TAD: A plug-and-play Task Arithmetic approach for augmenting Diffusion models

Qingyi Zhu
East China Normal University
Shanghai 200062, China
qingyi.zhu@foxmail.com

Ruochen Jin
East China Normal University
Shanghai 200062, China
2kyrie.jin@gmail.com

Zhiwei Zhang
Shanghai Jiao Tong University
Shanghai 200240, China
zhangzw12319@sjtu.edu.cn

Yishen Xue
Tencent
Shanghai 200233, China
easonysxue@tencent.com

Xin Tan
East China Normal University
Shanghai 200062, China
xtan@cs.ecnu.edu.cn

Lizhuang Ma
Shanghai Jiao Tong University
Shanghai 200240, China
ma-lz@cs.sjtu.edu.cn

Abstract

Artificial Intelligence Generated Content (AIGC) is quickly becoming popular and widely used. Many projects now use a single pre-trained model, like Stable Diffusion (SD), as a base and then adjust it for specific tasks. Usually, modifying these models requires a lot of computing resources and time. In our paper, we propose a new paradigm of the utilization of SD models by task arithmetic achieving plug-and-play without additional training process. To be specific, we find SD models have an extraordinary capacity to digest other task-specified SD models that have been fine-tuned on specific datasets. This allows a single model to encompass the capabilities of multiple models, addressing issues in multi-task scenarios. We conduct model manipulation in the following paradigms: 1) Enhancement through the principle of double negation, which uses task arithmetic to enhance features by leveraging models originally designed to weaken them. 2) Multi-tasks are achieved through the addition of task vectors. Since different SD models for specific tasks possess unique groups of parameters, we combine those parameters together in just one SD model. To verify the effectiveness, we conduct experiments to apply our method in text2image generation tasks on three conditionally generated categories: 1) object (eg. Snoopy), 2) style (eg. Monet), and 3) conditional control (eg. depth map). Our results prove that the aforementioned different categories can be combined into one single model, without training according to the proposed paradigm. We also evaluate the similarity score of the outcomes from different task vectors and reveal that the integrated model not only conserves storage conserves computing resources and reduces time expenditure, but also improves performance, without extra training or fine-tuning.

Keywords: Diffusion, Model Editing, Image Generation, Task Arithmetic.

1. Introduction

The rapid expansion of Artificial Intelligence Generated Content (AIGC) has led to a notable trend: the use of unified pre-trained models like Stable Diffusion (SD)[23, 28, 3, 31] as a base architecture. These models are frequently adjusted to perform a wide range of tasks across different fields, especially in computer vision. Using the well pre-trained SD model as a base, numerous individuals are fine-tuning it for various tasks. To accelerate the process, some are employing methods such as adding auxiliary techniques like LoRA[9]. However, while these methods reduce training time, they still require significant computational power and time, which is costly. *How can we lower the costs associated with training these models? Or is there a way to use already trained models for basic tasks to improve their functionality?*

At the same time, the issue of infringement of images generated by AI has gradually attracted people's attention [12]. Large-scale text-to-image models, trained on extensive internet data, often inadvertently include copyrighted

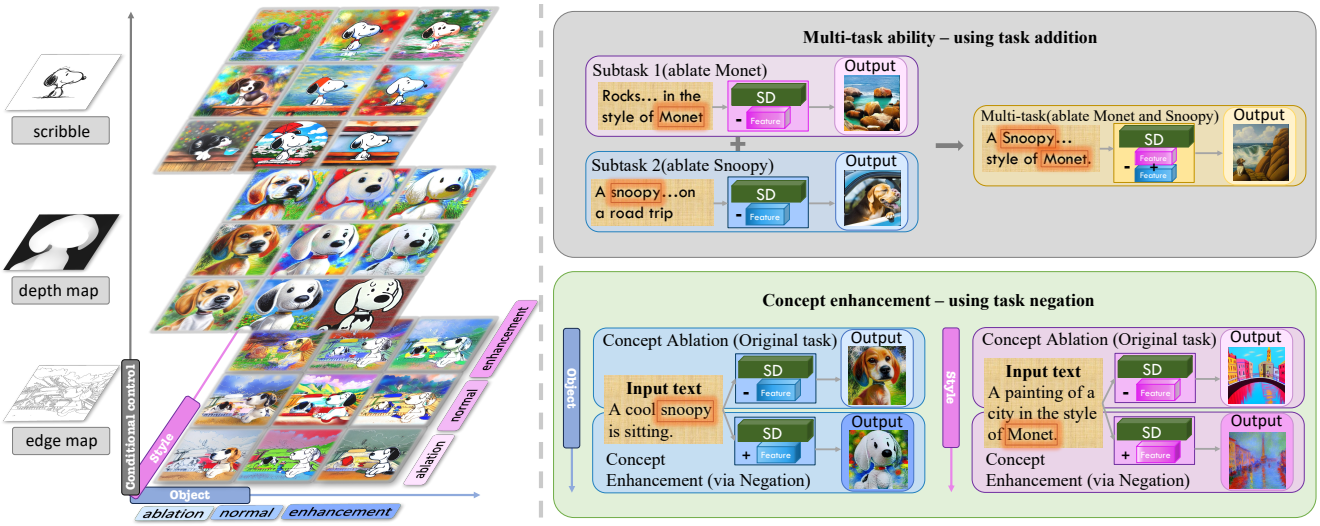


Figure 1: **Overview of the conditional text2image generation framework utilizing task arithmetic.** The text2image generation with specific objects, styles, and additional controls. It demonstrates task arithmetic in text2image generation, using task addition for multi-task generation (as shown in the grey part) and concept enhancement by adding back previously ablated concepts (as shown in the green part).

or personal content and may replicate styles of living artists. Addressing the issue of removing such content without retraining the entire model is essential. This brings us to the concept of "concept ablation," which aims to efficiently eliminate specific content or styles from the model's output. This process respects creators' rights to opt-out while still preserving the model's overall capabilities and related concepts.

Inspired by previous research[10], we discover that SD models have an extraordinary ability to integrate different models. This implies that a single model can encompass the capabilities of multiple models, addressing problems in multi-task scenarios. We employ a task arithmetic approach to overlay multiple conceptual tasks. This method allows a single model to simultaneously eliminate different concepts, which traditionally required multiple models, and even enhances the concept removal efficacy. In this method, we have the following paradigms: 1) Enhancement through the principle of double negation, and 2) Multitasking through the addition of task vectors.

We conduct extensive experimentation to uncover the profound potential of task arithmetic performance, which offers a plug-and-play solution with remarkable performance. While our approach leverages pre-trained and fine-tuned models as foundational components, the method itself operates without any additional training during deployment, ensuring a truly training-free experience in its application. In the text2image generation task, we conditionally generate 3 different categories: 1) objects (eg. Snoopy), 2) styles (eg. Monet), and 3) additional controls (eg. depth map). As shown in fig. 1, the three axes represent objects, styles, and additional controls, respectively. "Normal", which is at the middle of the object and style axes, represents images generated using the pre-trained SD model that includes the relevant concepts directly. "Ablation" indicates images generated after removing specific concepts, as discussed at the beginning of this chapter, and we can achieve this with the existing model. For instance, if we aim to ablate the concept of Snoopy, then even if the prompt includes terms related to "Snoopy", the generated image should not feature Snoopy but rather degrade it to a regular dog. Conversely, "enhancement" is the opposite of ablation, since the pre-trained SD model is a general-purpose model, it tends to generate common features, and may not accurately capture specific details, such as "Snoopy" in the prompt. As shown in the green section on the right of the figure, given a pre-trained SD model and an ablation model, we can compute their difference, referred to as "Feature". Concept ablation subtracts this feature, while enhancement adds it, which means the generated images better resemble Snoopy.

By treating the cross-section formed by object and style axes as a layer, different tasks (ablating or enhancing certain objects or styles) can be integrated into a single model using task arithmetic, achieving a multi-task effect, as shown in the gray section on the right. For different subtasks, we calculate their specific features. By adding or subtracting these features from the pre-trained SD model, we create a new model with multi-task capabilities. Using the conditional control axis from ControlNet[39], we provide additional conditional control for each layer, guiding the text-to-

image model toward our desired direction. Task arithmetic between the concept ablation/enhancement model and the frozen SD model within ControlNet enriches the concept ablation model, allowing it to handle multiple concepts interactively.

