ELDiff: When Evidential Learning Meets Text-to-Image Diffusion

Qingtao Pan[1,2], Kai Ye[2], Zhihao Dou[2], Bing Ji[1(✉)], and Shuo Li[2]

[1] Shandong University, Jinan, Shandong, China
b.ji@sdu.edu.cn
[2] Case Western Reserve University, Cleveland, OH, USA

Abstract. In multi-object text-to-image (T2I) diffusion, ensuring semantic consistency between textual prompts and generated visual content is crucial for image synthesis. However, such consistency constraint is often underemphasized in the denoising process of diffusion models. Although token supervised diffusion models can mitigate this issue by learning object-wise consistency between the image content and object segmentation maps, it tends to suffer from the problems of segmentation map bias and semantic overlap conflict, especially when involving multiple objects. In this paper, we propose ELDiff, a new evidential learning-supervised T2I diffusion model, which leverages the advantages of uncertainty metric and conflict detection to enhance the fault tolerance of unreliable segmentation maps and suppress semantic conflicts, strengthening object-wise consistency learning. Specifically, a pixel evidence loss is proposed to restrain overconfidence in unreliable labels through evidential regularization, and a token conflict loss is designed to weaken the contradiction between semantics through optimizing a measured conflict factor. Extensive experiments show that our ELDiff outperforms existing training based and train-free based T2I diffusion models on SD v1.4, SD v2.1, SDXL, SD v3.5, and Qwen-Image, without requiring additional inference-time manipulations. Notably, ELDiff can be seamlessly extended to the existing training pipeline of T2I diffusion models. Code can be found at `https://github.com/QingtaoPan/ELDiff`.

Keywords: Text-to-image diffusion · Evidential learning

1 Introduction

Recent advances in text-to-image (T2I) diffusion models [6,13,51,65,66] have demonstrated remarkable progress within computer vision applications, propelling the capabilities of general generative models to a new level. Leveraging their exceptional generation ability, these models enable users to synthesize highly realistic and creative visual content through user-specified text prompts. However, they often struggle to compose multiple objects into a coherent scene since the noise is predicted by a full text prompt in the denoising objective of T2I diffusion models [12, 61], which requires the model to be equipped with the ability of understanding both the full text prompt and individual linguistic concepts from the prompt.

Several studies [5, 12, 33, 37, 41, 61] have addressed the challenge of generating images including multiple objects. They enable the generated images better reflect the text prompts through utilizing fine-grained text information. As proposed in TokenCompose [61], the fundamental idea is to perform consistency constraint between the text prompts and the image contents in the finetuning stage by imposing training objectives at the token level. It adopts Grounding DINO [38] and Segment Anything (SAM) [30] to obtain the grounding segmentation map of each text token for

local supervision of each text token. Once fine-tuned, the model can generate images by composing different combinations of words from text prompts in the inference stage. Although these methods have shown great success, they still suffer from segmentation map bias and semantic overlap conflict, which is problematic particularly when the generated image involves multiple objects.

We hypothesize that the limitations come from two aspects. On the one hand, the masks (the segmentation maps) generated directly by SAM are not always reliable [22], which results in overfitting to spurious regions. On the other hand, semantic overlaps result in conflicts [28] among different text tokens since the consistency constraint between each noun token from the text prompt and its corresponding segmentation map is individually optimized. Therefore, we seek to address the following problems: *How to cope with segmentation map bias and semantic overlap conflict for more effective consistency constraint in T2I diffusion models?*

Motivation: Inspired by the above observation, we find it desirable to tackle such problems by considering the uncertainty metric of pixel-level predictions and quantifying the semantic overlap conflict. Evidential Deep Learning (EDL) [42, 55], which is able to collect evidence of each category and estimate the epistemic uncertainty in a single forward pass. Given this, we consider each pixel of each noun token from the text prompt of as a classification unit and provide evidence values per class to reflect both the confidence and the associated uncertainty. In this way, the model tends to produce low-confidence predictions in mislabeled regions of the biased segmentation map, thereby avoiding overfitting to incorrect supervision. Moreover, Dempster-Shafer Evidence Theory (DST) [9], an evidence combination rule, allows beliefs from different evidences to be combined to obtain a new belief [24, 56]. It is able to quantify the conflict level through introducing the conflict penalty item in overlapping areas. Inspired by this, we consider each noun token from the text prompt as different evidences and utilize DST combination rule to measure their conflict. Therefore, the model can perceive overlapping conflicts between different semantic evidences, effectively alleviating the semantic competition problem caused by independent optimization.

To this end, we propose ELDiff, an end-to-end fine-tuning strategy to enhance multi-object composition through a novel consistency constraint (i.e., EDL driven object-wise consistency constraint) between the text prompts and the image contents. Specifically, a pixel evidence loss is proposed to cope with the segmentation map bias. It models the per-token cross-attention maps as evidence distributions (i.e., Dirichlet distribution) and calculates the cross entropy using the mean value of the predicted Dirichlet distribution instead of directly using point predictions. In this way, when the uncertainty is high (i.e., the distribution is loose), the loss of EDL will not excessively penalize the model, allowing the model to remain cautious when facing biased segmentation map and avoid being forced to learn incorrectly. In addition, a token conflict loss is proposed to alleviate semantic overlap conflict. It regards per-token cross-attention map as multiple evidence embeddings and utilizes DST combination rule to aggregate such multiple evidence embeddings for conflict measurement. The measured conflict is utilized as an optimization target, enabling the model to adaptively adjust the spatial distribution of different tokens' cross-attention maps, thus mitigating semantic conflicts arising from attention map overlap. Our main contributions are summarized as the following:

- We formulate the problems of segmentation bias and semantic overlap conflict during the object-wise consistency constraint between the text prompts and the image contents. And we propose a new T2I finetuning pipeline, ELDiff, which introduces EDL based pixel-level uncertainty metric and DST based token-level conflict quantification to address these two problems. ELDiff can be seamlessly integrated into the existing T2I diffusion models.
- We design a pixel evidence loss to restrain overconfidence in unreliable labels through simultaneously considering pixel-level classification prediction and uncertainty estimation.

- We propose a token conflict loss to weaken the semantic overlap conflict through treating the conflict factor measured by the DST combination rule as the optimization objective.
- Extensive comparisons with previous outstanding methods demonstrate that ELDiff effectively improves the performance in composing multiple objects without additional inference cost.

2 Related Work

Text-to-Image Synthesis. The field of text-to-image synthesis [10, 18, 47, 59, 64, 68] has demonstrated remarkable generative power in various fields, such as text-to-image generation [53, 54], text-to-video generation [4, 58], and text-to-3D generation [34, 48]. By encoding text prompts into a condition vector using a pretrained CLIP [50], T2I diffusion models have achieved successful applications in image generation. Although progress has been made, compositional generation still struggles when text prompts involve multiple objects and intricate relationships [63]. One potential method to generate multi-target images is manipulating the latent and cross-attention maps [5, 7, 12, 25, 37, 41, 52, 61]. More recently, TokenCompose [61] optimizes the cross-attention maps based on segmentation maps to provide dense consistency constraint. It achieves strong compositionality and image quality without additional inference time. Despite its superiority, it suffers from the segmentation bias and optimizing multiple objects individually leads to semantic overlap conflicts, resulting in degraded image generation quality. In this work, we present a different approach to address these problems for more reliable multi-category image composition through consistency constraint with evidential supervision.

