Rethinking Diffusion-based Augmentation: Why Single Prompt Fails

Chanhee Lee Jeonghwan Cho Suhyun Kim Jeongyeon Kim
Sungkyunkwan University
Department of Applied Artificial Intelligence
{leechanhye, kyjjhh1, suhun0011, wjddus0203}@skku.edu

Abstract

Diffusion models have recently gained attention as a promising tool for data augmentation because they can synthesize new samples that resemble the original data while preserving class labels. However, despite the strong performance reported in prior works, it remains unclear which factors actually drive these gains. In particular, the contribution of the generated samples, the mixing procedure, and the diversity introduced by prompts has not been fully examined. Motivated by this gap, we conduct a controlled study in the one-prompt setting, where the diversity of the generated images is intentionally limited. Our analysis reveals that diffusion-based augmentation does not broaden the training distribution under this regime and often reduces performance when combined with real images. To investigate potential causes, we introduce controlled variants of DiffuseMix that isolate scheduling, fractal-based integration, and semantic guided blending. None of these components recover the performance of the vanilla model, suggesting that previously showed improvements are more closely tied to large scale offline generation than to semantic or saliency driven mixing. Our findings provide a clearer understanding of diffusion-based augmentation and highlight the need for further analysis in full prompt environments. The code is available at https://github.com/iontail/gdl_term.git.

1. Introduction

Data augmentation is a central technique for improving generalization in supervised learning. Many established approaches operate directly in the data space by applying transformations such as cropping, flipping, or color jittering. More advanced strategies include cut based augmentation, which masks or removes selected regions [3, 10, 12, 21, 25, 27], and mix based augmentation, which blends two images to generate intermediate samples [5, 8, 26]. While these methods enrich the training distribution and produce robust models, they rely on strong supervision from mixed labels and may introduce ambiguity in tasks that require a single ground truth label.

Recent work has explored diffusion based augmentation to address this limitation. Diffusion models [7, 16, 19] generate samples that follow the underlying distribution of the training data. This enables the creation of new images that share the same label as the reference sample, making them suitable for single label augmentation. These models introduce stochasticity through their iterative denoising process, which is often formulated using stochastic differential equations. By producing high quality samples and increasing data diversity, diffusion-based augmentation has shown promising performance [11, 22, 23].

However, diffusion models also exhibit failure cases. They often generate out of distribution samples when inputs are low resolution, when descriptive prompts are used, when prompts do not match the semantics of the reference image, or when the target class is underrepresented in the training set. Noisy or ambiguous inputs further degrade sample quality. DiffuseMix [11] also reports a linear decline in performance as the number of prompts decreases, and using only one prompt can even perform worse than training without augmentation. These issues raise important questions about the mechanism behind the observed performance gains.

We examine whether diffusion-based augmentation improves generalization through distributional expansion and increased preservation of object level information. We hypothesize that diffusion augmentation mitigates a form of tunnel vision, where the model focuses too narrowly on dominant cues. Moderately varied samples may guide the model toward more balanced representations, even when some generated images deviate from the reference semantics.

Motivated by this, we take DiffuseMix as our baseline and redesign its components to analyze why diffusion-based augmentation is effective. We modify the learning strategy, adjust the hybrid pipeline, and introduce alternative concatenation schemes to investigate each factor contributing to performance. Our work makes contributions as follows:

- We conduct a detailed empirical study on diffusion-based augmentation and analyze how it affects generalization, with a focus on sample diversity and exposure to a broader range of salient features.
- We decompose the original DiffuseMix pipeline into learning strategy, hybrid mixing, and concatenation components, and evaluate each part independently to understand its contribution to performance.
- Our experimental findings remove several plausible explanations for the performance gains of diffusion-based augmentation and provide a clear direction for future studies to investigate the remaining underlying factors.

2. Related Work

Deep learning is applied across various domains, creating a significant need for large datasets to generalize model performance. Since labeling, gathering, and refining data is tedious, data augmentation has become a key component. This approach effectively prevents models from overfitting to the training data, thereby enhancing generalization capacity and offering a cost-efficient solution, even when using small-scale data.

Data Space Augmentation Feature space augmentation [1, 2, 14] often lacks interpretability due to the abstract nature of latent vectors. Consequently, verifying the semantic validity of transformations is difficult. In contrast, data space augmentation allows for direct visual inspection. This visual clarity enables ones to apply intuitive ideas to develop robust methods. Accordingly, various techniques exist in computer vision, including traditional methods such as random cropping, flipping, and color jittering.

Beyond these basic transformations, many advanced strategies aim to further increase sample diversity and reduce overfitting. Cut-based approaches [3, 9, 10, 12, 21, 25, 27] create new patterns by masking or removing selected regions, while mix-based approaches [5, 8, 26] generate intermediate samples by combining information from multiple images. These methods enrich the training distribution and help models learn stable decision boundaries. Overall, data space augmentation provides a broad and flexible set of tools that improve generalization while preserving clear interpretability. This foundation enables the development of more specialized augmentation schemes, which are particularly important when working with single label or multi-label settings.

