy and Object and Attribute Consistency.

Quantitative Evaluation: VISOR. To evaluate spatial accuracy, we use the VISOR [6] to measure the ability of a model to position objects based on spatial cues (*e.g.,* above) in text prompts. VISOR benchmark contains over 25K prompts describing spatial relationships between objects.

Quantitative Evaluation: T2I-CompBench. The T2I-CompBench [9] evaluates object presence, spatial relations, and attribute alignment. We use its attribute and spatial subsets, averaging scores over 10+ random seeds. Attribute binding is measured with BLIP-VQA [10], and spatial relations with the UniDet metric [18]. As Table 2 shows, our images place objects correctly and preserve their attributes.

Quantitative Evaluation: User Studies. We create 10 custom text prompts with detailed objects, attributes, and spatial information, then generate images using random seeds. With 30 participants, our user study evaluates (1) object accuracy, (2) attribute matching, (3) spatial correctness, and (4) overall fidelity. In Table 3, our model shows surpassing results across the existing models of all categories.

Table 4. Ablation study on the impact of applying STO at different timesteps. Exp.#A0 is the SD baseline, and Exp. #A1 to #A4 apply STO over progressively broader timestep ranges: 19–24, 13–24, 7–24, and 1–24 (Ours).

# Exp.	OA (%)	VISOR					
		uncond	cond	1	2	3	4
A0	29.86	18.81	62.98	46.60	20.11	6.89	1.63
A1	46.62	37.48	80.38	74.11	45.73	22.30	7.82
A2	55.65	47.90	86.07	81.63	60.30	35.89	13.81
A3	59.85	53.62	89.60	83.61	67.53	44.12	19.31
A4 (Ours)	**61.01**	**57.58**	**94.39**	**85.93**	**69.71**	**49.01**	**25.70**

Figure 4. Comparison of results when applying STO at different timesteps. From left: baseline (no STO), STO at 19-24, 13-24, 7-24, and 1-24 (ours). Earlier application yields clearer separation and more accurate placement; see Table 4 for quantitative results.

Qualitative Analysis. Fig.3 compares our model with recent state-of-the-art methods[7, 13]. Using custom prompts that test both attribute fidelity and spatial accuracy, other models often misplace objects; for example, a `rabbit` overlaps a `backpack` instead of appearing to its right. Our positioning-focused approach, in contrast, aligns objects correctly and preserves their attributes.

3.2. Ablation Studies

Effects of Optimizing through Timestep. Table 4 compares STO applied at four timestep ranges (Exp.#A1 19–24, Exp.#A2 13–24, Exp.#A3 7–24, Exp.#A4 1–24). These experiments show that setting spatial configurations in the early stages is essential not only for positional accuracy but also for achieving higher OA. This improvement increases as spatial relationships are established earlier in the process. All configurations with STO outperform the baseline (Exp.#A0, without STO), demonstrating a substantial improvement in both object positioning and accuracy. In Fig. 4, tie placement is optimal when spatial cues are applied early, while delayed guidance leads to misplacements or suboptimal synthesis.

4. Conclusion

We introduced STORM, a framework that dynamically addresses spatial misalignment issues in training-free T2I synthesis. By leveraging the STO framework, which combines the ST Cost function in Optimal Transport theory, STORM not only resolves mislocated objects but also tackles missing objects and mismatched attributes. Extensive experiments show that STORM significantly improves spatial alignment and object accuracy, surpassing existing state-of-the-art methods.

References

[1] 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. 1, 3, 4

[2] Minghao Chen, Iro Laina, and Andrea Vedaldi. Training-free layout control with cross-attention guidance. In *WACV*, 2024. 1

[3] Guillaume Couairon, Marlene Careil, Matthieu Cord, Stéphane Lathuiliere, and Jakob Verbeek. Zero-shot spatial layout conditioning for text-to-image diffusion models. In *ICCV*, 2023. 1

[4] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *NeurIPS*, 2013. 3

[5] Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. *arXiv preprint arXiv:2212.05032*, 2022. 1, 4

[6] 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. 4

[7] 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. 3, 4

[8] Woojung Han, Chanyoung Kim, Dayun Ju, Yumin Shim, and Seong Jae Hwang. Advancing text-driven chest x-ray generation with policy-based reinforcement learning. *MICCAI*, 2024. 1

[9] 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. 4

[10] 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. 4

[11] Jiachen Li, Jitesh Jain, and Humphrey Shi. Matting anything. In *CVPR*, 2024. 1, 3, 4

[12] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *ECCV*, 2022. 1, 4

[13] 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. 1, 3, 4

[14] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 1

[15] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 3, 4

[16] Cédric Villani et al. Optimal transport: old and new. *Springer*, 338, 2009. 1

[17] Qiucheng Wu, Yujian Liu, Handong Zhao, Trung Bui, Zhe Lin, Yang Zhang, and Shiyu Chang. Harnessing the spatial-temporal attention of diffusion models for high-fidelity text-to-image synthesis. In *ICCV*, 2023. 1

[18] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Simple multi-dataset detection. In *CVPR*, 2022. 4