Evidential Deep Learning. EDL is gradually developed and refined based on Dempster-Shafer theory of evidence [9] and Subjective Logic theory [3]. The core idea of EDL is to collect evidence of each category and construct a Dirichlet distribution parameterized over the collected evidence to model the distribution of class probabilities. In addition to the probability of each category, the predictive uncertainty can be quantified from the distribution by Subjective Logic theory in the forward pass. EDL has been applied in a variety of research areas, e.g., EDL based classification [14,69] and segmentation [43,70], evidential models for regression [1,45], adversarial robustness [31] and calibration [60]. Most existing EDL approaches typically incorporate evidential loss along with evidence regularization to guide the uncertainty behavior [44], [57] of the evidence. In this work, we focus on EDL loss for segmentation and DST for multi-evidence aggregation. It is a promising way to tackle the problem of segmentation map bias and semantic overlap conflict, enabling more effective image synthesis involving multiple objects.

3 Method

Our ELDiff (Fig. 1) proposes a pixel evidence loss to restrain overconfidence in unreliable labels when conducting pixel-level supervision between the image content and the object segmentation map, and designs a token conflict loss to weaken the contradiction between semantics through optimizing a measured conflict factor. In Section 3.1, we illustrate the problem setup of the segmentation map bias and semantic overlap conflict. Subsequently, in Section 3.2, we detail how the pixel evidence loss address this through pixel-level uncertainty metric. Finally, in Section 3.3, we demonstrate how the token conflict loss alleviates the semantic conflict through DST combination rule.

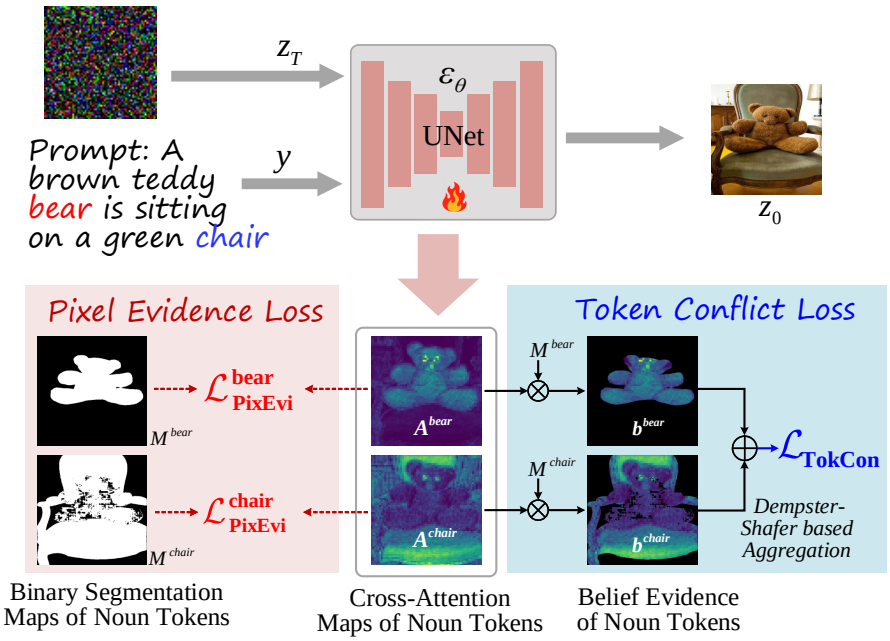

Fig. 1: Overview of ELDiff. The main components include: (1) The pixel evidence loss drives reliable consistency constraint between the text prompts and the image contents through pixel-level uncertainty estimation; (2) The token conflict loss weakens the contradiction between semantics.

3.1 Problem setup

Segmentation map bias. While consistency constraint based T2I diffusion methods have shown great success [61, 71], the generation effect heavily depends on the quality of the segmentation map used for consistency constraint. For example, when performing a training constraint between the text token guided cross-attention map and the corresponding binary segmentation map, the model overfits to incorrect regions if binary segmentation map is biased.

Semantic overlap conflict. Although $\mathcal{L}_{pixel}$ can focus the cross-attention map of nouns in text prompts on the target area, a side effect of this loss is that separately optimizing each cross-attention map is contradictory when different semantics overlap [29]. To be specific, this per-noun supervision strategy inherently assumes that the visual concepts of different nouns are spatially disjoint. For example, in the phrase "a cat on a chair," the noun "cat" and "chair" may correspond to overlapping regions in the image due to physical interaction or spatial proximity. Supervising each noun's cross-attention map with an independent segmentation map leads to semantic overlap conflicts, where the same pixel may be simultaneously assigned to multiple concepts. This introduces contradicting gradients during training, impairing the model's ability of concept-to-region mappings.

3.2 Pixel Evidence Loss

To address the segmentation bias, our key insight is to add uncertainty metric between the cross-attention map and the segmentation map to achieve more reliable consistency constraint. We accomplish this by leveraging the epistemic uncertainty ability of EDL, which can quantify the credibility

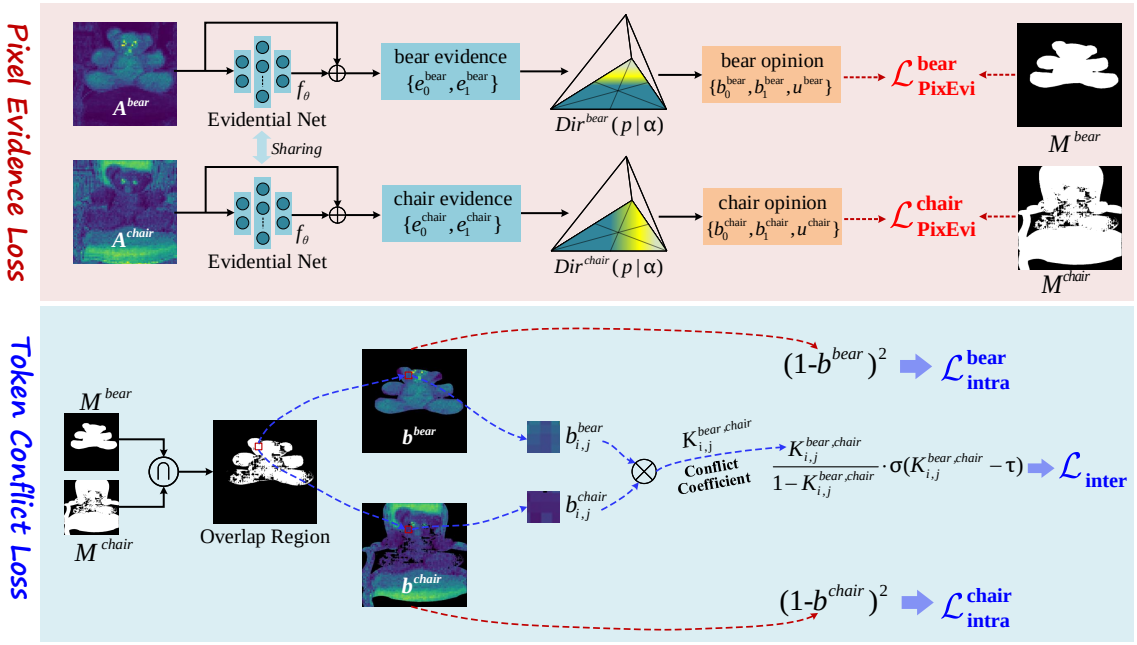

Fig. 2: Detailed diagram of pixel evidence loss and token conflict loss. For pixel evidence loss, the noun token's cross-attention maps are fed into the evidential Net to generate noun token's evidences. Then, these evidences are used to build the Dirichlet distribution, outputting pixel-level probabilities and the uncertainty for reliable pixel-level supervision. For token conflict loss, the conflict coefficient of the overlapping area between noun token's cross-attention maps is calculated for optimization. Moreover, the activations of different noun tokens are aggregated into respective corresponding spatial regions.

of its predictions. Instead of treating the attention score of each pixel as a deterministic prediction, we reinterpret it as evidence supporting its foreground or background class. As shown in Fig. 2, these evidence values are modeled via a Dirichlet distribution to construct an uncertainty-aware attention learning process.