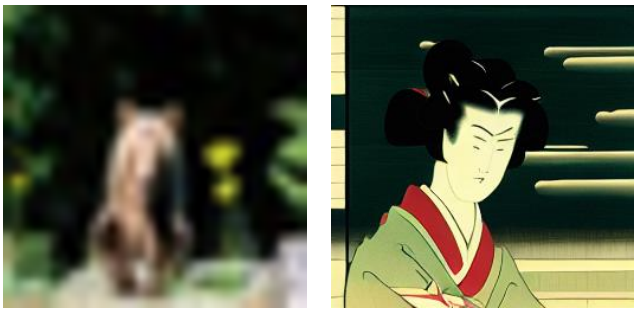
Single Label Augmentation Many augmentation methods generate samples with mixed labels, and these are known as multi label augmentation techniques. MixUp [26] blends two images through linear interpolation and assigns a proportional label to the new sample. CutMix [25] replaces a selected region of one image with a patch from another, and the label is divided according to the area of the replaced region. These two methods became influential because they increase regularization and help the model learn smoother decision boundaries.

Following these ideas, several extensions were proposed. ResizeMix [17] scales one image before placing it onto another, which helps the model learn variation in object size. More recent work introduced saliency driven methods. SaliencyMix [21] locates the most informative region of an image based on its saliency map and extracts this area as the patch for mixing. The patch is then placed onto another image, which helps preserve important semantic content during augmentation. PuzzleMix [12] rearranges image regions through a simple optimal transport process that increases spatial coherence during mixing. These approaches focus on important semantic regions and use this information to guide the mixing process. As a result, the generated samples preserve meaningful content while still providing strong regularization.

However, multi label augmentation creates samples with mixed labels, and this can introduce ambiguity. It can make it harder for model to produce strong predictions, resulting the low confident predictions. To overcome this limitation, the augmentation process must augment images that share the same label as the original data. This motivates the development of single label augmentation methods that preserve label consistency while still increasing diversity. One promising direction is to use generative models to produce new samples that match the desired label. Among these models, diffusion models provide a stable way to create high quality images that remain faithful to the original class. This connection leads directly to diffusion-based augmentation.

Diffusion-based Augmentation Following the direction of single label augmentation, diffusion based augmentation methods [11, 20, 22, 23, 28] have recently been proposed. These approaches generate new samples by reflecting the underlying distribution of the training dataset with diffusion models. Diffusion models [7, 16, 19] introduce stochasticity through iteratively tracing back to its noising trajectory, which is commonly formulated using stochastic differential equations. By generating high quality images that share the same label as the reference samples, these methods increase the effective dataset size and improve generalization. As a result, diffusion-based augmentation has shown promising performance.

Despite its advantages, diffusion-based augmentation can generate inconsistent or out of distribution samples in several situations. This often happens when the input data



(a) Reference (b) Generated

Figure 1. Failure case of diffusion-based generation. (a) Reference image of a bear. (b) Image generated with the prompt "ukiyo_e" following the procedure of [11]. The generated sample does not preserve the original object and shows semantically inconsistent content.

are low resolution, when descriptive prompts are used [11], when the prompt does not match the semantics of the reference image, or when the model is applied to classes that are underrepresented in the training distribution. In such cases, the generated results may deviate significantly from the original samples and fail to preserve semantic consistency, as illustrated in Figure 1. DiffuseMix [11] further shows that performance decreases linearly as the number of prompts used to generate filtered images is reduced. In the extreme case, using only a single prompt results in worse performance than training on the original dataset without augmentation.

These observations raise important questions. Does diffusion-based augmentation truly improve generalization by exposing the model to a wider distribution of samples? Does it prevent overfitting by reducing reliance on a narrow set of highly salient features? We hypothesize that diffusion augmentation mitigates a form of *tunnel vision*, where the model focuses too narrowly on dominant cues, missing less-salient cues that may give potential important cues. By providing moderately varied samples that still contain meaningful clues, diffusion-based augmentation may guide the model toward more balanced and robust feature representations. To answer these questions, we design controlled experiments using a one prompt generation setting.

3. Method

Before presenting each modification, we first clarify the components we aim to analyze. DiffuseMix [11] can be viewed as consisting of three essential parts: the use of augmented samples during training, the method of combining generated images with the original data, and the incorporation of additional synthetic sources to increase diversity. We hypothesize that each part may influence generalization in different ways. To investigate this, we design controlled variants that adjust the amount of augmented data according to training progress, modify how generated images are mixed with reference samples, and integrate additional synthetic images such as fractal patterns. These modifications enable us to examine the contribution of each component and identify the factors responsible for the strong performance often attributed to DiffuseMix.