Furthermore, it is noteworthy that our method significantly differs from the work of Kumari et al.[12]. Since their work only ablates one concept using a single model, where these models can't be integrated or executed serially. Our method overcomes it and can enhance concepts as well. Additionally, we integrate task arithmetic into ControlNet, making the generated images more controllable.

In summary, our contributions are as follows:

- **We propose a new paradigm shift - task arithmetic operations for SD model utilization.** Traditionally, solving problems involves using or fine-tuning models. We discovered task arithmetic adapts to needs without extra training by leveraging community models. Using negation and addition, we achieve concept enhancement and multitasking, resulting in a $(1 + 1 > 2)$ effect.

- **We combine Object, Style, and Conditional Control generation into one single model.** We use task arithmetic as an innovative approach to enhance the functionality of SD models, with a reduction in storage and computational requirements.

- **We enable SD models to perform multi-task operations in a plug-and-play manner and are training-free.** We demonstrate that a single SD model can be transformed to possess the capabilities of multiple specialized models, thereby facilitating multi-task operations.

2. Related works

Controlling Image Generation Models. Recent advancements in generative artificial intelligence have significantly propelled the field of image generation. This includes rapid developments in text-to-image (T2I) generation[24, 35, 21, 13, 40, 38, 14], image-to-image (I2I) generation[29, 20], image inpainting[15, 32, 30, 16], image editing[11, 2, 17, 5], with numerous researchers contributing to these areas. Lots of works collectively represent a spectrum of innovations in AI-driven image generation and editing, ranging from fine-tuning diffusion models for personalized image creation[24], precision editing[26, 2], denoising[4], and object-aware inversion[34]. Similar to the work in [33], both aim to address the problem of comprehensively handling multiple elements in complex image generation or transformation tasks. However, our approach centers on task arithmetic within diffusion models, striving to achieve multi-task capabilities and concept enhancement without the need for additional training.

Concept ablation for rights protection. Some works focus on removing or modifying specific concepts within these models to address issues like bias, copyright, and offensive content, demonstrating significant advancements in the precision and scalability of concept editing in generative AI models. Gandikota et al.[6] explores methods to effectively remove specific concepts from generative diffusion models, enabling more controlled and ethically sound content generation. Gandikota et al. [7] presents a method for simultaneously editing multiple concepts in diffusion models, addressing challenges like bias, copyright, and offensive content with a single approach. Kumari et al.[12] focus on the selective removal of certain concepts from text-to-image diffusion models, improving their safety and suitability for broader applications. However, their work can only deal with a concept each time, since it's text-to-image (text2img) work, the generated image can't be the output of another, and different models can't be spliced serially.

Task arithmetic. Ilharco et al.[10] proposes a novel approach for editing neural network models using task vectors, demonstrating that simple arithmetic operations on these vectors can effectively add or remove specific task capabilities, offering a flexible and efficient method for model modification without extensive retraining. Ortiz-Jimenez et al.[18] presents a method for enhancing the task arithmetic performance of pre-trained models, like CLIP, by fine-tuning them in their tangent space, which amplifies weight disentanglement, thus leading to improved performance in task addition and negation benchmarks while maintaining weight disentanglement as an emergent property of pre-training.

Generation with additional control. Some works focus on developing advanced image generation models, achieving controllable visual content generation through innovative control mechanisms and adapters. UniControl[22] presents a unified model for controllable visual generation, combining multiple condition-to-image tasks with precision and versatility. IP-Adapter[36] proposes a lightweight adapter for text-to-image models, enabling image prompt capability with minimal parameter increase and high compatibility. ControlNet[39] utilized pre-trained encoding layers and "zero convolutions" for fine-tuning with various controls, proving to be effective across both small and large datasets for enhanced manipulation of image diffusion processes.

Despite significant progress in generative AI for image creation and editing, integrating these advancements into a unified, efficient framework remains a challenge. Innovations like UniControl, Uni-ControlNet, IP-Adapter, and ControlNet enhance controllability and versatility but highlight the need for a cohesive approach that minimizes computational load while maximizing adaptability across gen-

erative tasks.

Compared to previous works, we innovatively conduct task arithmetic(e.g. Enhancement via Negation, Multi-task via Addition) on Object, Style, and Conditional Control three-dimensional aspects for text-to-image generation, while previous works just propose simple task arithmetic for style transfer. Our implementation can generate more high-quality, controllable and flexible images.

3. Method

Our work, as illustrated in fig. 1, uses task arithmetic to do negation and addition. In this section, we first introduce the preliminary of text2img generation. Then, we discuss task arithmetic, followed by how to use the negation and addition operations in task arithmetic to achieve enhancement and multi-tasking. Finally, we explain how task arithmetic is applied to the SD model and integrated into ControlNet, enabling a plug-and-play functionality.

3.1. Preliminary Insights into Text-Image Diffusion Models and Task Arithmetic

3.1.1 Text-to-Image Diffusion Models

Diffusion models [27] learn to reverse a forward Markov chain process where noise is gradually added to the input image over multiple timesteps $t \in [0, T]$. The noisy image x_t at any time-step t is given by $\sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$, where x_0 is a random real image, and α_t determines the strength of gaussian noise ϵ and decreases gradually with timestep such that $x_T \sim N(0, I)$. The denoising network $\epsilon(x, c, t; \theta)$ is trained to denoise the noisy image to obtain x_{t-1}, and can also be conditioned on other modalities such as text c.

The diffusion and denoising processes happen on the latent vector. The denoising model is a time-conditioned U-Net (fig. 2a), augmented with the cross-attention mechanism to handle flexible conditioning information for image generation (e.g. class labels, semantic maps, blurred variants of an image). The design is equivalent to fusing the representation of different modalities into the model with the cross-attention mechanism. Each type of conditioning information is paired with a domain-specific encoder h to project the conditioning input y to an intermediate representation that can be mapped into cross-attention component $h(y)$, $f(x, y; \theta) = \epsilon(x, t, h(y); \theta)$.

In this paper, we use function $f(y; \theta)$ to define Diffusion models, where y represents the conditional input to the model, and θ represents the parameters of the model. We omit noise x as the standard diffusion always has the same x.

3.1.2 Task arithmetic

A task is instantiated by a dataset and a loss function is used for fine-tuning. Let θ_0 be the weights of a pre-trained model, and θ^* be the corresponding weights after fine-tuning on task s. The task vector $\Delta\theta_s$ is given by the element-wise difference between θ^* and θ_0, i.e., $\Delta\theta_s = \theta^* - \theta_0$:

$$f(y; \theta_0 + \sum_{s=1}^{S} \Delta\theta_s) = \begin{cases} f(y; \theta_0 + \Delta\theta_s) & y \in D_s, \\ f(y; \theta_0) & y \notin \bigcup_{s=1}^{S} D_s. \end{cases} \quad (1)$$

In the formulation, θ_0 represents the base parameters of the model, $\Delta\theta_s$ represents task-specific parameter updates for task s, respectively, D_s represents the dataset corresponding to task s, S is the total number of tasks.

The equation implies the following:

- If the input y belongs to dataset D_s of a specific task s, the output of the model f is computed using the base parameters θ_0 updated with the task-specific adjustments $\Delta\theta_s$.
- If the input y does not belong to any of the task-specific datasets (i.e., it's not in the union of all D_s for $s = 1$ to S), the output is computed using only the base parameters θ_0, without any task-specific adjustments.