Evidence-based Attention Modeling. Given a cross-attention map A and the corresponding binary segmentation map M, A is fed into an MLP evidence network $f_\phi(\cdot)$ with an activation function (Softplus) to map A into the evidence space, thus outputting evidence values e. For each pixel $A_{i,j}$, its evidence is defined as $e_{i,j} \in \mathbb{R}_{\geq 0}$, from which the parameters of a Dirichlet distribution $D(p \mid \alpha)$ are obtained as $\alpha_{i,j} = e_{i,j} + 1$. The Dirichlet distribution $D(p \mid \alpha)$ is considered as the conjugate prior to the multinomial distribution. It provides a predictive distribution for the segmentation results, defned as follows:

$$D(p \mid \alpha) = \begin{cases} \frac{1}{\beta(\alpha)} \prod_{k=1}^{K} p_k^{\alpha_k - 1} & for \quad p \in \Omega \\ 0 & otherwise \end{cases} \tag{1}$$

where $p = [p_1, \ldots p_k]$ are the parameters of the multinomial distribution, $\beta(\alpha)$ is the high-dimensional multinomial beta function, and Ω is the $K-1$ dimensional unit simplex, defned as:

$$\Omega = \left\{ p \middle| \sum_{c=1}^{C} p^c = 1 \text{ and } 0 \leq p^1, \ldots, p^C \leq 1 \right\} \tag{2}$$

Subsequently, Subjective Logic [25] is applied for the optimization of Dirichlet distribution parameter optimization. It establishes a theoretical foundation linking Dirichlet distribution parameters to confidence and uncertainty quantification. Specifcally, for the predicted cross-attention map A, it provides a mass belief and uncertainty, satisfying:

$$u_{i,j}^{c} + \sum_{c=1}^{C} b_{i,j}^{c} = 1 \tag{3}$$

where $b_{i,j}$ and $u_{i,j}^{c}$ are the cross-attention map probability of the pixel (i,j) for the c-th class and the uncertainty of the pixel (i,j), respectively. The the mass of belief and uncertainty for the pixel (i,j) can be expressed as follows:

$$b_{i,j}^{c} = \frac{e_{i,j}^{c}}{S} = \frac{\alpha_{i,j}^{c} - 1}{S} \quad \text{and} \quad u_{i,j} = \frac{C}{S} \tag{4}$$

where $S = \sum_{c=1}^{C} \left(e_{i,j}^{c} + 1\right)$ is the Dirichlet strength. It describes that the higher the allocated belief mass, the more evidence is obtained for pixel (i,j). Conversely, the less evidence obtained, the greater the overall uncertainty for the segmentation of pixel (i,j).

Considering that, on the simplex, the ideal Dirichlet distribution should concentrate its mass on the vertex corresponding to the true class label. The distribution parameters should be close to 1 for all incorrect classes and significantly larger for the correct one. Therefore, it is necessary to design a loss function that guides the model to optimize the Dirichlet parameters in a way that minimizes segmentation uncertainty. Specifcally, We use the cross-entropy loss $\mathcal{L}_{ce} = \sum_{c=1}^{C} -y^{c} \log\left(p^{c}\right)$ to link the Dirichlet distribution with the belief representation in subjective logic. Based on the evidence, this loss can be reformulated to reflect the expected prediction under the Dirichlet distribution, enabling the model to learn both accurate and uncertainty-aware segmentations, defned as follows:

$$\begin{aligned} \mathcal{L}_{ice} &= \int \left[\sum_{c=1}^{C} -y^{c} \log(p^{c}) \right] \frac{1}{\beta(\alpha)} \prod_{c=1}^{C} \left(p^{c}\right)^{\alpha^{c}-1} dp \\ &= \sum_{c=1}^{C} y^{c} \left(\psi(S) - \psi(\alpha^{c})\right) \end{aligned} \tag{5}$$

where y^{c} is the ground truth labels and $\psi\left(\cdot\right)$ is digamma function. We further incorporate a Kullback-Leibler (KL) divergence loss to suppress evidence for incorrect classes while preventing the Dirichlet parameter of the ground-truth class from being forced to 1, defined as:

$$\begin{aligned} \mathcal{L}_{KL} = \log &\left(\frac{\Gamma\left(\sum_{c=1}^{C} \tilde{\alpha}^{c}\right)}{\Gamma(C) \sum_{c=1}^{C} \Gamma\left(\tilde{\alpha}^{c}\right)} \right) \\ &+ \sum_{c=1}^{C} \left(\tilde{\alpha}^{c} - 1\right) \left[\psi\left(\tilde{\alpha}^{c}\right) - \psi\left(\sum_{c=1}^{C} \tilde{\alpha}^{c} \right) \right] \end{aligned} \tag{6}$$

where $\tilde{\boldsymbol{\alpha}}^{c} = y^{c} + \left(1 - y^{c}\right) \odot \boldsymbol{\alpha}^{c}$ and $\Gamma\left(\cdot\right)$ is the gamma function.

The loss $\mathcal{L}_{\text{PixEvi}}$ in this section consists of cross-entropy loss $\mathcal{L}_{ce}$, $\mathcal{L}_{ice}$, and $\mathcal{L}_{KL}$.

$$\mathcal{L}_{\text{PixEvi}} = \frac{1}{N}\sum_{n=1}^{N}(\lambda_1 \mathcal{L}_{ce}^{(n)} + \lambda_2(\mathcal{L}_{ice}^{(n)} + \mathcal{L}_{KL}^{(n)})) \tag{7}$$

where n represents each text token that belongs to a noun within the text prompt, and λ_1, λ_2, and λ_3 are hyperparameters to balance these three losses. The cross-entropy loss $\mathcal{L}_{ce}$ is adopted to maximize the consistency between the cross-attention map and the binary segmentation map.

3.3 Token Conflict Loss

To solve the semantic overlap conflict, we propose to aggregate all noun-level cross-attention maps using Dempster-Shafer Evidence Theory (DST) [9] to explicitly quantify inter-noun conflicts (Fig. 2), guiding the model to learn semantically decoupled attention regions.