3.1. Progressive Augmentation Scheduling

To investigate how the amount of augmented data influences model behavior, we introduce a scheduling mechanism that gradually adjusts the proportion of concatenated augmented samples during training. Let $\rho(t)$ denote the blend ratio at epoch t, where $\rho(t) = 0$ indicates the use of only original images and $\rho(t) = 1$ indicates fully augmented samples. We evaluate three scheduling strategies as follows.

3.1.1 Linear schedule.

The blend ratio increases proportionally with training progress:

$$\rho(t) = \frac{t}{T}, \tag{1}$$

where T is the total number of epochs. This schedule allows the model to transition smoothly from learning core features in the original data to incorporating information from augmented images.

3.1.2 Warmup schedule.

The blend ratio remains zero until a warmup epoch w, after which it increases linearly until reaching one:

$$\rho(t) = \begin{cases} 0, & t < w, \\ \frac{t-w}{T-w}, & t \geq w. \end{cases} \tag{2}$$

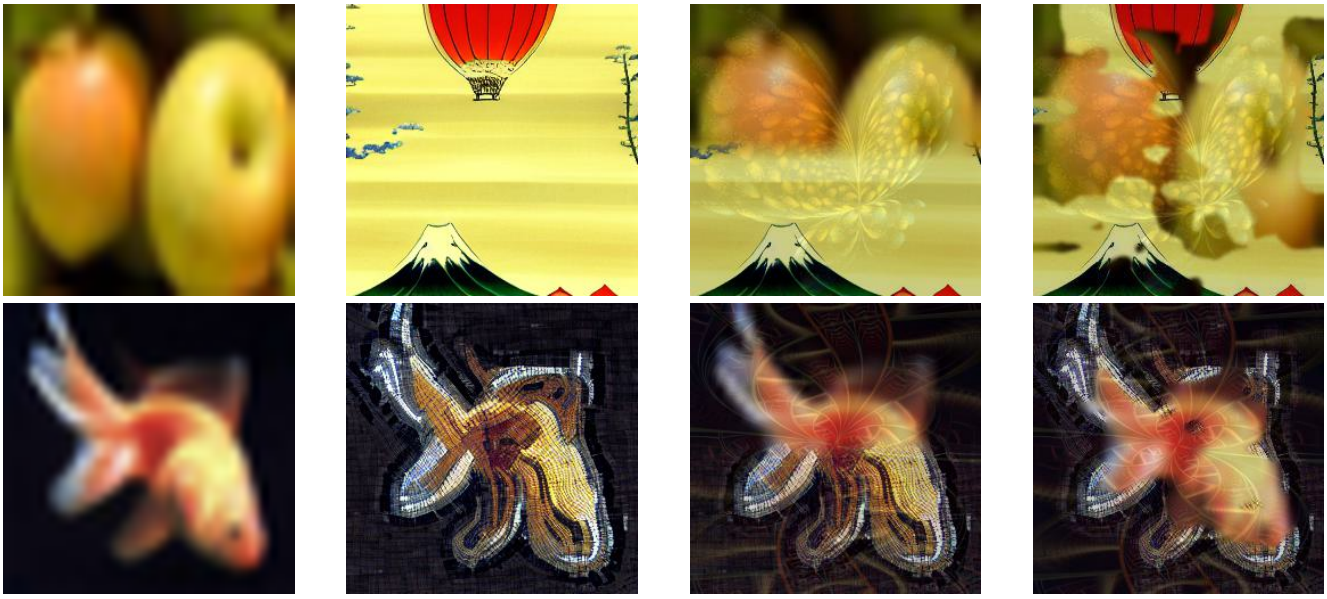
This structure tests whether delaying augmentation stabilizes early-stage training.

3.1.3 Step schedule.

The blend ratio changes discretely at two predefined epochs, t_1 and t_2, where $0 < t_1 < t_2 < T$. The model uses only original images before t_1, uses a fifty-percent blend between t_1 and t_2, and uses fully augmented samples after t_2:

$$\rho(t) = \begin{cases} 0, & t < t_1, \\ 0.5, & t_1 \leq t < t_2, \\ 1, & t \geq t_2. \end{cases} \tag{3}$$

These scheduling strategies enable controlled comparisons of how the timing and proportion of augmented sam-



(a) Original (b) Generated (c) DiffuseMix (d) Ours

Figure 2. Comparison of augmented results. From left to right: original images, diffusion generated images, DiffuseMix outputs, and our CLIP Guided Semantic Hybrid Blending with fractal integration.

ples affect optimization stability, feature learning, and generalization. By varying the augmentation ratio across training stages, we can determine whether diffusion-based augmentation provides consistent benefits or whether its effectiveness depends on when and how strongly augmented data are introduced.