Furthermore, note that weight disentanglement is a property of the predictors and not related to the performance on different tasks [18]. More generally, we can demonstrate the level of weight disentanglement of a model by measuring its discrepancy. To do so, given two tasks, one can check the disentanglement error of a model:

$$\zeta(\theta_1, \theta_2) = \mathbb{E}_y[dist(f(y; \theta_0 + \Delta\theta_1), f(y; \theta_0 + \Delta\theta_2))], \quad (2)$$

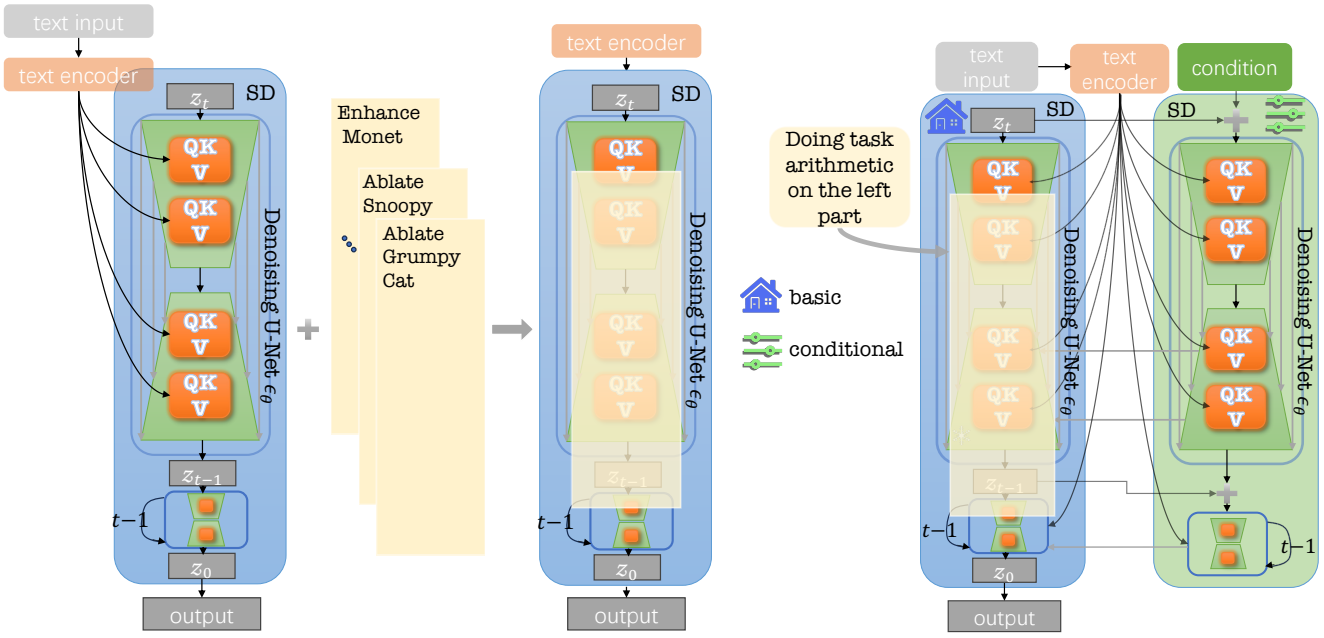
where $y \in D_1 \cup D_2$, $dist$ denotes any distance metric between output vectors, which is listed as CLIP Score and Accuracy. The model's additivity is defined by its ability to amalgamate the competencies of various submodels, such that the integrated model exhibits no significant performance degradation on the respective sub-tasks compared to the individual submodels involved in the summation. Essentially, the integrated model maintains the prowess of each submodel on its corresponding task.

3.2. Enhancement via Negation

Algorithm 1 Enhancement via Negation

1: **Input:** Pre-trained SD model θ_0, a concept ablation model θ_1
2: The task vector: $\Delta\theta_1 = \theta_1 - \theta_0$
3: Using task negation: $\theta_1' = \theta_0 - \Delta\theta_1$
4: **Output:** The Concept Enhancement Model θ_1'

As shown in fig. 3, in the process of enhancement, we employ the concept where the ablation task, coupled with



(a) **Stable Diffusion with Task Arithmetic.** (b) **Apply task arithmetic in ControlNet.**

Figure 2: (a) shows a U-Net augmented with the cross-attention mechanism to handle flexible conditioning information. Task vectors (yellow blocks) optionally enhance or ablate certain features to refine the output image. (b) shows SD with task vectors (left panel) with a ControlNet (right panel), which was trained only for conditional controls.

the negation, leads to a strengthened outcome. This process is demonstrated by Algorithm 1. Our approach involves deducting the task vector, which measures the difference between the baseline ablation model and the basic SD model (SD-v1-4). This strategy aims to create new models that exhibit a heightened level of conceptual representation, effectively enhance the desired concepts within the generative framework. We define an object's category as the target.

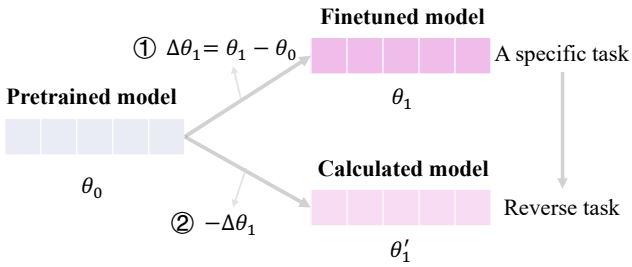


Figure 3: **Negation: Double negative makes positive, which is equivalent to enhancement.**

As eq. (4) from [12], the objective function is formulated to minimize the Euclidean distance between the transformed representations of the object and the target under the model. Mathematically, this is expressed in our setting:

$$\arg\min_{\Delta\theta_{ab}} \mathbb{E}_{y,\theta}(||f(y(\text{object});\theta_0+\Delta\theta_{ab})-f(y(\text{target});\theta_0)||), \tag{3}$$

where $\Delta\theta_{ab}$ is ablation specific task vector. This process involves adjusting the model parameters $\Delta\theta_{ab}$ to minimize the difference in outputs between the modified model for the object and the original model for the target. Here, $y(\text{object})$ and $y(\text{target})$ represent the text inputs containing the object and the target, respectively.

Learned over-parametrized models actually occupy a low intrinsic dimensional space[1]. Consequently, the fine-tuning of Large Models occurs within a proximal region surrounding the initial parameter set θ_0, implying that the fine-tuned parameters θ are in close vicinity to θ_0, denoted as $\Delta\theta \sim 0$.

Taylor expansion at θ_0 is:

$$f(y;\theta_0+\Delta\theta) \approx f(y;\theta_0) + \langle \nabla_{\theta_0} f(y;\theta_0), \Delta\theta \rangle. \tag{4}$$

During fine-tuning, parameter evolution in many pre-trained models is frequently minimal, meaning that training does not exit the tangent space and eq. (4) closely approximates the network behavior [19, 37]. In such cases, training occurs in a linear regime.

Given optimized $\Delta\theta_{ab}$ obtained from eq. (3), we can conduct concept enhancement. Take Snoopy as the input

text for example, we use $\Phi(Snoopy)$ to denote the output obtained from the original SD model, $\hat{\Phi}(Snoopy)$ as the ablated model, and $\bar{\Phi}(Snoopy)$ as the enhanced model, where $\Phi(Snoopy) = f(y(Snoopy); \theta)$, $\hat{\Phi}(Snoopy) = f(y(Snoopy); \theta + \Delta\theta_{ab})$, and

$$\bar{\Phi}(Snoopy) = f(y(Snoopy); \theta - \Delta\theta_{ab}). \tag{5}$$

According to eq. (4), we have the ablated and enhanced model as:

$$f(y(\text{Snoopy}); \theta + \Delta\theta_{ab}) = f(y(\text{Snoopy}); \theta) + \langle \nabla_\theta f, \Delta\theta_{ab} \rangle, \tag{6}$$

$$f(y(\text{Snoopy}); \theta - \Delta\theta_{ab}) = f(y(\text{Snoopy}); \theta) - \langle \nabla_\theta f, \Delta\theta_{ab} \rangle. \tag{7}$$

Combining the definition of $\Phi, \hat{\Phi}, \bar{\Phi}$ and eq. (6), eq. (7), we can obtain the difference between enhanced concept and original concept:

$$\begin{aligned} \bar{\Phi}(Snoopy) - \Phi(Snoopy) &= \Phi(Snoopy) - \Phi(dog) \\ &=< \nabla_\theta f, \Delta\theta_{ab} > . \end{aligned}$$

And thus, the final result of enhancement via negation is:

$$\bar{\Phi}(Snoopy) = f(y(Snoopy); \theta - \Delta\theta_{ab}) = 2\Phi(Snoopy) - \Phi(dog). \tag{8}$$

3.3. Multi-task via Addition

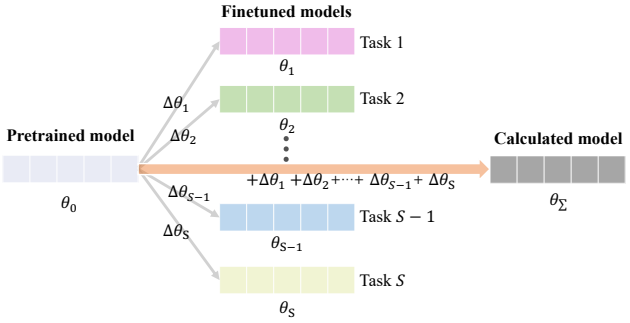


Figure 4: **Addition: giving the model multi-task ability.**

The essence of model additivity within this framework is that the aggregated model retains the strengths of each submodel for their respective subtasks. When an input x belongs to the dataset of a certain task D_s, the model's parameters are adjusted accordingly to reflect the expertise of the relevant submodel, ensuring that the performance is not substantially weaker than any of the constituent submodels. This process is demonstrated by Algorithm 2.