Cross-Attention Map Aggregation. Given a set of per-noun cross-attention maps $\{A^1, A^2, \ldots, A^N\}$, we construct basic belief assignment function based on DST.

$$b^n = \gamma^n A^n \cdot M^n \tag{8}$$

where b^n is the belief evidence belonging to the n-th noun category. $\gamma^n \in (0, 1)$ represents learnable factor. M^n is the corresponding binary segmentation map. For tow nouns (v, w), the conflict coefficient within the overlapping area $M^v \cap M^w$ is calculated by:

$$K_{i,j}^{v,w} = \begin{cases} b_{i,j}^v \cdot b_{i,j}^w, & \text{if}(i,j) \in M^v \cap M^w \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

Subsequently, we define a intra-object consistency loss $\mathcal{L}_{\text{intra}}$ to aggregate its activations of the cross-attention map into certain subregions of its target regions and a inter-object conflict loss $\mathcal{L}_{\text{inter}}$ to punish conflicting activations in overlapping areas, defined as:

$$\mathcal{L}_{\text{intra}} = \sum_{n}(1 - b^n)^2 \tag{10}$$

$$\mathcal{L}_{\text{inter}} = \sum_{(i,j)\in M^v \cap M^w} \frac{K_{i,j}^{v,w}}{1 - K_{i,j}^{v,w}} \cdot \sigma(K_{i,j}^{v,w} - \tau) \tag{11}$$

where σ is Sigmoid function and τ is constraint threshold. The final token conflict loss is the sum of $\mathcal{L}_{\text{intra}}$ and $\mathcal{L}_{\text{inter}}$, defines as:

$$\mathcal{L}_{\text{TokCon}} = \frac{1}{N}\sum_{n=1}^{N}(\eta_1 \mathcal{L}_{\text{intra}}^{(n)} + \eta_2 \mathcal{L}_{\text{inter}}^{(n)}) \tag{12}$$

Finally, our ELDiff is jointly optimized by $\mathcal{L}_{LDM}$, $\mathcal{L}_{\text{ELDiff}}$, and $\mathcal{L}_{\text{TokCon}}$. The training objective is $\mathcal{L}_{\text{ELDiff}} = \mathcal{L}_{LDM} + \mathcal{L}_{\text{PixEvi}} + \mathcal{L}_{\text{TokCon}}$

Table 1: Evaluation results of our method against other methods based on SD v1.4. ELDiff holistically demonstrates the best performance regarding Multi-category Instance Composition, Photorealism, Realism and Efficiency. The best score is in blue , with the second-best score in green. (C) denotes the COCO instance validation set and (F) denotes the Flickr30K instance validation set.

Model	Multi-category Instance Composition					Photorealism				Efficiency
	OA↑	MG2↑	MG3↑	MG4↑	MG5↑	FID(C)↓	FID(F)↓	CLIP(C)↑	CLIP(F)↑	Latency↓
SD v1.4 [25]	0.2986	$0.9072_{1.33}$	$0.5074_{0.89}$	$0.1148_{0.45}$	$0.0088_{0.21}$	22.06	57.24	0.3022	0.3082	$7.54_{0.17}$
Composable [37]	0.2783	$0.6333_{0.59}$	$0.2187_{1.01}$	$0.0325_{0.45}$	$0.0023_{0.18}$	21.71	60.32	0.3123	0.2987	$13.81_{0.15}$
Layout [7]	0.4359	$0.9322_{0.69}$	$0.6015_{1.58}$	$0.1949_{0.88}$	$0.0227_{0.44}$	23.18	58.77	0.2851	0.3042	$18.89_{0.20}$
Structured [12]	0.2964	$0.9040_{1.06}$	$0.4864_{1.32}$	$0.1071_{0.92}$	$0.0068_{0.25}$	22.09	56.73	0.2706	0.2931	$7.74_{0.17}$
Attend-Excite [5]	0.4513	$0.9364_{0.76}$	$0.6510_{1.24}$	$0.2801_{0.90}$	$0.0601_{0.61}$	21.57	57.05	0.3018	0.3044	$25.43_{4.89}$
CoMat [23]	0.4732	$0.9241_{0.33}$	$0.7044_{1.11}$	$0.2566_{0.77}$	$0.0211_{0.24}$	22.43	58.49	0.3122	0.3044	$7.54_{0.33}$
IterComp [67]	0.4891	$0.9560_{1.21}$	$0.6742_{1.54}$	$0.2487_{0.41}$	$0.0167_{0.15}$	21.68	57.66	0.3047	0.3152	$7.54_{0.35}$
SynGen [52]	0.4570	$0.9570_{1.44}$	$0.6915_{0.68}$	$0.2638_{0.79}$	$0.0241_{0.46}$	22.35	56.94	0.3105	0.3201	$9.21_{0.26}$
InitNO [17]	0.3703	$0.9144_{0.67}$	$0.6731_{0.66}$	$0.2544_{0.30}$	$0.0311_{0.47}$	21.50	57.31	0.3120	0.3234	$12.14_{0.47}$
Self-Cross [49]	0.4461	$0.9357_{0.56}$	$0.7144_{1.68}$	$0.2615_{0.62}$	$0.0262_{0.38}$	23.11	57.26	0.3145	0.3154	$10.35_{0.53}$
TokenCompose [61]	0.5215	$0.9808_{0.40}$	$0.7616_{1.04}$	$0.2881_{0.95}$	$0.0328_{0.48}$	21.12	56.93	0.3112	0.3205	$7.56_{0.14}$
ELDiff(Ours)	0.5563	$0.9755_{0.65}$	$0.8041_{1.15}$	$0.3305_{1.58}$	$0.0931_{0.47}$	19.25	55.13	0.3276	0.3359	$7.54_{0.24}$

Table 2: Comparison results with training-free methods on T2I-CompBench.

Method	Attribute Binding			Object Relationship		Complex↑	Latency↓
	Color↑	Shape↑	Texture↑	Spatial↑	Non-Spatial↑		
SDXL	0.6734	0.5064	0.6243	0.2073	0.3166	0.3575	8.07s
SynGen [52]	0.7267	0.5249	0.6476	0.2387	0.3172	0.3944	11.62s
InitNO [17]	0.6823	0.5229	0.6547	0.2553	0.3147	0.4077	19.42s
R2F [46]	0.7135	0.5194	0.6625	0.2447	0.3170	0.4031	10.14s
Ours	0.7768	0.5663	0.6941	0.2496	0.3261	0.4252	8.07s
SD v2.1	0.5309	0.4447	0.4929	0.1355	0.3099	0.3185	3.25s
Self-Cross [49]	0.6344	0.5347	0.5681	0.2048	0.3116	0.3664	13.05s
CountGui [27]	0.5387	0.5371	0.5211	0.1339	0.3077	0.3296	11.74s
Ours	0.6604	0.5125	0.6287	0.2238	0.3204	0.3841	3.25s

4 Experiments

4.1 Datasets and Evaluation

Dataset for Finetuning. We follow the dataset setting in TokenCompose [61], which is a subset of COCO image-caption pairs [35]. To be specific, all unique images are from the Visual Spatial Reasoning dataset [36] since these image-caption pairs have relatively low ambiguity in its visual language and the rich diversity of object categories. The CLIP model [50] is used to choose the caption that has the highest semantic correspondence to its paired image. The Grounded-SAM [30, 39] is utilized to generate binary segmentation maps of all nouns from the captions. Finally, 4526 image-caption pairs and their corresponding binary segmentation maps are finally obtained.
Evaluation metrics. We measure the following metrics: VISOR [16], MULTIGEN [61], Fréchet Inception Distance (FID) [19], CLIP Score [32], Efficiency, T2I-CompBench [21], T2I-CompBench++ [20], GenEval [15], and GenEval 2 [26].

4.2 Implementation Details and Baseline Methods

Implementation Details. We follow the experiment settings outlined in TokenCompose [61], finetuning our ELDiff based on Stable Diffusion v1.4 [53]. All experiments are implemented with

Table 3: To evaluate the model generalization, we apply our finetuning strategy to SD v2.1 and SD v3.5 Medium, and report results on multi-category instance composition, FID, and CLIP score.