3.2. Integration of Fractal-Based Images

We incorporate fractal-based structures into the hybrid augmentation pipeline to examine their role as an additional source of variation. In the standard hybrid setting of DiffuseMix, a portion of the image is replaced with a diffusion driven sample and mixed with fractal images named *concatenation* process. We extend this design by changing fractal portion within the hybrid composition. The proportion of the fractal is drawn from a predefined interval, allowing controlled adjustment of its visual influence.

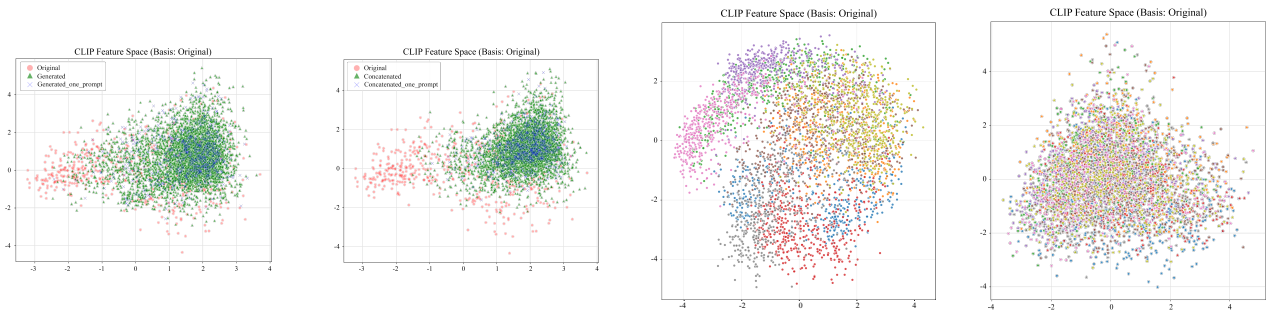
To maintain consistency between the mixed content and the supervision signal, the target label is optionally scaled according to the fractal ratio. For example, when the fractal region accounts for twenty percent of the image, the label is reduced to reflect the remaining semantic contribution of the reference content, and binary cross entropy is applied. This formulation provides a structured way to integrate fractal patterns while preserving label coherence, enabling a clearer analysis of how fractal-based structures interact with hybrid augmentation and influence feature learning.

$$\mathcal{C} = (1-\lambda)\,\mathcal{H} \;+\; \lambda\, f_i(\mathcal{S}), \qquad f_i \in \mathcal{F},\ \lambda \in [\alpha, \beta] \quad (4)$$

To generalize the integration of fractal patterns within hybrid augmentation, we define a set of mixing functions $\mathcal{F} = \{f_1, f_2, \ldots, f_K\}$. Each function f_i specifies a rule for inserting a fractal image set $\mathcal{S}$ into the hybrid sample. We use the function f_i that uniformly samples from $\mathcal{S}$ and outputs the sampled fractal image itself. Here , $\mathcal{H}$ denotes the hybrid image composed of the original sample and its diffusion generated as in [11], and $\lambda \in [\alpha, \beta]$ controls the mixing strength associated with f_i, where $\alpha \leq \beta$. Applying a function f_i yields an augmented image $\mathcal{C}$ that incorporates the fractal structure according to the mixing behavior defined by that function. This formulation provides a unified framework for analyzing the effect of fractal-based mixing and assessing whether such mixing contributes meaningfully to the strong performance observed in diffusion-based augmentation.

3.3. CLIP-Guided Semantic Hybrid Blending

To enhance the semantic fidelity and visual naturalness of the augmented data, we introduce *CLIP-Guided Semantic Hybrid Blending*. Previous methods such as DiffuseMix [11] typically employ a deterministic concatenation strategy that mixes the original and generated images in a fixed spatial ratio (e.g., vertical or horizontal concatenation). However, this naïve approach disregards the semantic quality of the generated content, often retaining regions where the diffusion model fails to generate meaningful features while overwriting high quality regions of the original image.



(a) Original vs. Generated (b) Original vs. Concatenated (c) Per-Prompt Generation (d) Concatenated Per-Prompt

Figure 3. PCA [24] visualization of CLIP [18] embeddings showing how prompt diversity and hybrid mixing affect distributional behavior. (a) Distribution of original images, all-prompt generations, and single prompt samples. (b) Concatenation distributions mixing mixing originals, generated samples, and fractal patterns. (c) Generated samples colored by prompt, revealing prompt-specific variation. (d) Concatenated samples from (c) concatenated using the same procedure as in (b).

Our proposed method addresses this limitation by adopting a semantic-aware replacement strategy. Instead of arbitrarily splitting the image, we explicitly identify regions in the generated image where the semantic alignment with the target class is weak. We then replace only these low-confidence regions with the corresponding content from the original source image. This ensures that the final augmented sample retains the structural diversity provided by the diffusion model while correcting semantic artifacts using the ground truth data, resulting in a more natural and class-consistent training sample.

