MITIGATE REPLICATION AND COPYING IN DIFFUSION MODELS WITH GENERALIZED CAPTION AND DUAL FUSION ENHANCEMENT

Chenghao Li*, Dake Chen*, Yuke Zhang*, Peter A. Beerel

University of Southern California, Los Angeles, CA

ABSTRACT

While diffusion models demonstrate a remarkable capability for generating high-quality images, their tendency to 'replicate' training data raises privacy concerns. Although recent research suggests that this replication may stem from the insufficient generalization of training data captions and duplication of training images, effective mitigation strategies remain elusive. To address this gap, our paper first introduces a generality score that measures the caption generality and employ large language model (LLM) to generalize training captions. Subsequently, we leverage generalized captions and propose a novel dual fusion enhancement approach to mitigate the replication of diffusion models. Our empirical results demonstrate that our proposed methods can significantly reduce replication by 43.5% compared to the original diffusion model while maintaining the diversity and quality of generations.

Index Terms— Generative model, diffusion model, training data privacy, replication mitigation

1. INTRODUCTION

While denoising diffusion models excel in generative tasks [1, 2, 3, 4] and are popular in commercial applications, they face ongoing legal challenges and privacy risks, notably for their tendency to memorize training data, a phenomenon known as *replication* [5, 6]. Replication is more likely to appear when fine-tuning a large pre-trained model with a small dataset [5]. The common notion attributes the issue of data replication primarily to the lack of general caption of the training images [6], and the duplication of training images [7]. Specifically, diffusion models tend to memorize images that are either associated with more specific captions [6] or displayed multiple times within the training dataset [5]. Inspired by these insights, we aim to generalize the captions and curtail the use of duplicate images during training to diminish data replication and mitigate the risk of privacy breaches.

The concept of the generality of natural language is abstract and lacks widely accepted quantitative metrics due to the subjective variations among human evaluators. In response, we introduce a *generality score* which comprises four

---

*Authors contributed equally

metrics: specificity score, broadness score, tense modality score, and abstraction score, to quantify the generality of a sentence, and facilitate an analysis of its correlation with data replication.

Although simplifying captions to a more general form is a direct in approach, it becomes labor-intensive when dealing with a large volume of captions. The emergence of Large Language Models (LLMs) [8] offers a solution, reducing the need for costly and time-intensive manual processes across multiple research domains [9, 10, 11, 12, 13]. For example, WizardLM [9] uses the LLM model, to iteratively regenerate instructions to refine LLaMA [14]. Inspired by this development, we employ LLM to generalize the captions of the diffusion model, achieving a 32.18% reduction in replication compared to the original training captions.

Furthermore, we introduce a novel dual fusion enhancement training approach, which involves the stochastic fusion of specific object features with the original image in the latent space, while concurrently fusing the corresponding label embedding with the caption. Besides reducing the duplication of training images, this dual fusion process introduces diversity and randomness into the training data, effectively alleviating data replication concerns. Our empirical results demonstrate that the proposed dual fusion approach, in comparison to the original model, can reduce replication by 43.5% while preserving the comparable generation diversity and quality.

We summarize our contributions as follows: (1) We introduce a generality score combining attributes from four dimensions that measure caption generality. (2) We generate more general captions using LLM and empirically present the correlation between generality, specificity, and replication. (3) We propose a novel dual fusion enhancement approach to mitigate replication, and the experimental results demonstrate its effectiveness.

2. PRELIMINARIES

2.1. Problem statement

We denote a text-to-image diffusion model as *Diff*. The diffusion forward process involves iteratively introducing noise to each training image and transforming it into Gaussian noise. The training objective is to predict the distribution of the in-

corporated noise at different time steps. The generative process is expressed as $g_n = \mathit{Diff}(p_n, r)$, where g_n denotes the generated image, p_n corresponds to the associated caption, and r represents the initial random noise. In line with Somepalli et al.'s investigation [6], it has been observed that $X = \{x_n\}_{n=1}^N$ and $G = \{g_n\}_{n=1}^N$ exhibit a significant degree of memorization, defined as replication. Prior work has proposed several defense methods [6]. The Multiple Captions (MC) uses BLIP [15] to generate 20 alternative captions for each image and randomly pick one during training. Gaussian Noise (GN) add Gaussian noise to the text embedding. Random Caption Replacement (RC) substitutes an image's caption with a random sequence of words. Caption Word Repetition (CWR) randomly choose a word from the provided caption and randomly insert it at a different position within the same caption.