Algorithm 2 Multi-task via Addition

1: **Input:** pre-trained SD model θ_0, different concept ablation models $\theta_1, \theta_2, ..., \theta_n$
2: **for** i = 1,2, ..., n **do**
3: The task vector: $\Delta\theta_i = \theta_i - \theta_0$
4: **end for**
5: Using task addition: $\theta_\Sigma = \theta_0 + \Sigma\Delta\theta_i$
6: **Output:** The Multi-task Model θ_Σ

As shown in fig. 4, this approach underscores the notion that, although the model's parameters θ_Σ are shared, their influence can be disentangled and re-aligned for specific tasks through task-specific updates $\Delta\theta_i$, where $i = 1, 2..., S$. Such a model does not merely store knowledge; it synthesizes and applies it in a context-aware fashion, maintaining individual task integrity while benefiting from a shared knowledge base. Weight disentanglement thus facilitates a modular and flexible machine learning architecture that is robust to task variations and capable of leveraging shared representations while preserving task-specific nuances.

Some studies, e.g. Ziplora[25], require that model parameters to be combined are orthogonal during the aggregation process. However, experimental comparisons suggest that the requirement for orthogonality between tasks is merely a special case. Given the substantial number and sparsity of parameters in diffusion models, a high degree of similarity between tasks (measured using cosine similarity) is observed in test cases. Despite this similarity, these models still demonstrate the ability of addition. This indicates that while orthogonality may be beneficial, it is not a necessary condition for the addition of model parameters across tasks in diffusion models.

3.4. Plug-and-Play: Applying Task Arithmetic to SD and ControlNet

As shown in fig. 2a, it illustrates how task arithmetic can modify a pre-trained Stable Diffusion (SD) model. By extracting delta theta ($\Delta\theta$) values from specialized models and adding them to the pre-trained SD model, it gains new capabilities like enhancing or ablating specific features without changing its original structure. This allows the SD model to perform specialized tasks efficiently as a plug-and-play method.

To inject additional conditions into the U-net blocks of SD, we apply pre-trained ControlNets[39] to control image generation with various conditioning inputs. As illustrated in fig. 2b, the model consists of two parts: the basic part (left), directly utilizing the pre-trained SD, and the conditional part (right), to apply additional conditioning images (e.g. canny edge map, depth map, and so on) to a single instance of Stable Diffusion. Task arithmetic is applied to the

basic part, while the conditional part introduces finer control, enabling more precise and flexible image generation.

4. Experiment

In this section, we rigorously investigate the efficacy of model arithmetic in the context of stable diffusion models. Our focused evaluation examines the performance of task enhancement and task addition capabilities, employing a series of extensive and meticulous experiments to ascertain the task arithmetic performance of diffusion models.

4.1. Experimental Settings

Baseline. In our study, we establish our experimental foundation upon the checkpoints presented in "Ablating Concepts in Text-to-Image Diffusion Models"[12] and "ControlNet"[39]. The former introduces an efficient method to ablate concepts in pre-trained models, crucial for preventing the generation of target concepts without retraining from scratch, while the latter work offers conditional controls to image generation.

Evaluation metrics. In our research, we utilize the CLIP Score and CLIP Accuracy [8] measures to quantitatively assess the alignment between the images generated by our model and the input text concepts. These metrics are crucial for evaluating the model's effectiveness in either enhancing or ablating particular features in response to textual prompts, like making them look more like "Snoopy" or more like a regular "dog".

Models' preparation. Since the task arithmetic method is plug-and-play, requiring no additional training or fine-tuning. Leveraging this method, we utilized already trained models to carry out computational tasks, enabling the generation of new models with modified capabilities. To enhance the generative capabilities of diffusion models, we differentiate original ablation models from the baseline SD model to craft new variants with improved and diversified concept augmentation skills. Using calculated disparities $(\Delta_1, \Delta_2, \Delta_3, \ldots)$, we fuse these with the foundational SD model, resulting in versatile ablation models with sophisticated multi-concept modulation, such as simultaneously removing distinct styles or elements from various characters and themes.

Further refining our technique, we blend negation and addition strategies into the computational framework and embed them within ControlNet's architecture. This dual-pronged approach, employing both subtractive and additive deltas, empowers the models to intensify and fine-tune the generation of multiple nuanced concepts, thus broadening the scope of controlled image synthesis.

4.2. Quantitative and qualitative results

In this section, we use the generation of Monet-related concepts as an example to illustrate style synthesis and Snoopy-related concepts to exemplify instance generation. We investigate the role and efficacy of task arithmetic within diffusion model generation from both qualitative and quantitative perspectives.

The Impact of Task Arithmetic on Stylistic and Instance Generation. To assess the effectiveness of different models in replicating Monet's style and generating specific image instances, we conducted experiments with prompts tailored to each scenario. For style generation, we used "A painting of a city in the style of Monet," and for instance generation, "A playful Snoopy splashing around in a puddle." The results, illustrated in fig. 5, showcase the initial outcomes from the basic SD model, results from an ablated (baseline for comparison) and an enhanced model, and the outputs from models employing Monet & Snoopy ablation. What's more, we use 'A playful dog splashing around in a puddle.' to generate from an enhanced model (in the middle), which illustrates that our enhancement will not have any side effect on the original concept.

The Monet & Snoopy ablation model demonstrates its ability to emphasize Monet's iconic features - vivid brushstrokes and softened edges - or to generate a playful Snoopy image, depending on the task. The enhancement model further enhances these elements by focusing on the differences between the basic SD model and the ablation models.

Quantitative results, presented in table 1a and table 1b, indicate the models' success in closely aligning with Monet's style and accurately generating Snoopy instances, as evidenced by higher CLIP Scores and CLIP Accuracy. "S&M" denotes Snoopy and Monet, while "S&G&N&R&M" includes additional characters for a broader comparison.

These experiments highlight the models' capability to not only perform concept ablation tasks effectively but also to surpass baseline performances, showcasing the potential of task arithmetic in enhancing stylistic generation and image instance creation. Despite challenges in distinguishing between generic and specific instances, such as between a regular dog and Snoopy, and the limitations of CLIP scores as a linear measure, our approach significantly improves the representation of targeted subjects and styles.