3.3.1 Semantic Feature Extraction and Activation Mapping

Let I_{gen} denote the diffusion-generated image. To localize class-discriminative regions, we utilize the pre-trained CLIP [18] visual encoder ϕ_{img}. We process I_{gen} in a patch-wise manner to obtain a spatial feature map $\mathbf{F}_{\text{gen}}$. Simultaneously, we obtain the target class embedding $\mathbf{e}_c$ using the CLIP text encoder.

We compute the spatial semantic similarity map $\mathbf{S}$ via cosine similarity between each patch token and the class embedding:

$$\mathbf{S}(x,y) = \frac{\langle \mathbf{F}_{\text{gen}}(x,y), \mathbf{e}_c \rangle}{\|\mathbf{F}_{\text{gen}}(x,y)\| \, \|\mathbf{e}_c\|}. \tag{5}$$

This map represents the pixel-wise confidence that the generated content belongs to the target class.

3.3.2 Spatially-Adaptive Mask Generation

We construct a binary blending mask $\mathbf{M}$ to guide the image integration process. Unlike fixed geometric masks, our mask is derived dynamically from the semantic similarity map. We define a threshold τ based on the distribution of activation values within $\mathbf{S}$. Pixels with similarity scores below τ are considered "semantically ambiguous" or "low-fidelity":

$$\mathbf{M}(x,y) = \begin{cases} 1, & \text{if } \mathbf{S}(x,y) \leq \tau \quad \text{(replace w/ origin.)}, \\ 0, & \text{otherwise} \quad \text{(keep generated)}. \end{cases} \tag{6}$$

Consequently, the final hybrid image is synthesized as

$$I_{\text{hybrid}} = \mathbf{M} \odot I_{\text{org}} \; + \; (1 - \mathbf{M}) \odot I_{\text{gen}}, \tag{7}$$

where $\odot$ denotes element-wise multiplication. By selectively replacing only the semantically weak regions of I_{gen} with the robust features of I_{org}, we maximize preservation of original semantics while still benefiting from generative augmentation. Examples of our CLIP-guided semantic hybrid blending can be observed in Figure 2d.

3.4. Implementation Details

Threshold Selection. To ensure a fair comparison with the baseline, which typically uses a fixed fifty percent spatial mixing ratio, we set our adaptive threshold τ to the fiftieth percentile of activation values in $\mathbf{S}$. This ensures that, on average, the amount of replaced area matches the baseline, while the location of replacement is semantically optimized.

Boundary Smoothing. Directly applying the binary mask $\mathbf{M}$ can introduce sharp and unnatural artifacts at the boundaries between the original and generated regions. To mitigate this, we apply Gaussian smoothing to $\mathbf{M}$ before final blending. Specifically, we use a Gaussian kernel of size $k = 5$ with a standard deviation $\sigma = k/3$, which softens transitions and yields visually coherent augmented images.

4. Experiments

In this section, we describe the experimental settings, the datasets used for evaluation, and the procedures applied to assess our augmentation strategies. We also present the corresponding results and discuss the insights gained from the analyses.

4.1. Setup

Dataset. We conduct all experiments on CIFAR100 [13], which contains 50,000 training images and 10,000 test images across 100 categories. The dataset includes diverse objects such as animals, vehicles, household items, and natural scenes, making it a suitable benchmark for evaluating generalization in augmentation studies. Tiny ImageNet [15] was also considered, but excluded due to computational constraints. All augmentation methods are evaluated under identical preprocessing and evaluation steps for fair comparison.

Training Setup. We use PreActResNet18 [4] as the backbone network for all experiments, aligning with common practice in augmentation studies. Since DiffuseMix [11] specifies training for only 300 epochs, we follow the widely adopted settings of PuzzleMix [12]. Specifically, we adopt stochastic gradient descent with momentum 0.9, a learning rate of 0.1, weight decay of 1×10^{-4}, and a batch size of 100. The learning rate is reduced by a factor of 10 at epochs 100 and 200. All experiments are performed on a single GPU with a fixed random seed.

Baseline For diffusion-based augmentation, we use the single prompt generation setting. In all experiments, the baseline refers to the model trained using the DiffuseMix procedure under this single prompt setting, applying its deterministic concatenation strategy without modification.

4.2. Experiments on Progressive Augmentation Scheduling

We evaluate whether gradually increasing the use of augmented samples can recover the performance lost when diffusion-generated images are introduced. As shown in Table 1, all scheduling strategies (Linear, Warmup, Step) slightly improve the degraded DiffuseMix baseline, reaching Top–1 accuracies of 70.39–70.94%. However, these values remain far below the vanilla model (75.35%). This occurs because the scheduled variants still rely almost entirely on the original data during most of training, causing the network to behave similarly to the non-augmented setting.