2.2. Evaluation of replication

Keep in line with [6], we use *replication score*, denoted as R, to assess the content replication of a diffusion model *Diff*. Given the original training dataset X and the generated image set G, R is defined as the threshold satisfying: $Pr(S < R) = F_S(S < R) = 0.95$, where S is the random variable of the similarity score of two images x and g from X and G, respectively, whose probability density function (PDF) is $f_S(S)$ and cumulative distribution function (CDF) is $F_S(S < s)$. The similarity score is calculated based on features extracted by the SSCD [16], a self-supervised copy detection model, with the official checkpoint. There is a 95% probability that the sample of S will fall below the cutoff point R. This ensures that low similarity scores do not distort the overall dataset replication score. Specifically, a higher value of R signifies a greater extent of replication.

3. STRATEGIES FOR MITIGATING REPLICATION IN DIFFUSION MODELS

In this section, we introduce our proposed generality score for captions, the process of generating generalized captions using an LLM, and our proposed dual fusion enhancement training framework.

3.1. Semantic generality metrics

Generality is inherently abstract in nature. To quantify it, we introduce four fundamental and intuitive metrics: *specificity of information*, *broadness of terms*, *tense and modality*, and *degree of abstraction*, as mathematically defined in Table 1. Their details are further elaborated below.

Specificity of Information (SI). In essence, a sentence tends to be more general when it contains fewer details. These specific pieces of information are typically presented

Metric	Equation	Range
SI	$1 - \frac{Ent + Num}{N_{word}}$	$[0, 1]$
BT	$min\left(\frac{\sum_i hypo(noun_i)}{N_{word} \times 2Avg_{global}}, 1\right)$	$[0, 1]$
TM	$\frac{Count_{pres\&ind}*}{V_{total}}$	$[0, 1]$
DA	$min\left(\frac{DA_{caption}}{2DA_{global}}, 1\right)$	$[0, 1]$

$*$ TM is set to 0.5, indicating neutrality, when $V_{total} = 0$.

Table 1. Mathematical description of the proposed metrics

as named entities such as individual names, locations, specific times, and numerical values [17]. For instance, *"Johnson went to USC at 10 am"* is more specific than *"Students always go to school in the morning"*. The metric computation is shown in Table 1 and is based on the number of named entities in a caption Ent, the number of numerical details Num, and the length of a caption N_{word}, all obtained using the spaCy library [18].

Broadness of Terms (BT). In hierarchical semantics, broader categories are typically situated higher in the hierarchy and include more specific terms which are referred to as hyponyms [19]. For instance, *"animal"* is more general than *"dog"*. We quantify this aspect as in Table 1 using the number of hyponyms of each noun $hypo(noun)$ in the caption using WordNet [19]. Avg_{global} is the global average hyponyms count across a lexicon of $30k$ words, ignoring the less-relevant terms [19].

Tense and Modality (TM). Verbs play a crucial role in determining generality, with sentences in the present tense and indicative mood often being more general. For instance, *"Birds can fly"* is more general than *"The bird was flying"*. To quantify this aspect, we calculate the proportion of verbs in the present tense and indicative mood in the sentence using spaCy library [18].

Degree of Abstraction (DA). Abstract concepts or words often possess greater generality compared to their concrete counterparts. For example, *"Kindness is a good virtue"* is more general than *"It's good to often greet people with good morning"*. To compute this we use WordNet [19] to provide a depth measure indicating the degree of abstraction associated with each noun in the caption and compute the ratio of $DA_{caption}$, the average depth of the synonym sets(synsets) for a caption, to DA_{global}, the global average depth.