Task arithmetic's impact on multifaceted image generation. The illustration of fig. 6 demonstrates task arithmetic across three distinct tasks: object representation, style rendering, and conditional control. The image showcases displays a visual matrix where different models generate variations of "A Snoopy, painting in Monet style" and "A Grumpy Cat, painting in Salvador Dali style," using a depth map and scribble as conditional controls. Starting from the origin on the coordinate axes, we observe concept ablation, default, and enhancement, respectively. fig. 7 and fig. 8 present more visualization results. They show different popular characters, Snoopy and Grumpy Cat, each

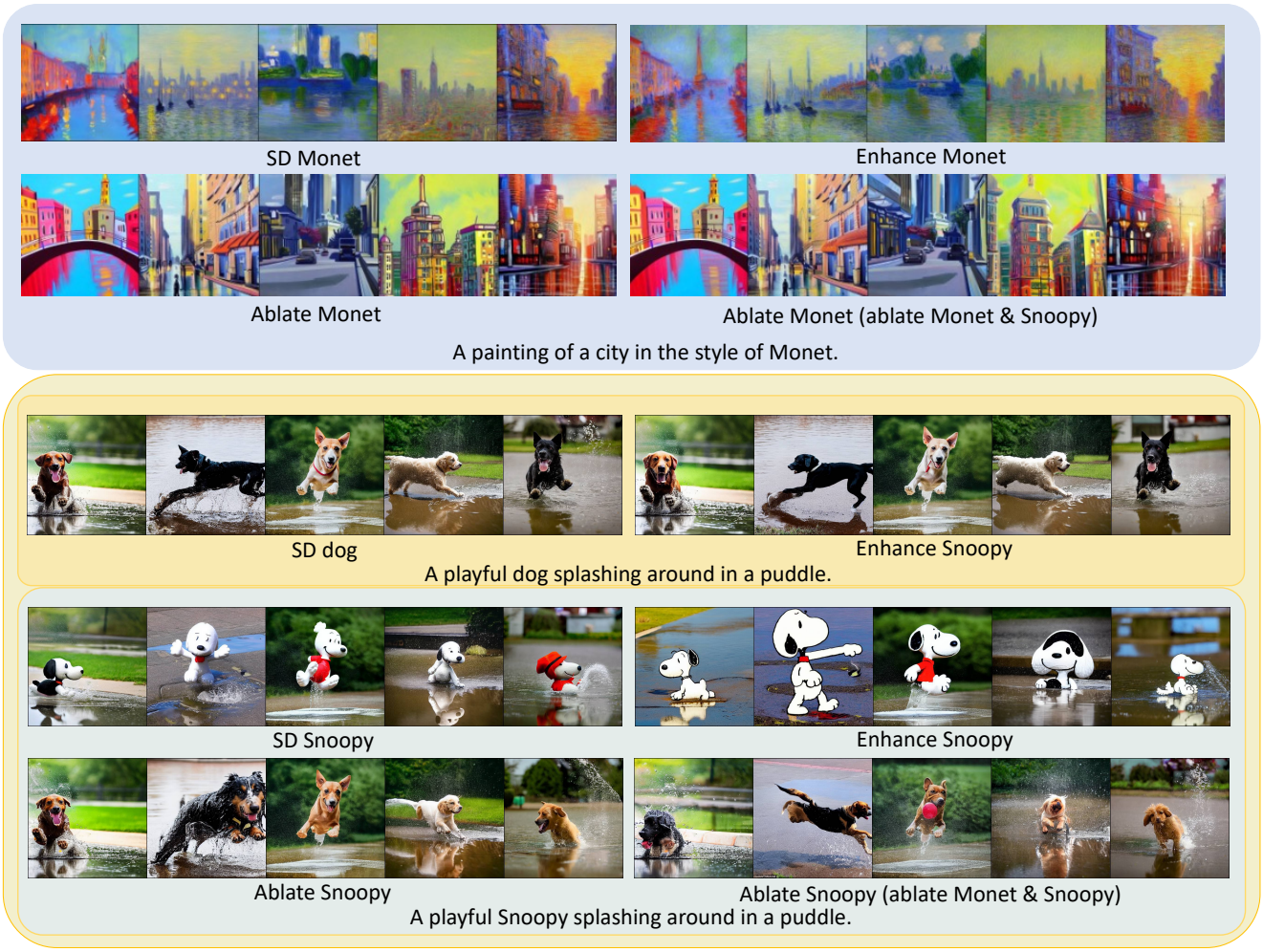


Figure 5: **Visualization of different models related to style and object generation.** Using Monet's style as an example of style and Snoopy as the object example, visualizations of both task addition and enhancement have achieved notable results.

Table 1: **Quantitative results of different models related to style and object generation.**Taking Monet generation as an example of style generation and Snoopy generation as an example of object generation, the application of both task addition and task enhancement has shown significant achievements.

(a) **Quantitative results of different models related to Monet generation.**

Experiments(↑)(↓)	CLIP Score	CLIP Acc.
SD Monet	0.763	1.00
Ablate Monet	0.637	0.54
Ablate Monet(ab S&M)	0.629	0.46
Ablate Monet(ab S&G&N&R&M)	0.625	0.5
Enhance Monet	0.787	1.00
Enhance Monet(en S&G&N&R&M)	0.810	1.00

(b) **Quantitative results of different models related to Snoopy generation.**

Experiments(↑)(↓)	CLIP Score	CLIP Acc.
SD Snoopy	0.746	0.94
Ablate Snoopy	0.576	0.04
Ablate Snoopy(ab S&M)	0.581	0.02
Ablate Snoopy(ab S&G&N&R&M)	0.572	0.00
Enhance Snoopy	0.802	1.00
Enhance Snoopy(en S&G&N&R&M)	0.819	1.00

modified through various artistic styles and conditional controls(canny edge map or scribble) to ablate or enhance certain features.

Models that are not orthogonal can also yield good results in task addition. To delve deeper into the vast potential of task arithmetic within stable diffusion models, we

Figure 6: **Visualization of task arithmetic applied to three types of tasks(object, style and conditional control).** Task arithmetic enabled models designed for three specific tasks to exhibit commendable performance within a single model, without training and no extra model storage.

conducted integration tests on a greater variety of models. Unlike task addition for instances and styles, we overlaid multiple instance concept-ablation models. Task vector cosine similarity was used to measure the distance between tasks. During experiments, the average distances of style and instance task vectors were calculated. A cosine similarity close to 0 indicates orthogonality, and the results shown above indicate that task arithmetic performs well in such cases. Further experimentation revealed that even with high task vector cosine similarity among instances, close to 1, indicating high task similarity, the fusion still achieved good results and demonstrated effective weight disentanglement of the model. As shown in fig. 9, the unified model was then tested across each concept-ablation task. The results confirm the model's effectiveness, successfully omitting the relevant concept in every task, a clear indication of achieving weight disentanglement and multi-task proficiency.

4.3. Ablation Study

In our ablation study, we employ visual demonstrations to showcase the generative outcomes of models derived from various combinations of Snoopy instances and Monet style elements. Furthermore, we delve into a quantitative analysis, taking Snoopy as a focal point, to examine the impact of different magnitudes of task combination on the model's output. This investigation allows us to better understand the intricate relationship between task weighting and generative quality.

The illustration of the process of ablation experiments is shown in fig. 11 in two segments. In part (a), the experimental setup showcases the fusion of evaluation prompts from Monet and Snoopy, creating a series of prompts that contain both Monet style and Snoopy for input. Part (b) shows the outcomes of using task enhancement on the basic SD model and Monet ablation and Snoopy ablation models. This procedure yields intensified models for both Monet and Snoopy. Subsequently, various combinations of these models are integrated, resulting in seven distinct blends: the basic SD model, models that ablate or enhance Snoopy and Monet simultaneously or selectively, and models that either enhance or suppress one while leaving the other unaffected.

In fig. 13, we showcase the performance of seven different models in creating images based on the prompt "A Monet-inspired painting of a sunset, with a grateful Snoopy giving its owner a loving look." The first row features the basic SD model's interpretation. The second row shows models that remove (ablate) both Snoopy and Monet's style, adding a unique capability. The third row amplifies both elements, acknowledging that Snoopy rendered in Monet's style is an unusual combination, and our model's attempt to visualize the described scene could be limited by the foundational SD model's capabilities in this artistic area. Rows four and five focus on selectively removing Monet and Snoopy's influences, respectively. The last two rows highlight these characters individually, offering a detailed look at the balance between removing and enhancing specific aspects of artistic style and character representation.