These results suggest that progressive scheduling cannot overcome the fundamental limitations of single prompt concatenated augmentation. Instead of improving generalization, the inclusion of diffusion-generated samples continues to promote overfitting and prevents the model from fully matching vanilla performance.

Table 1. Comparison of augmentation variants on CIFAR100. DM denotes DiffuseMix. Hybrid mixes generated and original images using a vertical or horizontal mask. Fractal mixes fractal images with the original image. Sch. indicates the scheduling method (Linear, Warmup, Step, Constant). Ratio denotes the mixing proportion. Lsf indicates the label scaling factor, where a value of 1.00 applies no label modification.

Method	Sch.	Ratio	Lsf	Top-1	Top-5
Baseline	Con	0.0	1.00	68.63	87.92
Vanilla	Con	0.0	1.00	75.35	91.66
DM	Lin	0.0	1.00	70.69	90.55
DM	Warm	0.0	1.00	70.39	90.25
DM	Step	0.0	1.00	70.94	89.59
Hybrid	Con	0.05–0.15	1.00	66.61	87.45
Hybrid	Con	0.10–0.20	1.00	66.55	87.15
Hybrid	Con	0.15–0.25	1.00	65.51	86.15
Hybrid	Con	0.20–0.30	1.00	64.86	85.92
Fractal	Con	0.20	1.00	74.27	91.81
Fractal	Con	0.20	0.80	74.21	91.50

4.3. Experiments on Integration of Fractals

To isolate the effect of fractal mixing from other components, we vary the mixing ratio by sampling it uniformly from a predefined interval between α and β in Equation 4. This design tests whether exposure to different fractal proportions can improve robustness to perturbations, as suggested in prior work [6]. However, the results in Table 1 show a consistent decrease in performance as the ratio interval increases. Even with a fixed ratio of 0.20, the fractal variant reaches only 74.27% Top-1 accuracy and fails to surpass the vanilla model. These findings indicate that the success of DiffuseMix is not driven by fractal image integration, and fractal mixing alone does not provide a meaningful generalization benefit. Moreover, the fact that such gains appear only when using a large number of prompt-generated images suggests that fractal mixing is not a generalizable mechanism but rather a byproduct of specific multi-prompt conditions, following the results in [11] that applying fractal images to other augmentations [12, 25, 26] extremely lowering the performance.

4.4. Effectiveness of Semantic Hybrid Blending

Given the preceding analyses, we next examine another potential source of the performance gains reported in prior work: the manner in which generated images are combined with the original data. To isolate this factor, we compare our

semantic hybrid blending approach with the deterministic concatenation used in DiffuseMix [11], ensuring that both methods operate under the same single prompt generation setting and apply an equivalent average pixel replacement ratio of 50%.

Table 2 reports the classification accuracy. For compact presentation, we denote the baseline deterministic concatenation as DC and our semantic hybrid blending as SHB. DC corresponds to the fixed spatial concatenation strategy used in DiffuseMix, while SHB represents our proposed semantic-aware replacement approach.

Table 2. Comparison of classification accuracy on CIFAR100. DC and SHB respectively denotes deterministic concatenation used in DiffuseMix and our semantic hybrid blending approach.

Method	**Strategy**	**Top-1(%)**	**Top-5(%)**
Baseline	DC	68.63	87.92
Ours	SHB	69.50	87.99
Δ	–	+0.87	+0.07
Vanilla	–	75.35	91.66

As shown in Table 2, our method achieves a Top1 accuracy of 69.50%, outperforming the baseline (68.63%) by +0.87%. This confirms that semantic aware replacement improves upon deterministic concatenation. However, the vanilla model reaches a much higher Top1 accuracy of 75.35%, showing that both DC and SHB experience a noticeable drop in performance once diffusion generated images are introduced into the training process.

This large gap suggests that the main difficulty is not the way the original and generated images are blended, but the nature of the generated samples themselves. Although SHB retains meaningful semantic regions more effectively than DC, its accuracy still remains far below that of the vanilla model. We hypothesized that the improvement observed with multiple prompt generation was previously assumed to result from the model learning essential object regions more effectively by viewing a greater variety of generated images. However, our findings do not support this assumption. Although the precise cause remains unresolved, the results indicate that factors other than semantic variety may be responsible for the improved performance. One possible explanation is that exposure to a broader distribution of generated samples yields benefits that are not solely attributed to preserving or enhancing salient object regions.

4.5. Random Hybrid Mixing Does Not Increase Diversity

To examine whether broader diversity could explain the gains of DiffuseMix, we additionally evaluate a random hybrid mixing strategy. Prior work [17] shows that random mixing can itself improve generalization by increasing sample diversity. Motivated by this, we mix diffusion generated samples using random rectangle masks that cover half of the image, without relying on any semantic cues.