Model	Multi-category Instance Composition				Photorealism			
	OA↑	MG3↑	MG4↑	MG5↑	FID(C)↓	FID(F)↓	CLIP(C)↑	CLIP(F)↑
SD v2.1 frozen	0.4728	0.7014	0.2557	0.0327	21.79	56.83	0.3045	0.3159
SD v2.1 ft. w. $\mathcal{L}_{LDM}$	0.5509	0.7643	0.3207	0.0473	21.94	57.02	0.3122	0.3231
SD v2.1 ft. w. CoMat	0.5844	0.7858	0.3279	0.0496	21.44	56.52	0.3168	0.3185
SD v2.1 ft. w. IterComp	0.6341	0.8072	0.3544	0.0543	20.85	56.31	0.3155	0.3237
SD v2.1 ft. w. TokenCompose	0.6010	0.8048	0.3669	0.0571	20.77	56.43	0.3209	0.3264
SD v2.1 ft. w. Ours	0.7116	0.8870	0.5401	0.1392	20.09	55.12	0.3241	0.3348
SD v3.5 frozen	0.8084	0.9996	0.9696	0.7328	18.21	54.33	0.3185	32.66
SD v3.5 ft. w. $\mathcal{L}_{LDM}$	0.8010	0.9996	0.9702	0.7345	18.94	55.14	0.3174	32.43
SD v3.5 ft. w. CoMat	0.8122	0.9997	0.9733	0.7456	18.45	55.01	0.3148	32.93
SD v3.5 ft. w. IterComp	0.8208	0.9997	0.9814	0.7644	18.21	54.85	0.3183	32.52
SD v3.5 ft. w. TokenCompose	0.8233	0.9997	0.9839	0.7569	18.05	54.87	0.3198	32.87
SD v3.5 ft. w. Ours	0.8644	0.9997	0.9964	0.8171	17.66	54.79	0.3247	32.78

Table 4: Evaluation results about compositionality on T2I-CompBench. Based on SD v2.1 and SD v3.5 Medium, ELDiff consistently shows the best performance regarding attribute binding, object relationships, numeracy and complex.

Model	Attribute Binding			Object Relationship		Numeracy↑
	Color↑	Shape↑	Texture↑	Spatial↑	Non-Spatial↑	
SD v2.1 frozen	0.5309	0.4447	0.4929	0.1355	0.3099	0.4896
SD v2.1 ft. w. $\mathcal{L}_{LDM}$	0.5514	0.5007	0.5376	0.1644	0.3192	0.5042
SD v2.1 ft. w. CoMat	0.6146	0.4835	0.5691	0.1543	0.3144	0.5078
SD v2.1 ft. w. IterComp	0.6354	0.5011	0.6124	0.1674	0.3169	0.5136
SD v2.1 ft. w. TokenCompose	0.6063	0.4934	0.6131	0.1789	0.3185	0.5094
SD v2.1 ft. w. Ours	0.6604	0.5325	0.6487	0.2238	0.3204	0.5366
SD v3.5 frozen	0.8033	0.5831	0.7154	0.3102	0.4010	0.6043
SD v3.5 ft. w. $\mathcal{L}_{LDM}$	0.8152	0.5641	0.7184	0.3361	0.4122	0.5961
SD v3.5 ft. w. CoMat	0.8124	0.5914	0.7283	0.3681	0.4451	0.6037
SD v3.5 ft. w. IterComp	0.8177	0.6035	0.7358	0.3422	0.4295	0.6254
SD v3.5 ft. w. TokenCompose	0.8023	0.5748	0.7463	0.3651	0.4474	0.6243
SD v3.5 ft. w. Ours	0.8452	0.6241	0.7435	0.3925	0.4640	0.6613

PyTorch and carried out with NVIDIA GeForce RTX 3090 GPU. The number of training steps is 24000, the initial learning rate is 5e-6 with AdamW [40], and the batch size is 1 and 4 gradient accumulation steps on a single GPU. Apart from the original denoised target $\mathcal{L}_{LDM}$, $\mathcal{L}_{PixEvi}$ and $\mathcal{L}_{\text{PixEvi}}$ and $\mathcal{L}_{\text{TokCon}}$ are utilized in the finetuning process, where $(\lambda_1, \lambda_2, \lambda_3)$ is set to (5e-5, 5e-7, 5e-7$\cdot min\{1, n_{epoch}/100\}$ for $\mathcal{L}_{\text{PixEvi}}$ and (η_1, η_2) is set to (1e-4, 1e-4) for $\mathcal{L}_{\text{TokCon}}$. Binary masks are generated once offline using SAM before diffusion training, not in training process. No mask needed at testing. Training cost: one 3090 GPU, about 22 hours, comparable to Stable Diffusion.

Baseline Methods. To validate the effectiveness of ELDiff, we conduct comparisons on multiple diffusion backbones, including SD v1.4 [25], SD v2.1, SDXL [47], SD v3.5 [11], and Qwen-Image [62]. On these backbones, we compare ELDiff with several representative T2I methods, including Composable Diffusion [37], Layout Guidance Diffusion [7], Structured Diffusion [12], Attend-and-Excite [5], CoMat [23], IterComp [67], SynGen [52], InitNO [17], Self-Cross [49], and TokenCompose [61], R2F [46], and CountGui [27].

4.3 Main Results

Results of Multi-category Instance Composition: Object Accuracy (OA) & MULTI-GEN We quantitatively evaluate the compositionality of ELDiff based on OA (a metric from

Table 5: Ablation studies of different objectives with $\mathcal{L}_{LDM}$, $\mathcal{L}_{\text{PixEvi}}$, and $\mathcal{L}_{\text{TokCon}}$.

Model	$\mathcal{L}_{LDM}$	$\mathcal{L}_{\text{PixEvi}}$	$\mathcal{L}_{\text{TokCon}}$	Multi-category Instance Composition					Photorealism			
				OA↑	MG2↑	MG3↑	MG4↑	MG5↑	FID(C)↓	FID(F)↓	CLIP(C)↑	CLIP(F)↑
SD v1.4				0.2986	$0.9072_{1.33}$	$0.5074_{0.89}$	$0.1148_{0.45}$	$0.0088_{0.21}$	22.06	57.24	0.3022	0.3082
SD v1.4	✓			0.3821	$0.9144_{0.21}$	$0.6372_{1.11}$	$0.1956_{0.94}$	$0.1897_{0.32}$	24.57	60.14	0.3028	0.3106
SD v1.4	✓	✓		0.5466	$0.9821_{0.30}$	$0.7648_{0.88}$	$0.2934_{1.14}$	$0.3361_{0.21}$	21.17	55.98	0.3178	0.3246
SD v1.4	✓	✓	✓	0.5563	$0.9736_{0.65}$	$0.7660_{1.15}$	$0.2962_{1.58}$	$0.3530_{0.47}$	20.84	55.36	0.3204	0.3299

Table 6: Comparison results on GenEval using the stronger Qwen-Image backbone.