We conducted further exploration of how different levels of the ablation task impact the resemblance of the generated images to the original Snoopy concept. Using the

Figure 7: **Visualization of Grumpy Cat-Dali ablation/ enhancement experiments under the conditional control of a scribble.**

Experiments(↓)	CLIP Score	CLIP Accuracy
SD S	0.746	0.94
Ablate S	0.576	0.04
Ablate 0.9S	0.579	0.02
Ablate 0.7S	0.599	0.04
Ablate 0.5S	0.622	0.14
Ablate 0.3S	0.658	0.38
Ablate 0.1S	0.713	0.86
Ablate 0.01S	0.743	0.94
Ablate 0.001S	0.747	0.94
Ablate 0.0001S	0.745	0.94

Table 2: **Quantifying the impact of gradual ablation.** Taking Snoopy as an example, denoted by 'S', we employed the method of task arithmetic to examine the effects of various degrees of Snoopy ablation on model-generated images.

pre-trained SD model, we performed experiments by adding different coefficients of delta. We compare the CLIP Score

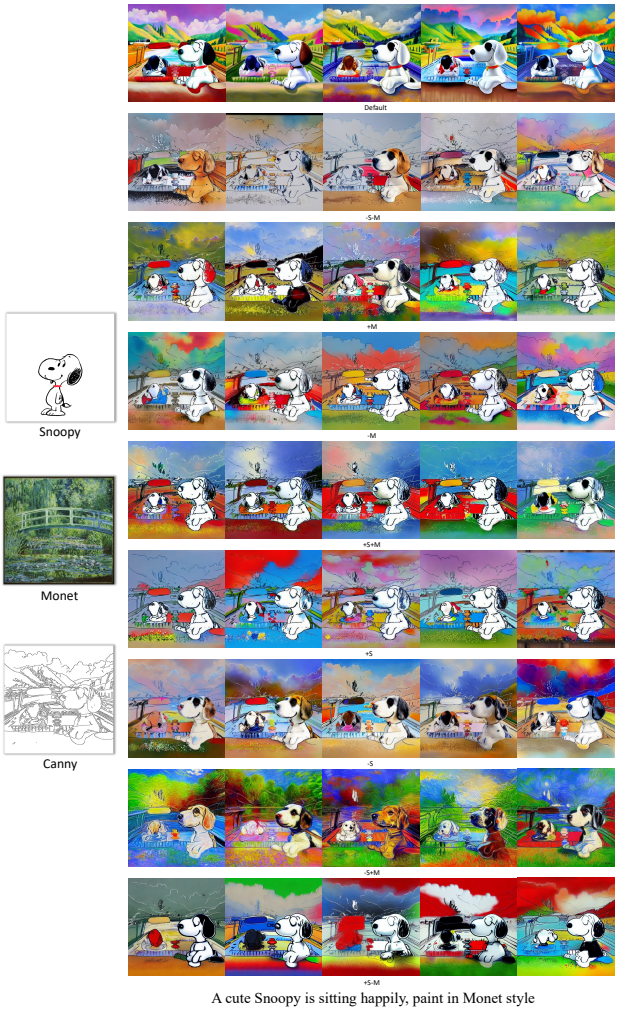


Figure 8: **Visualization of Snoopy-Monet ablation/ enhancement experiments under the conditional control of an edge map.**

and CLIP Accuracy across different levels of Snoopy ablation in table 2. The scores reflect how closely the generated images align with the concept of Snoopy, while the accuracy indicates the model's ability to correctly generate or ablate the Snoopy concept. As the intensity of the ablation decreases (indicated by decreasing multipliers like 0.9, 0.7, down to 0.0001), the CLIP Score generally increases, showing a trend towards more Snoopy-like features in the images.

4.4. Weight disentanglement emerges during pre-training

Task arithmetic is not exclusive to diffusion models. In fact, task arithmetic can also be performed on pre-trained text transformers and convolutional neural networks [18]. To investigate this, we investigate the task addition of ControlNets with randomly initialized convolution layers (convs). The results in fig. 12 reveal that task arithmetic

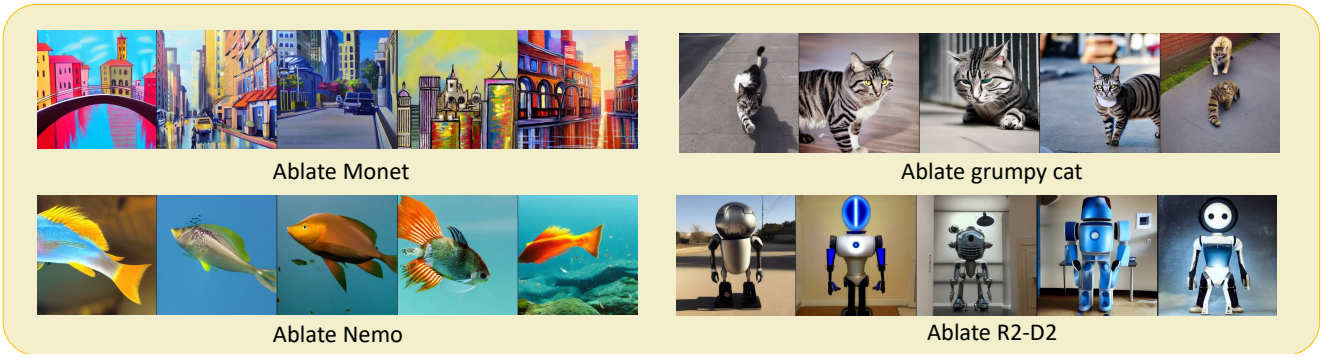


Figure 9: **Visualization of an integrated model ablating different concepts.** Through the method of task addition, we merged models ablated for Monet, Grumpy Cat, Nemo, R2-D2, and Snoopy into a single model capable of multi-task operations.

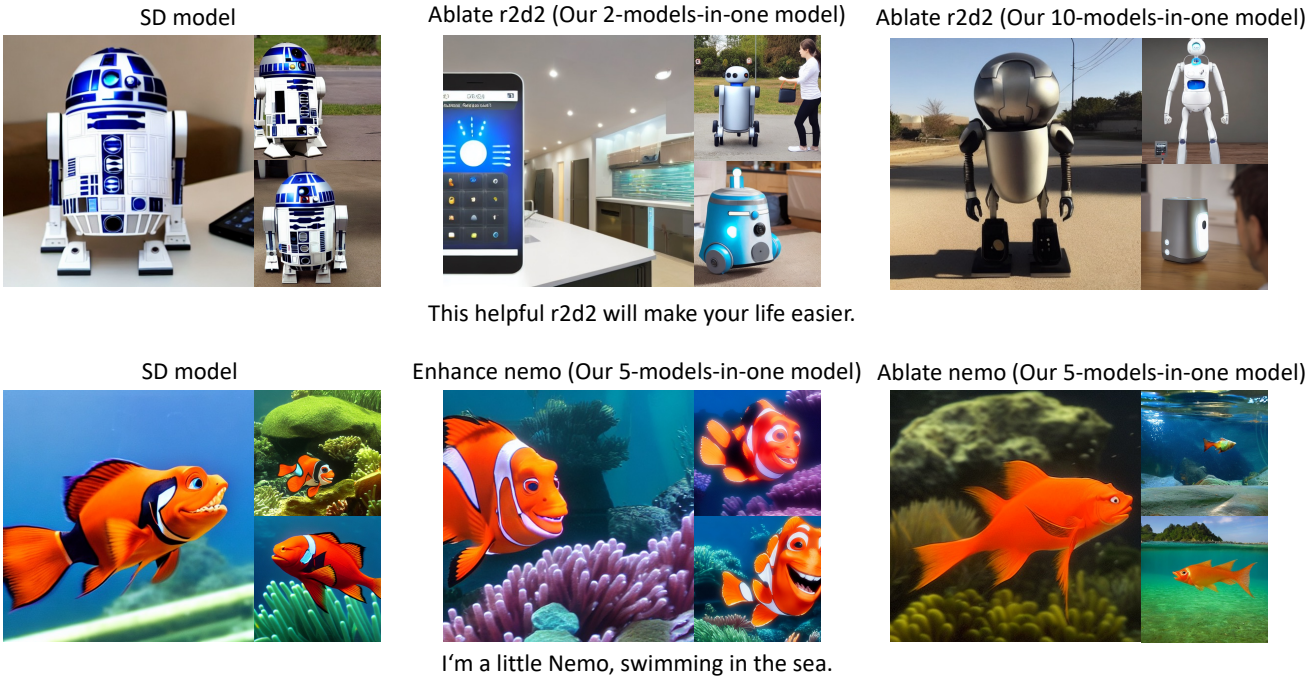


Figure 10: **Visualization of more experiments.**

is not achievable on randomly initialized convs. Indeed, adding task vectors obtained from a random initialization does not result in significant improvements in multi-task performance over random chance. To the best of our knowledge, it violates the assumptions of eq. (4), as a result, task arithmetic is a property acquired during pre-training.