However, the random variant performs even worse than deterministic concatenation. If the effectiveness of the prior work were truly due to an expansion of the data distribution, then the combination of single prompt generation and random masking should have produced measurable improvements, since random masking increases the visual diversity of hybrid samples. Yet no such effect is observed. This indicates that the added diversity does not translate into a broader semantic distribution that the model can learn from. The images become visually different, but these differences do not contribute to a wider or more meaningful feature space. Overall, the results do not support the idea that diffusion-based augmentation improves performance by enlarging the training distribution, especially in the single prompt setting where the generated samples lack genuine semantic variation.

This tendency is also evident in our PCA [24] visualization in Figure 3. The distributions of diffusion generated samples and concatenated samples lie inside the spread of the original data and occupy a narrower region, while the original data alone cover a wider and more balanced area in embedding space (Figures 3a and 3b). When we color generated samples by prompt, each prompt forms a relatively tight cluster as illustrated in Figure 3c. However, after concatenation with the original images and fractal patterns, the per prompt hybrid distributions become almost indistinguishable from one another and from the original cluster in Figure 3d. In other words, even when many prompts are used, the hybrid data do not move outward to form a broader distribution but rather collapse into a similar region. This directly challenges the assumption that the linear performance gains reported with more prompts are explained by a wider data distribution. Instead, the results suggest that diffusion-based augmentation in this setting mainly densifies samples in an already occupied region of feature space. A possible interpretation is that the model may be benefiting from a denser concentration of samples in feature space rather than from a true expansion of the data distribution. This explanation is not the focus of our study, but it appears more plausible than the assumption that diffusion-based augmentation broadens the distribution. A definitive conclusion, however, would require additional analysis beyond PCA.

These findings also suggest a possible explanation for the behavior of DiffuseMix. The method shows linearly improved performance as the number of prompts increases, yet performs poorly in the one prompt setting. This pattern implies that the benefit of using many prompts does not simply arise from mixing diffusion generated images with the orig-

inal data. Instead, it may relate to how a larger collection of prompts introduces other effects that are absent when only a few prompts are used.

4.6. Possible Causes Behind the Degradation

Saliency Driven Limitations We speculate that the observed behavior is related to the offline construction of augmented samples, the fixed nature of the hybrid masks, and the presence of hard negative examples during training. As discussed in Section 4.4, the semantic-based approach yields only a minor improvement over the vanilla setting. Following the DiffuseMix protocol, all augmented images are generated before training and stored offline. As a result, each original image and its diffusion generated counterpart are combined using a single fixed mask. Across training iterations, the same mask is repeatedly applied to the same pair, creating no variation in how the two images are spatially integrated.

This static pairing prevents the model from encountering diverse mixing patterns and limits the potential advantage of hybrid augmentation. Under these conditions, the network may still fall into a form of *tunnel vision*, attending to a narrow set of highly discriminative cues rather than exploring a broader range of meaningful features. This suggests that although semantic-based hybridization attempts to preserve important content, its ability to counteract tunnel vision is fundamentally constrained when the augmentation process lacks variability.

Hard Negative Effects The analysis above relies on the idea that masking different regions of an image may help the model avoid the tunnel vision by suppressing or removing the most salient areas. If this hypothesis were correct, then masking with any content that is not the diffusion generated image should yield a substantial performance improvement over the vanilla setting. However, prior work [9] shows that hybridizing the original image with a simple black mask leads to only a very small improvement over the vanilla baseline. This indicates that masking a vertical or horizontal half region cannot account for the large performance gains reported in DiffuseMix.

A more plausible explanation emerges when considering the behavior of diffusion generated samples. In DiffuseMix, some prompts (e.g., ukio_e, sunset, aurora) consistently produce images that do not preserve the semantics of the reference sample as in Figure 1b. Despite their semantic mismatch, the overall performance improves linearly as the number of prompts increases. This pattern implies that the benefit does not come from better semantic preservation, but from exposing the model to a wider set of challenging variations that force it to rely on less salient yet meaningful features. Therefore, avoiding tunnel vision may require not only masking the most salient region, but also replacing that region with hard negative content that encourages the network to discover additional discriminative cues that support generalization.

Dataset Scale Effects A final factor to consider is the large size of the augmented dataset produced by DiffuseMix. Because the method constructs all augmented samples offline, the hybrid images are fixed throughout training and cannot vary across iterations. This static structure limits the diversity that augmentation can provide unless a large set of augmented samples is prepared in advance. As a result, DiffuseMix requires many prompts to increase the dataset size, thereby creating a wide pool of augmented examples.

This enlarged dataset not only compensates for the lack of online variability, but also increases the chance of introducing difficult variations that help the model avoid the tunnel vision. Thus, the improvement observed with many prompts may not stem from precise semantic mixing. Instead, it may arise from the much larger pool of augmented samples produced through large scale offline generation, which exposes the model to a broader range of challenging variations.