Model	Single Object	Two Object	Counting	Colors	Position	Attribute Binding	Overall
Qwen-Image [62]	0.99	0.92	0.89	0.88	0.76	0.77	0.87
ft. w. TokenCompose	1.00	0.93	0.89	0.89	0.77	0.79	0.88
ft. w. Ours	1.00	0.96	0.90	0.92	0.80	0.79	0.90

VISOR) [16] and MULTIGEN [61] compared to the outstanding T2I models. As demonstrated in Table 1, ELDiff achieves state-of-the-art performance on OA, which is 0.0348 higher than the suboptimal TokenCompose [61]. For MG2-5, it is clear that our ELDiff and TokenCompose show significant improvements apart from MG2. To verify whether our method is benefit for finetuning other SD versions, we add our fine-tuning strategy to Stable Diffusion v2.1 and SD v3.5 Medium. As shown in Table 3, the results of Multi-category Instance Composition and Photorealism indicate that our ELDiff significantly outperforms SD v2.1, SD v3.5, and other finetuning methods. On SD v2.1, compared to TokenCompose, our ELDiff improved OA by 0.1106 and MG4 by 0.1732. In addition, on SD v3.5, our method has also achieves competitive performance. These improvements are largely attributed to the consistency constraint between the text prompts and the image contents, which greatly enhances the model's ability of composing multiple object categories. By finetuning the Stable Diffusion through pixel evidence loss and token conflict loss, ELDiff provides reliable consistency constraint during the denoising process and achieves outstanding generation results involving multiple objects. For qualitative experiments, as shown in Fig. 4, ELDiff achieves a high level of image quality in conjunction with multi-category instance composition, compared with other T2I models.

Results of Photorealism: FID & CLIP Score As shown in Table 1, our ELDiff consistently outperforms other T2I models in both FID [19] and CLIP Score [32]. Specifically, ELDiff is 1.87 and 1.60 lower than the suboptimal method on Fid(C) and Fid(F), respectively. It also 0.0131 and 0.0125 higher than the suboptimal method on CLIP Score(C) and CLIP Score(F) respectively. We attribute these performance advantages to the proposed pixel evidence loss and token conflict loss, which enhances image photorealism when composing multiple categories of instances.

Results of Efficiency: Latency Compared to the standard T2I diffusion, our ELDiff does not require additional inference time during the inference stage. As shown in Table 1, ELDiff achieves the best generation performance using the least amount of inference time. The latency results in Table 1 represent the number of seconds required to generate an image with 50 DDIM steps.

Results of Attribute Binding and Object Relationship: T2I-CompBench As illustrated in the Table 4, our ELDiff observes significant gains in all seven evaluation tasks, compared with other finetuing methods. Based on SD v2.1, ELDiff outperforms the second-best method by 0.025, 0.0314, 0.0356, 0.0449, and 0.0230 on color, shape, texture, spatial, and numeracy. On SD3.5, ELDiff also consistently outperforms the compared approaches. It is notable that ELDiff improves attribute binding and object relationship by enhancing the model's multi-object composition ability without specifically optimizing color, shape, texture, and the relationships between objects.

Table 7: Comparison of Qwen-Image on GenEval 2.

Methods	Attribute↑	Count↑	Object↑	Position↑	Verb↑	Soft-TIFA$_{AM}$ ↑	Soft-TIFA$_{GM}$ ↑
Qwen-Image	84.11	67.98	99.30	60.82	37.50	81.58	34.57
ft. w. $\mathcal{L}_{LDM}$	84.33	68.47	99.46	60.93	38.07	81.90	35.11
ft. w. TokenCom	84.42	68.55	99.40	61.54	37.88	81.79	35.83
ft. w. Ours	85.41	70.67	99.58	61.97	39.40	83.13	37.94

Table 8: Comparison of Qwen-Image on T2I-CompBench++.

Methods	Color↑	Shape↑	Texture↑	2D Spatial↑	3D Spatial↑	Non-Spatial↑	Numeracy↑	Complex↑
Qwen-Image	83.92	59.41	75.68	45.12	46.31	31.26	77.05	39.44
ft. w. $\mathcal{L}_{LDM}$	84.74	59.48	75.83	45.69	46.44	31.34	77.40	39.50
ft. w. TokenCom	84.44	59.57	76.41	46.37	46.86	31.31	77.64	39.62
ft. w. Ours	85.25	61.67	76.17	48.55	47.61	31.53	78.12	40.24

4.4 Ablation Study

Importance of Pixel Evidence Loss $\mathcal{L}_{\text{PixEvi}}$ As shown in Table 5, we show the Multi-category Instance Composition and Photorealism results, aiming to identify the effectiveness of pixel evidence loss. It is clear that without the use of the pixel evidence loss, Multi-category Instance Composition and Photorealism have significantly worsened. This is because the text prompt and the object in the image are not aligned during the denoising process, leading to poor multi-categories generation.
Effectiveness of Token Conflict Loss $\mathcal{L}_{\text{TokCon}}$ We demonstrate the advantages of token conflict loss by introducing the token conflict loss into the pixel evidence loss. As shown in Table 5, compared to before adding the token conflict loss, the performance metrics of Multi-category Instance Composition and Photorealism after adding the token conflict loss achieves consistent improvements apart from MG2. This is because the token conflict loss avoids multi-object semantic overlap conflicts in the image when conducting the consistency constraint between text prompts and image contents.

The ablation visualization (Fig. 3) clearly shows that finetuning Stable Diffusion with only $\mathcal{L}_{LDM}$ struggles to accurately capture the target objects. When the pixel evidence loss $\mathcal{L}_{PixEvi}$ is incorporated, the model improves the reliability of object localization. Furthermore, adding the token conflict loss $\mathcal{L}_{TokCon}$ results in a enhancement of text-to-object consistency. The ablation visualization demonstrates the effectiveness of the proposed loss in capturing the objects corresponding to nouns in the text.

4.5 Comparison with Training-Free Methods

Table 2 reports the comparison with representative training-free methods on T2I-CompBench under SDXL and SD v2.1 backbones. Our ELDiff achieves the best overall performance across most metrics. On SDXL, compared to the second-best method, ELDiff improves attribute binding (color, shape, and texture) by 0.0504, 0.0414, and 0.0316, and obtains the highest complex score while maintaining competitive latency. Similar trends are observed on SD v2.1, where our method consistently outperforms Self-Cross and CountGui in both attribute binding and spatial relationships.

4.6 Comparison with Stronger T2I Backbone

Table 6 shows the GenEval results on the stronger Qwen-Image backbone. Specifically, we integrate both our method and the TokenCompose finetuning strategy into Qwen-Image for comparison.

Table 9: Results of user study on SD v2.1 and SD v3.5.

Method	Image Realism↑	Image Compositionality↑	Overall↑
SD v2.1	12.2%	5.8%	19.7%
SD v2.1 ft. w. TokenCompose	34.3%	36.8%	31.7%
SD v2.1 ft. w. Ours	53.5%	57.4%	48.6%
SD v3.5	30.1%	25.0%	29.6%
SD v3.5 ft. w. TokenCompose	27.4%	35.7%	31.5%
SD v3.5 ft. w. Ours	42.5%	42.6%	38.9%

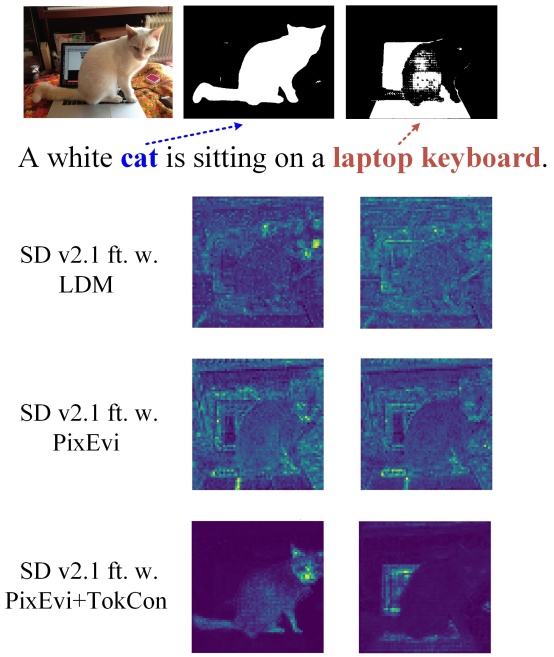

Fig. 3: Ablation visualization demonstrates that finetuning the Stable Diffusion with only $\mathcal{L}_{LDM}$ does not clearly identify the target objects. By introducing $\mathcal{L}_{PixEvi}$ and $\mathcal{L}_{TokCon}$, the model shows substantial improvement in text-object correspondence.