4.5. Comparison with Unified Concept Editing

Images in fig. 14 present a comparison of different model outputs for the prompt at the bottom of the figure. The symbols used are defined as follows:

- ”Ab” denotes ”ablate”, signifying the removal or weakening of a specific concept in the generated image.
- ”En” represents ”enhance”, indicating the strengthening or augmentation of a particular concept.
- ”S” refers to Snoopy, ”M” to Monet, ”G” to Grumpy, and ”R” to R2D2.
- In parentheses, the use of ”&” to connect these abbreviations (e.g., S&M, S&G&N&R&M) indicates that the corresponding model is a fusion of the respective single-task models associated with those concepts.

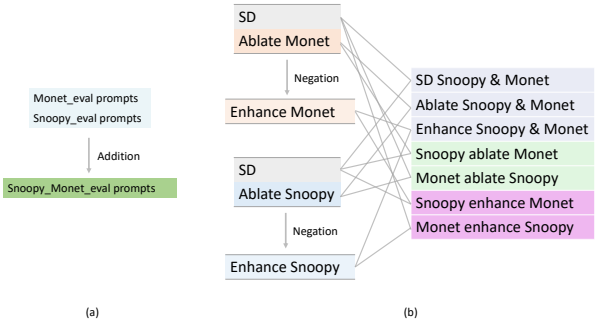


Figure 11: **The process of ablation experiments.** Cross-experimentation involving style enhancement/ablation (using Monet as an example) and object enhancement/ablation (with Snoopy as the example) demonstrates the intricate interplay between style and object representation.

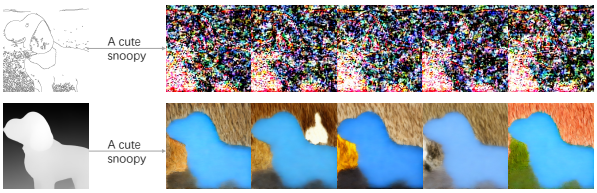


Figure 12: **Failure cases of violating the constraints.** After randomly initializing convolution layers in the right part of ControlNet, we engage in task arithmetic to combine tasks involving an edge map and a depth map. Subsequently, by using only a single conditional control as input (with the upper part of the results utilizing an edge map and the lower part a depth map), the output does not achieve the anticipated target.

The green-framed section at the top provides reference images for contrast. We can see the images generated by Unified Concept Editing[7], though we edit it to ablate Monet style, it's more like Monet, especially the unobtrusive shading and outline strokes. Below that, in the purple frame, are the results of a single multi-in-one model applied to a coupling task. The image on the left results from the fusion of two models, while the image on the right shows the output when multiple models are fused, indicating a multi-model integration approach. The bottom section, encased in blue, displays models that have undergone negation followed by addition operations to enhance the Monet style. Similarly, the left image in this section represents the fusion of two models, and the right image demonstrates the fusion of multiple models.

In fig. 10, further experiments validated that our model enhanced concepts through task negation and after merging multiple models using task addition, its performance on the original tasks not only remained stable but even improved.

5. Discussion and Limitations

We emphasize the substantial potential of task arithmetic in stable diffusion models, highlighting its effectiveness in multi-task operations and weight disentanglement. The study showcases the model's ability to integrate and amplify various concepts. The experiments demonstrate that task arithmetic, including both enhancement and addition, can significantly enhance model functionality without additional training, suggesting a new paradigm in model utilization for complex image generation tasks.

We carried out experiments in task arithmetic on specific tasks spanning multiple categories, including objects, styles, and conditional control. In the realms of objects and styles, we employed task enhancement to achieve concept enhancement. All outcomes aligned perfectly with our expectations.

In our experiments, we identified certain limitations associated with task arithmetic. Specifically, if the base model begins with a random initialization, subtracting this model from one trained on a specific task does not accurately produce the intended task vector. As a result, conducting task arithmetic operations with these imprecise task vectors fails to achieve the anticipated outcomes. However, employing a pre-trained model such as SD-v-1.4 as the base model for task arithmetic yields favorable results.

Acknowledgement

We are grateful to Jinkun Hao, Haoming Chen, Zhengyuan Peng, Fulin Qi for their helpful discussion. This work was supported in part by the National Natural Science Foundation of China (62302167, U23A20343, 62222602, and 62176092); in part by Shanghai Sailing Program (23YF1410500); in part by Chenguang Program of Shanghai Education Development Foundation and Shanghai Municipal Education Commission (23CGA34).

References

[1] A. Aghajanyan, L. Zettlemoyer, and S. Gupta. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*, 2020.

[2] T. Brooks, A. Holynski, and A. A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023.

[3] Y. Cao, S. Li, Y. Liu, Z. Yan, Y. Dai, P. S. Yu, and L. Sun. A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. *arXiv preprint arXiv:2303.04226*, 2023.

[4] H. Chen, J. Gu, Y. Liu, S. A. Magid, C. Dong, Q. Wang, H. Pfister, and L. Zhu. Masked image training for generalizable deep image denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1692–1703, 2023.

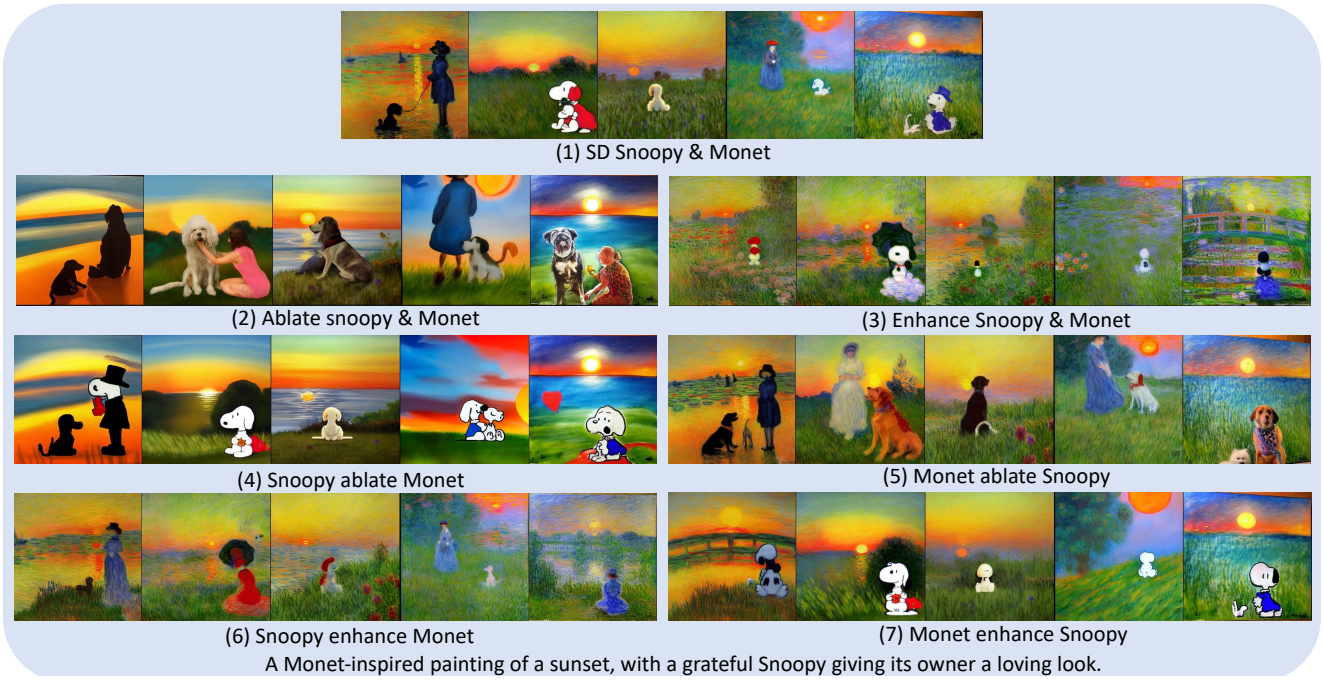


Figure 13: **Visualization of different models generating Snoopy and Monet style.** Same integration (addition) model on different tasks.

[5] G. Couairon, J. Verbeek, H. Schwenk, and M. Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022.

[6] R. Gandikota, J. Materzynska, J. Fiotto-Kaufman, and D. Bau. Erasing concepts from diffusion models. *arXiv preprint arXiv:2303.07345*, 2023.

[7] R. Gandikota, H. Orgad, Y. Belinkov, J. Materzyńska, and D. Bau. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 5111–5120, 2024.