Generality Score (GS). We aggregate these metrics into a unified measure for each caption by first scaling them linearly into the range $[0, 10]$ and then computing their average.

3.2. Generalize captions

Creating captions that are general yet convey the semantics of visual images presents a challenge in NLP due to the absence of efficient caption-generalization frameworks. Manual labeling is costly and lacks uniform standards. However, Large
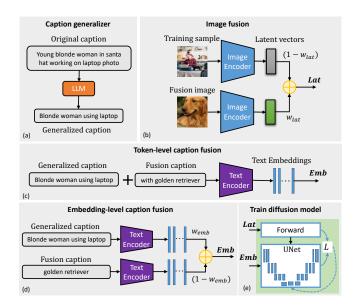
Fig. 1. Overview of the proposed methods. (a) Generalize captions with LLM, (b) image fusion, (c) token-level caption fusion, (d) embedding-level caption fusion, (e) train the diffusion model with dual fusion enhancement.

Language Models (LLMs) like GPT [20, 21, 14, 8] have proven to be a cost-effective solution, with their inference capabilities leading to effective text generalization. Therefore, we instructed GPT-3.5 to generalize captions at two distinct levels of generality using the following guidelines:

General caption: Convert this caption of an image to a more general caption:[given caption];

5-word caption: Make this caption of an image extremely general (result in less than 5 words): [given caption].

After generating, the generalized captions are assessed with the proposed generality score, allowing us to evaluate the effectiveness of the generalization method. Table 2 contrasts SI, BT, TM, DA, and GS values between an original caption and its LLM-generalized versions.

3.3. Dual fusion enhancement

An overview of the proposed dual fusion enhancement is illustrated in Fig. 1. This method utilizes generalized captions to enhance caption generality and incorporates a fusion dataset. This dataset supplies image-text pairs that, when fused with the fine-tuning samples, boost diversity and randomness while minimizing potential image duplication in the fine-tuning dataset.

We define the fine-tuning dataset as $D^{ft} = \{(x_i^{ft}, y_i^{ft})\}_{i=1}^I$ and the fusion dataset as $D^{fu} = \{(x_j^{fu}, y_j^{fu})\}_{j=1}^J$, where (x_i^{ft}, y_i^{ft}) and (x_j^{fu}, y_j^{fu}) are the image-text pair in D^{ft} and D^{fu} respectively. During the dual fusion training process, both the fine-tuning image x_i^{ft} and a randomly selected fusion image x_j^{fu} are passed through a visual encoder, whose

function is denoted as VE, yielding their respective latent representations. Subsequently, we weighted-fuse these two latent representations to obtain an enhanced latent representation, which is expressed as follows:

$$Lat_i = (1 - w_{lat})VE(x_i^{ft}) + w_{lat}VE(x_j^{fu}), \quad (1)$$

where w_{lat} is the image latent fusion weight.

Regarding text conditioning, we introduce two fusion modes. In *token-level fusion*, we append the fusion caption y_j^{fu} directly to the end of the original caption y_i^{ft}. The resulting concatenated caption is then fed into the text encoder, whose function is denoted as TE, to obtain the updated text embedding. In *embedding-level fusion*, we separately pass both y_i^{ft} and y_j^{fu} through the text encoder, resulting in two distinct text embeddings, which are then weighted-combined at the embedding level to create a unified text representation, represented as follows:

$$Emb_i = (1 - w_{emb})TE(y_i^{ft}) + w_{emb}TE(y_j^{fu}), \quad (2)$$

where w_{emb} is the text embedding fusion weight. The loss function component for diffusion model is as follows:

$$L(\theta) = E_{\varepsilon,t}[||\varepsilon - \varepsilon_\theta(Lat_i^{(t)}, t, Emb_i)||^2], \quad (3)$$

where ε is the noise added during the forward pass, ε_θ is the predicted noise, and t is the time step.