Our method consistently outperforms Qwen-Image and TokenCompose. The overall score increases from 0.87 to 0.90. In addition, Table 7 shows that our finetuning method consistently outperforms Qwen-Image on GenEval 2, with gains of 1.55 and 3.37 on Soft-TIFA$_{\text{AM}}$ and Soft-TIFA$_{\text{GM}}$. We used Qwen3-VL-8B [2] as the VLM evaluator. On T2I-ComBench++ (Table 8), compared with Qwen-Image, it improves Color, Shape, Texture, 2D Spatial, 3D Spatial, Non-Spatial, Numeracy, and Complex scores by 1.33, 2.26, 0.49, 3.43, 1.30, 0.27, 1.07, and 0.80, respectively. Additionally, compared with SFT loss ($\mathcal{L}_{LDM}$), our method also shows consistent improvements. This demonstrates that our ELDiff has strong generalization capability on the stronger backbone.

4.7 User Study

We also conducted user studies on SD v2.1 and SD v3.5. This study involves 11 volunteers with expertise in image processing. Participants were asked to select the best image based on realism,

Fig. 4: Qualitative comparison between our ELDiff and the other T2I models.

compositionality, and overall quality. We use 1,000 prompts from DSG1K [8] for evaluation. Voting results are shown in Table 9. Across two backbones, ELDiff consistently ranked first in user evaluations.

5 Conclusion

In this paper, we formulate the problems of segmentation map bias and semantic overlap conflict in performing the consistency constraint between text prompts and image contents, and propose ELDiff, an end-to-end T2I diffusion model fine-tuning strategy equipped with consistency constraint between text prompts and image contents. We identify the causes of these two problems and propose two key components to explicitly address them. The pixel evidence loss judges the reliability of consistency supervision through pixel-level uncertainty metric. Besides, we introduce the token conflict loss to address the semantic overlap conflict among objects in the image. Despite achieving superior generation results, ELDiff shortcomings in attribute binding task. In future work, we will continue to improve this framework by considering applying constraints on color, shape, texture, etc.

Acknowledgements

This work was partly supported by Taishan Scholars Program of Shandong Province and Shandong Province "Double Hundred Talent Plan" on 100 Foreign Experts and 100 Foreign Expert Teams (Grant No.WSR2023049).

References

1. Amini, A., Schwarting, W., Soleimany, A., Rus, D.: Deep evidential regression. Advances in neural information processing systems **33**, 14927–14937 (2020)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Bao, W., Yu, Q., Kong, Y.: Evidential deep learning for open set action recognition. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 13349–13358 (2021)
4. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22563–22575 (2023)
5. Chefer, H., Alaluf, Y., Vinker, Y., Wolf, L., Cohen-Or, D.: Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. ACM transactions on Graphics (TOG) **42**(4), 1–10 (2023)
6. Chen, J., Jincheng, Y., Chongjian, G., Yao, L., Xie, E., Wang, Z., Kwok, J., Luo, P., Lu, H., Li, Z.: Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In: The Twelfth International Conference on Learning Representations
7. Chen, M., Laina, I., Vedaldi, A.: Training-free layout control with cross-attention guidance. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision. pp. 5343–5353 (2024)
8. Cho, J., Hu, Y., Garg, R., Anderson, P., Krishna, R., Baldridge, J., Bansal, M., Pont-Tuset, J., Wang, S.: Davidsonian scene graph: Improving reliability in fine-grained evaluation for text-to-image generation. arXiv preprint arXiv:2310.18235 (2023)
9. Dempster, A.P.: Upper and lower probabilities induced by a multivalued mapping. In: Classic works of the Dempster-Shafer theory of belief functions, pp. 57–72 (2008)
10. Du, C., Li, Y., Qiu, Z., Xu, C.: Stable diffusion is unstable. Advances in Neural Information Processing Systems **36**, 58648–58669 (2023)
11. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
12. Feng, W., He, X., Fu, T.J., Jampani, V., Akula, A.R., Narayana, P., Basu, S., Wang, X.E., Wang, W.Y.: Training-free structured diffusion guidance for compositional text-to-image synthesis. In: The Eleventh International Conference on Learning Representations
13. Feng, Z., Zhang, Z., Yu, X., Fang, Y., Li, L., Chen, X., Lu, Y., Liu, J., Yin, W., Feng, S., et al.: Ernie-vilg 2.0: Improving text-to-image diffusion model with knowledge-enhanced mixture-of-denoising-experts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10135–10145 (2023)
14. Fu, W., Chen, Y., Liu, W., Yue, X., Ma, C.: Evidence reconciled neural network for out-of-distribution detection in medical images. In: International conference on medical image computing and computer-assisted intervention. pp. 305–315 (2023)
15. Ghosh, D., Hajishirzi, H., Schmidt, L.: Geneval: An object-focused framework for evaluating text-to-image alignment. Advances in Neural Information Processing Systems **36**, 52132–52152 (2023)
16. Gokhale, T., Palangi, H., Nushi, B., Vineet, V., Horvitz, E., Kamar, E., Baral, C., Yang, Y.: Benchmarking spatial relationships in text-to-image generation. arXiv preprint arXiv:2212.10015 (2022)
17. Guo, X., Liu, J., Cui, M., Li, J., Yang, H., Huang, D.: Initno: Boosting text-to-image diffusion models via initial noise optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9380–9389 (2024)
18. Hao, Y., Chi, Z., Dong, L., Wei, F.: Optimizing prompts for text-to-image generation. Advances in Neural Information Processing Systems **36**, 66923–66939 (2023)
19. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)