[8] J. Hessel, A. Holtzman, M. Forbes, R. L. Bras, and Y. Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.

[9] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[10] G. Ilharco, M. T. Ribeiro, M. Wortsman, L. Schmidt, H. Hajishirzi, and A. Farhadi. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*, 2022.

[11] B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, and M. Irani. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6007–6017, 2023.

[12] N. Kumari, B. Zhang, S.-Y. Wang, E. Shechtman, R. Zhang, and J.-Y. Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22691–22702, 2023.

[13] B. Li, X. Qi, T. Lukasiewicz, and P. Torr. Controllable text-to-image generation. *Advances in Neural Information Processing Systems*, 32, 2019.

[14] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li, and Y. J. Lee. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22511–22521, 2023.

[15] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022.

[16] C. Meng, R. Rombach, R. Gao, D. Kingma, S. Ermon, J. Ho, and T. Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14297–14306, 2023.

[17] R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6038–6047, 2023.

[18] G. Ortiz-Jimenez, A. Favero, and P. Frossard. Task arithmetic in the tangent space: Improved editing of pre-trained models. *Advances in Neural Information Processing Systems*, 36, 2024.

[19] G. Ortiz-Jiménez, A. Modas, S.-M. Moosavi-Dezfooli, and P. Frossard. What can linearized neural networks actually say about generalization? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

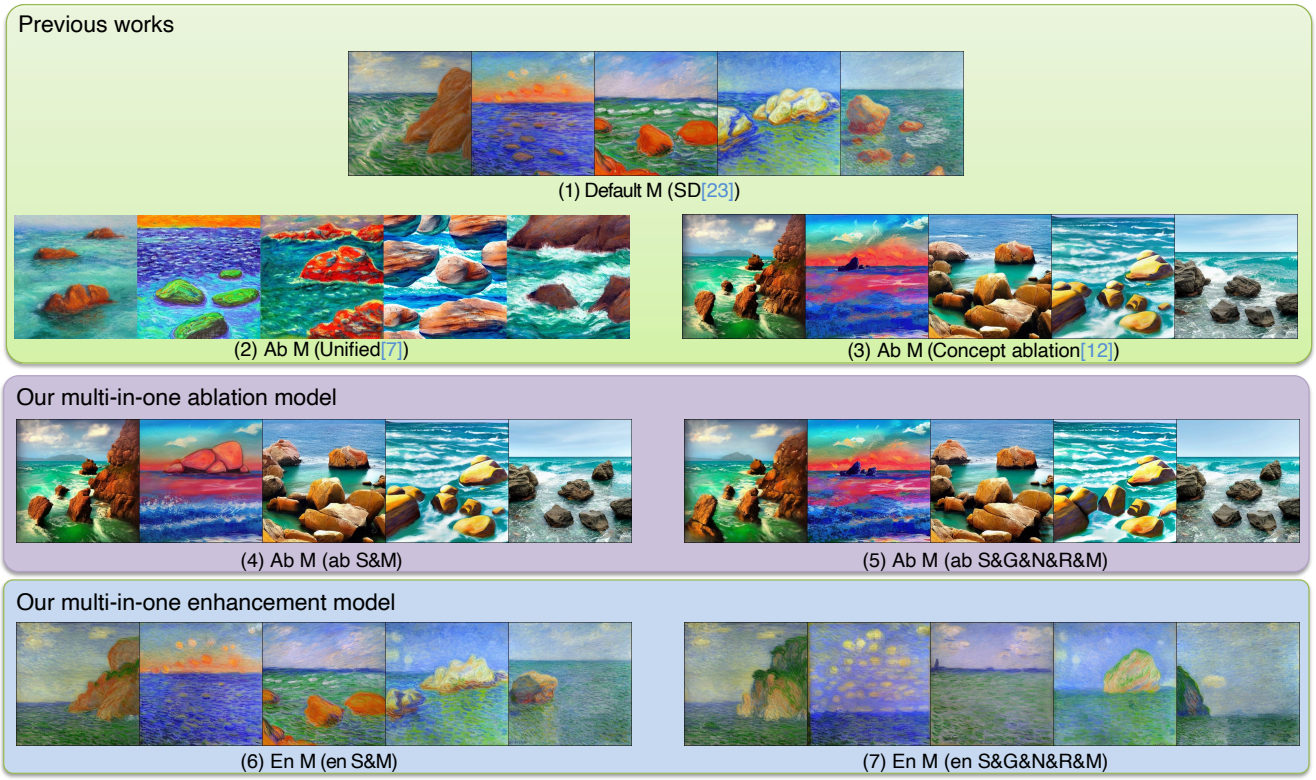


Figure 14: **Comparison of our multi-in-one model with other models.** This image illustrates a comparison of a multi-in-one model achieved through task arithmetic against other works on a singular task. The comparison underscores the efficacy of task arithmetic in creating a multi-in-one model approach, showcasing its capability to handle a single task with enhanced effect and artistic flair.

[20] G. Parmar, K. Kumar Singh, R. Zhang, Y. Li, J. Lu, and J.-Y. Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–11, 2023.

[21] T. Qiao, J. Zhang, D. Xu, and D. Tao. Mirrorgan: Learning text-to-image generation by redescription. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1505–1514, 2019.

[22] C. Qin, S. Zhang, N. Yu, Y. Feng, X. Yang, Y. Zhou, H. Wang, J. C. Niebles, C. Xiong, S. Savarese, et al. Unicontrol: A unified diffusion model for controllable visual generation in the wild. *arXiv preprint arXiv:2305.11147*, 2023.

[23] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. 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.

[24] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22500–22510, 2023.

[25] V. Shah, N. Ruiz, F. Cole, E. Lu, S. Lazebnik, Y. Li, and V. Jampani. Ziplora: Any subject in any style by effectively merging loras. *arXiv preprint arXiv:2311.13600*, 2023.

[26] S. Sheynin, A. Polyak, U. Singer, Y. Kirstain, A. Zohar, O. Ashual, D. Parikh, and Y. Taigman. Emu edit: Precise image editing via recognition and generation tasks. *arXiv preprint arXiv:2311.10089*, 2023.

[27] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, 2015.

[28] E. Spolaore and R. Wacziarg. The diffusion of development. *The Quarterly journal of economics*, 124(2):469–529, 2009.

[29] N. Tumanyan, M. Geyer, S. Bagon, and T. Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023.

[30] S. Wang, C. Saharia, C. Montgomery, J. Pont-Tuset, S. Noy, S. Pellegrini, Y. Onoe, S. Laszlo, D. J. Fleet, R. Soricut, et al. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18359–18369, 2023.

[31] J. Wu, W. Gan, Z. Chen, S. Wan, and H. Lin. Ai-generated content (aigc): A survey. *arXiv preprint arXiv:2304.06632*, 2023.

[32] S. Xie, Z. Zhang, Z. Lin, T. Hinz, and K. Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22428–22437, 2023.

[33] S.-Z. Xu and Y.-K. Lai. Simultaneous multi-attribute image-to-image translation using parallel latent transform networks. In *Computer Graphics Forum*, volume 39, pages 531–542. Wiley Online Library, 2020.

[34] Z. Yang, G. Ding, W. Wang, H. Chen, B. Zhuang, and C. Shen. Object-aware inversion and reassembly for image editing. In *The Twelfth International Conference on Learning Representations*, 2023.

[35] Z. Yang, J. Wang, Z. Gan, L. Li, K. Lin, C. Wu, N. Duan, Z. Liu, C. Liu, M. Zeng, et al. Reco: Region-controlled text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14246–14255, 2023.

[36] H. Ye, J. Zhang, S. Liu, X. Han, and W. Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

[37] G. Yüce, G. Ortiz-Jiménez, B. Besbinar, and P. Frossard. A structured dictionary perspective on implicit neural representations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[38] C. Zhang, C. Zhang, M. Zhang, and I. S. Kweon. Text-to-image diffusion model in generative ai: A survey. *arXiv preprint arXiv:2303.07909*, 2023.

[39] L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

[40] Y. Zhou, R. Zhang, C. Chen, C. Li, C. Tensmeyer, T. Yu, J. Gu, J. Xu, and T. Sun. Towards language-free training for text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17907–17917, 2022.