4. EXPERIMENTS

Experimental setup. In line with [5, 6], we focus on reducing replication when fine-tuning a pre-trained diffusion model with small datasets. We fine-tune Stable Diffusion v2.1 [2], pre-rained on LAION [22], with a subset of LAION-2B [22] with $10k$ randomly selected samples. We use Tiny-Imagenet [23] as the fusion dataset in dual fusion enhancement, which consists of 200 classes, with 500 images in each class. We freeze all other modules and only fine-tune the UNet component in the Stable Diffusion. To ensure comparability, we follow the the model hyperparameters in [6], including $100k$ iterations and a learning rate of $5e^{-6}$. We use replication score R to evaluate replication on $10k$ generated images, and the Frechet Inception Distance (*FID*), which serves as an indicator of the quality and diversity of the model's generated images [24].

Caption generality. We evaluate the performance of original, general, and 5-word captions in Table 4. The 5-word captions standout, registering the highest generality score and the lowest replication score, while maintaining a competitive *FID*. This highlights the role of caption generality in reducing replication. Thus, we selected the 5-word caption for our subsequent experiments.

Comparison with prior-art. In Table 3, we compare our approach to prior mitigation methods, namely multiple captions (MC), Gaussian noise (GN), random caption replacement (RC), and caption word repetition (CWR) [6]. We use

LLM guide	Caption	SI↑	BT↑	TM↑	DA↑	GS↑
Original	"2-Ton Multi-Directional Roller Head Pipe Welding Stands"	8.18	3.67	5.0	5.15	5.50
General	"Versatile Pipe Welding Stands for Heavy-Duty Applications"	10.0	3.96	10.0	5.34	7.33
5-word	"Pipe welding stands in action"	10.0	10.0	10.0	4.61	**8.65**

Table 2. Example of generalized captions and their generality score

Method	Baseline	MC [6]	GN [6]	RC [6]	CWR [6]	**Ours**
R↓	0.662	0.420	0.596	*0.565*	*0.614*	**0.374**
FID↓	17.394	16.831	19.504	-	-	**15.997**

∗ The italic replication scores are taken from the paper [6]

Table 3. Comparison of replication mitigation strategies

Caption Metrics	Original	General	**5-word**
Generality score↑	5.202	5.732	**5.846**
R↓	0.662	0.529	**0.449**
FID↓	17.394	**16.837**	17.406

Table 4. Comparison of caption generality levels

embedding-level fusion with $w_{lat} = 0.1$ and $w_{emd} = 0.5$ and achieves a significant reduction in replication by 43.50% compared to baseline and 10.95% compared to the leading method, MC, with competitive *FID*.

Visualize the generated images. Figure 2 showcases five training samples alongside the generated images from both the original diffusion model and our optimized model. Notably, our model considerably diminishes replication, producing images that deviate more from the original training data.

The effects of the fusion weights. We report replication scores and *FID* values for various fusion weights in Table 5, covering both token-level and embedding-level scenarios. For token-level dual fusion enhancement, there exists a trade-off between replication score and *FID* as w_{lat} increases. This trade-off emerges since a higher w_{lat} allows the fusion image to influence the fine-tuning sample more profoundly, potentially compromising image generation quality. In the embedding-level scenario, using consistent fusion weights for both visual and text elements, we observed a similar balance between replication and *FID*. To curtail the influence of the fusion image, we set the latent weight at 0.1 and progressively adjusted the embedding weight. Our findings suggest that by limiting the dominance of the fusion image and elevating the contribution of the fusion caption, we can strike a balance between R and *FID*. This balance is achieved since minimal image fusion can effectively prevent training image duplication, while a greater degree of caption fusion introduces more caption generality.

Ablation study. In Table 6, we present the ablation study of our proposed methods. The 5-word captions yield a notable 32.18% improvement in replication over the baseline. Additionally, the dual fusion enhancement with $w_{lat} = 0.1$ and $w_{emb} = 0.5$ contributes another 16.7% improvement in replication, all while preserving a competitive *FID*.