20. Huang, K., Duan, C., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-compbench++: An enhanced and comprehensive benchmark for compositional text-to-image generation. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(5), 3563–3579 (2025)
21. Huang, K., Sun, K., Xie, E., Li, Z., Liu, X.: T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. Advances in Neural Information Processing Systems **36**, 78723–78747 (2023)
22. Ji, W., Li, J., Bi, Q., Liu, T., Li, W., Cheng, L.: Segment anything is not always perfect: An investigation of sam on different real-world applications. Machine Intelligence Research **21**(4), 617–630 (2024)
23. Jiang, D., Song, G., Wu, X., Zhang, R., Shen, D., Zong, Z., Liu, Y., Li, H.: Comat: Aligning text-to-image diffusion model with image-to-text concept matching. Advances in Neural Information Processing Systems **37**, 76177–76209 (2024)
24. Jøsang, A., Hankin, R.: Interpretation and fusion of hyper opinions in subjective logic. In: 2012 15th International Conference on Information Fusion. pp. 1225–1232 (2012)
25. Jsang, A.: Subjective Logic: A formalism for reasoning under uncertainty. Springer Publishing Company, Incorporated (2018)
26. Kamath, A., Chang, K.W., Krishna, R., Zettlemoyer, L., Hu, Y., Ghazvininejad, M.: Geneval 2: Addressing benchmark drift in text-to-image evaluation. arXiv preprint arXiv:2512.16853 (2025)
27. Kang, W., Galim, K., Koo, H.I., Cho, N.I.: Counting guidance for high fidelity text-to-image synthesis. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 899–908. IEEE (2025)
28. Ke, L., Tai, Y.W., Tang, C.K.: Deep occlusion-aware instance segmentation with overlapping bilayers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4019–4028 (2021)
29. Kim, J., Esmaeili, E., Qiu, Q.: Text embedding is not all you need: Attention control for text-to-image semantic alignment with text self-attention maps. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8031–8040 (2025)
30. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
31. Kopetzki, A.K., Charpentier, B., Zügner, D., Giri, S., Günnemann, S.: Evaluating robustness of predictive uncertainty estimation: Are dirichlet-based models reliable? In: International Conference on Machine Learning. pp. 5707–5718 (2021)
32. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1931–1941 (2023)
33. Liew, J.H., Yan, H., Zhou, D., Feng, J.: Magicmix: Semantic mixing with diffusion models. arXiv preprint arXiv:2210.16056 (2022)
34. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 300–309 (2023)
35. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755 (2014)
36. Liu, F., Emerson, G., Collier, N.: Visual spatial reasoning. Transactions of the Association for Computational Linguistics **11**, 635–651 (2023)
37. Liu, N., Li, S., Du, Y., Torralba, A., Tenenbaum, J.B.: Compositional visual generation with composable diffusion models. In: European conference on computer vision. pp. 423–439 (2022)
38. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European conference on computer vision. pp. 38–55 (2024)
39. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Jiang, Q., Li, C., Yang, J., Su, H., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In: European conference on computer vision. pp. 38–55. Springer (2024)

40. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representations
41. Ma, W.D.K., Lahiri, A., Lewis, J.P., Leung, T., Kleijn, W.B.: Directed diffusion: Direct control of object placement through attention guidance. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 4098–4106 (2024)
42. Malinin, A., Gales, M.: Predictive uncertainty estimation via prior networks. Advances in neural information processing systems **31** (2018)
43. Pan, Q., Li, Z., Yang, G., Yang, Q., Ji, B.: Evivlm: When evidential learning meets vision language model for medical image segmentation. IEEE Transactions on Medical Imaging **45**(4), 1369–1382 (2025)
44. Pandey, D.S., Yu, Q.: Multidimensional belief quantification for label-efficient meta-learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14391–14400 (2022)
45. Pandey, D.S., Yu, Q.: Evidential conditional neural processes. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 37, pp. 9389–9397 (2023)
46. Park, D., Kim, S., Moon, T., Kim, M., Lee, K., Cho, J.: Rare-to-frequent: Unlocking compositional generation power of diffusion models on rare concepts with llm guidance. In: International Conference on Learning Representations
47. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952 (2023)
48. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. In: The Eleventh International Conference on Learning Representations
49. Qiu, W., Wang, J., Tang, M.: Self-cross diffusion guidance for text-to-image synthesis of similar subjects. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23528–23538 (2025)
50. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763 (2021)
51. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
52. Rassin, R., Hirsch, E., Glickman, D., Ravfogel, S., Goldberg, Y., Chechik, G.: Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. Advances in Neural Information Processing Systems **36**, 3536–3559 (2023)
53. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
54. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding. Advances in neural information processing systems **35**, 36479–36494 (2022)
55. Sensoy, M., Kaplan, L., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. Advances in neural information processing systems **31** (2018)
56. Sentz, K., Ferson, S.: Combination of evidence in dempster-shafer theory (2002)
57. Shi, W., Zhao, X., Chen, F., Yu, Q.: Multifaceted uncertainty estimation for label-efficient deep learning. Advances in neural information processing systems **33**, 17247–17257 (2020)
58. Singer, U., Polyak, A., Hayes, T., Yin, X., An, J., Zhang, S., Hu, Q., Yang, H., Ashual, O., Gafni, O., et al.: Make-a-video: Text-to-video generation without text-video data. In: The Eleventh International Conference on Learning Representations
59. Sun, J., Fu, D., Hu, Y., Wang, S., Rassin, R., Juan, D.C., Alon, D., Herrmann, C., Van Steenkiste, S., Krishna, R., et al.: Dreamsync: Aligning text-to-image generation with image understanding feedback. arXiv preprint arXiv:2311.17946 (2023)
60. Tomani, C., Buettner, F.: Towards trustworthy predictions from deep neural networks with fast adversarial calibration. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 9886–9896 (2021)

61. Wang, Z., Sha, Z., Ding, Z., Wang, Y., Tu, Z.: Tokencompose: Text-to-image diffusion with token-level supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8553–8564 (2024)
62. Wu, C., Li, J., Zhou, J., Lin, J., Gao, K., Yan, K., Yin, S.m., Bai, S., Xu, X., Chen, Y., et al.: Qwen-image technical report. arXiv preprint arXiv:2508.02324 (2025)
63. Wu, T.H., Lian, L., Gonzalez, J.E., Li, B., Darrell, T.: Self-correcting llm-controlled diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6327–6336 (2024)
64. Xu, Y., Zhao, Y., Xiao, Z., Hou, T.: Ufogen: You forward once large scale text-to-image generation via diffusion gans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8196–8206 (2024)
65. Xue, Z., Song, G., Guo, Q., Liu, B., Zong, Z., Liu, Y., Luo, P.: Raphael: Text-to-image generation via large mixture of diffusion paths. Advances in Neural Information Processing Systems **36**, 41693–41706 (2023)
66. Yang, L., Liu, J., Hong, S., Zhang, Z., Huang, Z., Cai, Z., Zhang, W., Cui, B.: Improving diffusion-based image synthesis with context prediction. Advances in Neural Information Processing Systems **36**, 37636–37656 (2023)
67. Zhang, X., Yang, L., Li, G., Cai, Y., Xie, J., Tang, Y., Yang, Y., Wang, M., Cui, B.: Itercomp: Iterative composition-aware feedback learning from model gallery for text-to-image generation. In: International Conference on Learning Representations. pp. 8608–8628 (2025)
68. Zhang, Y., Xing, J., Lo, E., Jia, J.: Real-world image variation by aligning diffusion inversion chain. Advances in Neural Information Processing Systems **36**, 30641–30661 (2023)
69. Zhao, X., Chen, F., Hu, S., Cho, J.H.: Uncertainty aware semi-supervised learning on graph data. Advances in neural information processing systems **33**, 12827–12836 (2020)
70. Zhong, J., Pan, Q., Zhou, X., Lin, J., Zhuang, X.: Evidential learning driven breast tumor segmentation with stage-divided vision-language interaction. Neurocomputing p. 132411 (2025)
71. Zhou, Y., Zhou, D., Wang, Y., Feng, J., Hou, Q.: Maskdiffusion: Boosting text-to-image consistency with conditional mask. International Journal of Computer Vision **133**(5), 2805–2824 (2025)