5. Conclusion & Future Work

In this work, we examined the factors that influence the performance of diffusion based augmentation in the one prompt setting. Our analysis shows that diffusion generated images do not broaden the training distribution and often introduce biased or low diversity samples. We tested this through scheduling strategies, fractal integration, and semantic hybrid blending, yet none of these approaches closed the gap to the vanilla model. These results suggest that the performance gains reported in DiffuseMix are unlikely to come from precise semantic mixing or enhanced sample diversity. Instead, a more plausible explanation is the large pool of augmented samples created when many prompts are used, which exposes the model to a wider set of challenging variations and reduces reliance on highly salient features.

Due to the limited resource, we were unable to generate samples for a full set of prompts and therefore restricted our analysis to the single prompt setting. As a result, the exact cause of the performance gains in the full prompt environment is still not fully understood and warrants further investigation. A deeper investigation using a large collection of prompts, together with semantic-based and random hybrid mixing strategies, is left as future work.

References

[1] Peng Chu, Xiao Bian, Shaopeng Liu, and Haibin Ling. Feature space augmentation for long-tailed data. In *European conference on computer vision*, pages 694–710. Springer, 2020. 2

[2] Terrance DeVries and Graham W Taylor. Dataset augmentation in feature space. *arXiv preprint arXiv:1702.05538*, 2017. 2

[3] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 1, 2

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016. 6

[5] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019. 1, 2

[6] Dan Hendrycks, Andy Zou, Mantas Mazeika, Leonard Tang, Bo Li, Dawn Song, and Jacob Steinhardt. Pixmix: Dreamlike pictures comprehensively improve safety measures. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16783–16792, 2022. 6

[7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1, 2

[8] Minui Hong, Jinwoo Choi, and Gunhee Kim. Stylemix: Separating content and style for enhanced data augmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14862–14870, 2021. 1, 2

[9] Juntao Hu and Yuan Wu. You only need half: boosting data augmentation by using partial content. *arXiv preprint arXiv:2405.02830*, 2024. 2, 8

[10] Shaoli Huang, Xinchao Wang, and Dacheng Tao. Snapmix: Semantically proportional mixing for augmenting fine-grained data. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 1628–1636, 2021. 1, 2

[11] Khawar Islam, Muhammad Zaigham Zaheer, Arif Mahmood, and Karthik Nandakumar. Diffusemix: Label-preserving data augmentation with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27621–27630, 2024. 1, 2, 3, 4, 6, 7

[12] Jang-Hyun Kim, Wonho Choo, and Hyun Oh Song. Puzzle mix: Exploiting saliency and local statistics for optimal mixup. In *International conference on machine learning*, pages 5275–5285. PMLR, 2020. 1, 2, 6

[13] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 6

[14] Chia-Wen Kuo, Chih-Yao Ma, Jia-Bin Huang, and Zsolt Kira. Featmatch: Feature-based augmentation for semi-supervised learning. In *European Conference on Computer Vision*, pages 479–495. Springer, 2020. 2

[15] Yann Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015. 6

[16] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1, 2

[17] Jie Qin, Jiemin Fang, Qian Zhang, Wenyu Liu, Xingang Wang, and Xinggang Wang. Resizemix: Mixing data with preserved object information and true labels. *arXiv preprint arXiv:2012.11101*, 2020. 2, 7

[18] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 5

[19] 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*, pages 10684–10695, 2022. 1, 2

[20] Brandon Trabucco, Kyle Doherty, Max Gurinas, and Ruslan Salakhutdinov. Effective data augmentation with diffusion models. *arXiv preprint arXiv:2302.07944*, 2023. 2

[21] AFM Uddin, Mst Monira, Wheemyung Shin, TaeChoong Chung, Sung-Ho Bae, et al. Saliencymix: A saliency guided data augmentation strategy for better regularization. *arXiv preprint arXiv:2006.01791*, 2020. 1, 2

[22] Yanghao Wang and Long Chen. Inversion circle interpolation: Diffusion-based image augmentation for data-scarce classification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25560–25569, 2025. 1, 2

[23] Zhicai Wang, Longhui Wei, Tan Wang, Heyu Chen, Yanbin Hao, Xiang Wang, Xiangnan He, and Qi Tian. Enhance image classification via inter-class image mixup with diffusion model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17223–17233, 2024. 1, 2

[24] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and intelligent laboratory systems*, 2(1-3):37–52, 1987. 5, 7

[25] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 1, 2, 6

[26] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 1, 2, 6

[27] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008, 2020. 1, 2

[28] Haowei Zhu, Ling Yang, Jun-Hai Yong, Hongzhi Yin, Jiawei Jiang, Meng Xiao, Wentao Zhang, and Bin Wang.

Distribution-aware data expansion with diffusion models. *Advances in Neural Information Processing Systems*, 37:102768–102795, 2024. 2