Fig. 2. Comparison of generated images

Token-level		Embedding-level				Embedding-level			
w_{lat}	R↓	FID↓	w_{lat}	w_{emb}	R↓	FID↓	w_{lat}	w_{emb}	R↓	FID↓
0.1	0.454	**14.383**	0.1	0.1	0.417	**15.680**	0.1	0.1	0.417	**15.680**
0.25	0.286	60.767	0.25	0.25	0.361	19.642	0.1	0.3	0.384	16.665
0.75	**0.226**	73.030	0.5	0.5	**0.257**	61.953	0.1	0.5	**0.374**	15.997

Table 5. The effects of fusion weights

Model	Ablation		Metrics	
	Generalized Caption	Dual Fusion	Replication	FID
Original	✗	✗	0.662	17.394
Ablated	✗	✓	0.416	16.869
Ablated	✓	✗	0.449	17.406
Ours	✓	✓	0.374	15.997

Table 6. Result of ablation study (✗ indicates the module is suppressed and ✓ means the module is applied)

5. CONCLUSIONS

In this paper, we propose a novel approach to bolster the privacy of diffusion models utilizing broader captions and introduce a metric for quantifying caption generality. Our findings underscore that employing generalized captions can substantially reduce data replication of diffusion models. Additionally, our proposed dual fusion enhancement method significantly mitigates data duplication while maintaining the diversity and quality of generations. The extensive experimental results show that our mitigation strategies surpass state-of-the-art methodologies in effectively reducing data replication. In future work, we will explore the use of the generality score to guide the caption generalization process and iteratively enhance the generality of each caption.

6. REFERENCES

[1] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

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

[3] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, pp. 3, 2022.

[4] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala, "Adding conditional control to text-to-image diffusion models," .

[5] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein, "Diffusion art or digital forgery? investigating data replication in diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6048–6058.

[6] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein, "Understanding and mitigating copying in diffusion models," *arXiv preprint arXiv:2305.20086*, 2023.

[7] Nikhil Vyas, Sham Kakade, and Boaz Barak, "Provable copyright protection for generative models," *arXiv preprint arXiv:2302.10870*, 2023.

[8] OpenAI, "Gpt-4 technical report," 2023.

[9] Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang, "WizardLM: Empowering large language models to follow complex instructions," *arXiv preprint arXiv:2304.12244*, 2023.

[10] Ziyang Luo, Can Xu, Pu Zhao, Qingfeng Sun, Xiubo Geng, Wenxiang Hu, Chongyang Tao, Jing Ma, Qingwei Lin, and Daxin Jiang, "WizardCoder: Empowering code large language models with evol-instruct," *arXiv preprint arXiv:2306.08568*, 2023.

[11] Suvir Mirchandani, Fei Xia, Pete Florence, Brian Ichter, Danny Driess, Montserrat Gonzalez Arenas, Kanishka Rao, Dorsa Sadigh, and Andy Zeng, "Large language models as general pattern machines," *arXiv preprint arXiv:2307.04721*, 2023.

[12] Zhen Qin et al., "Large language models are effective text rankers with pairwise ranking prompting," *arXiv preprint arXiv:2306.17563*, 2023.

[13] Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen, "Large language models as optimizers," 2023.

[14] Hugo Touvron et al., "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[15] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*. PMLR, 2022, pp. 12888–12900.

[16] Ed Pizzi, Sreya Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze, "A self-supervised descriptor for image copy detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14532–14542.

[17] Annie Louis and Ani Nenkova, "General versus specific sentences: automatic identification and application to analysis of news summaries," 2011.

[18] Matthew Honnibal and Ines Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing," 2017.

[19] George A Miller, "WordNet: a lexical database for English," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[20] Long Ouyang et al., "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27730–27744, 2022.

[21] Yuntao Bai et al., "Training a helpful and harmless assistant with reinforcement learning from human feedback," *arXiv preprint arXiv:2204.05862*, 2022.

[22] Christoph Schuhmann et al., "LAION-5B: An open large-scale dataset for training next generation image-text models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25278–25294, 2022.

[23] Ya Le and Xuan S. Yang, "Tiny ImageNet Visual Recognition Challenge," 2015.

[24] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter, "GANS trained by a two time-scale update rule converge to a local Nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.