

Генерация изображений на основе текстовых описаний с сохранением уникальных характеристик субъекта

Курсовой исследовательский проект

Выполнил:

Теплов Андрей Романович БКНАД231

Руководитель проекта:

Мамедов Ильгар Габирович



Subject-Driven Image Generation

- есть 1–4 исходных фотографий конкретного объекта
- хотим научиться генерировать этот же объект, а не просто похожий объект
- объект должен появляться в новых позах, местах, стилях и сценариях
- при малом числе фотографий модель быстро запоминает не только объект, но и фон, освещение, ракурс и позу



Input images



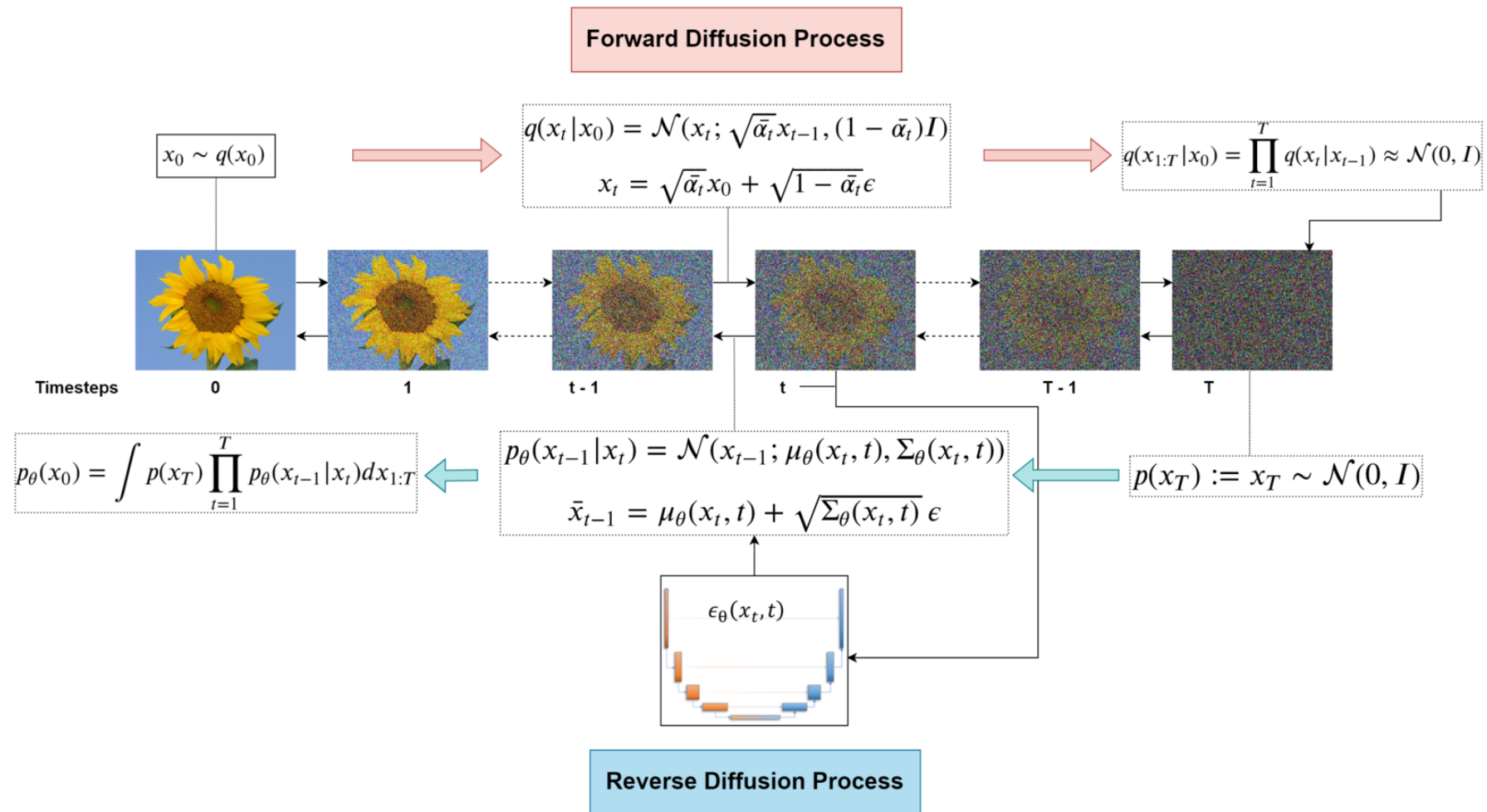


Обзор литературы





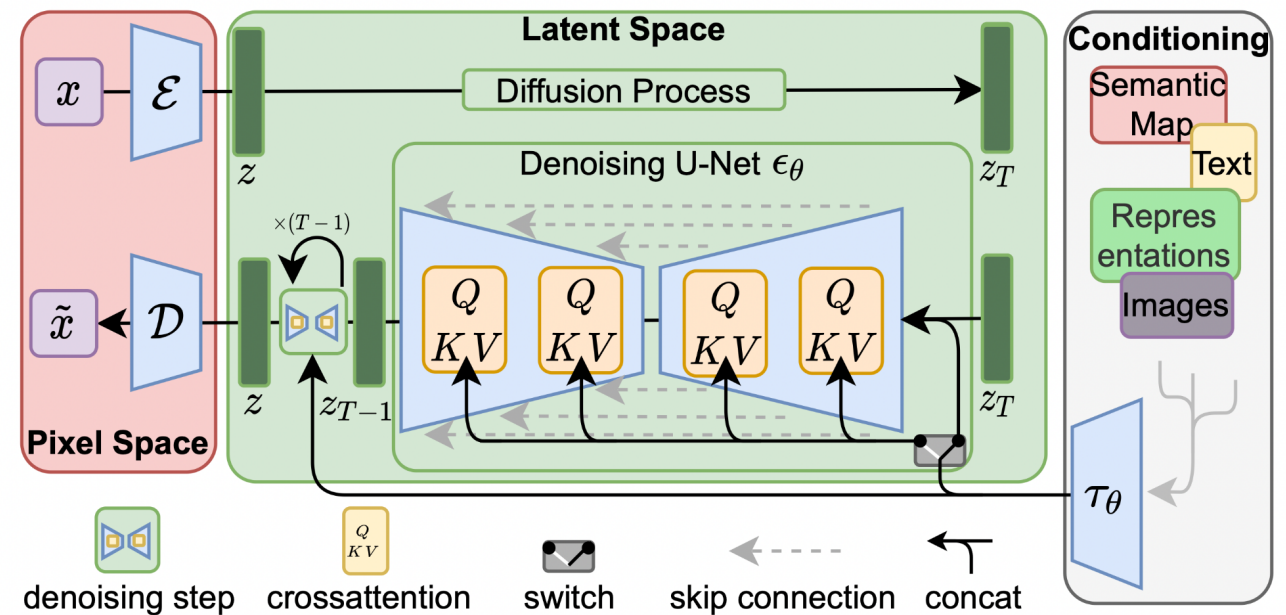
DDPM





Stable Diffusion

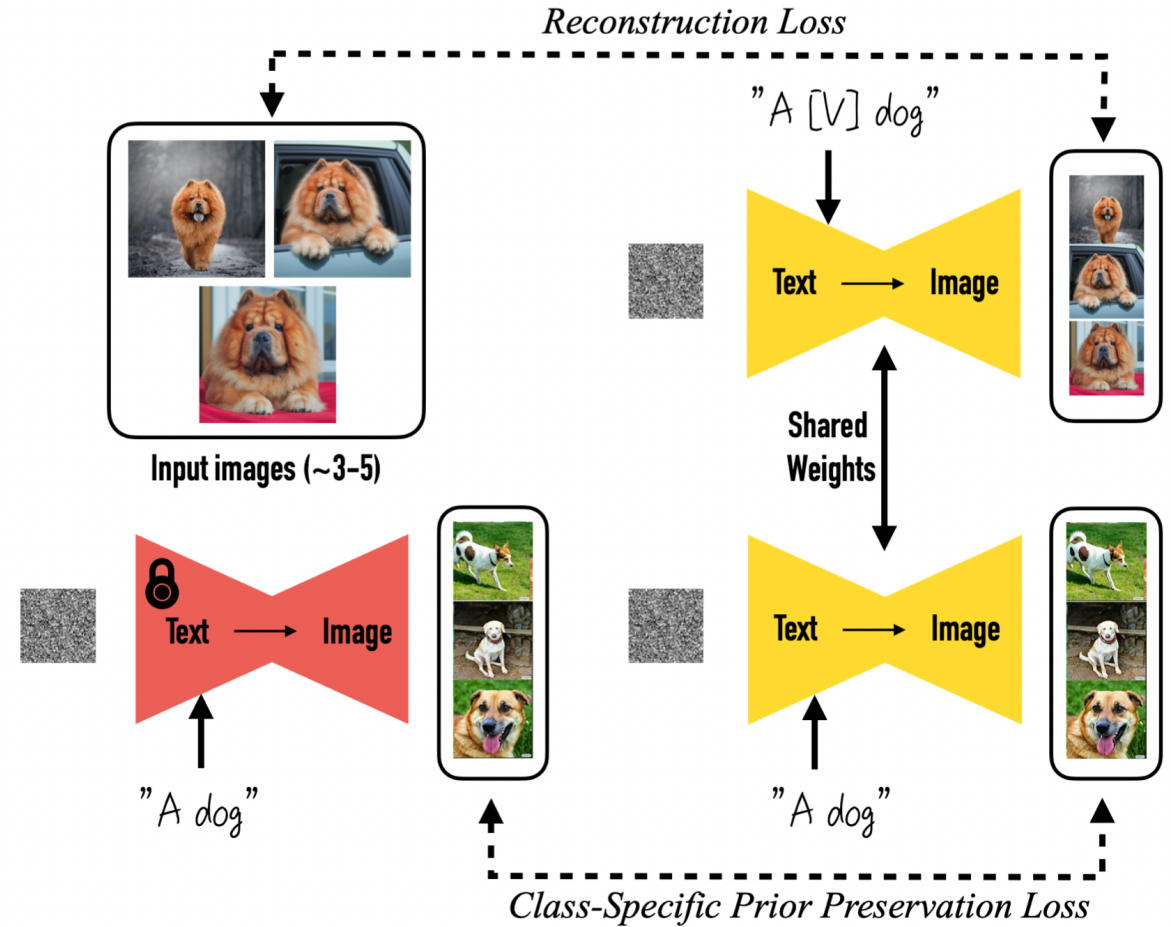
- изображение кодируется с помощью VAE латентное представление;
- к латентному представлению постепенно добавляется шум, модель учится удалять шум;
- U-Net на каждом шаге получает зашумлённый латент, timestep и prompt;
- prompt кодируется с помощью CLIP и передаётся в U-Net через cross-attention;
- после завершения обратного диффузионного процесса полученный латент декодируется через VAE в изображение;





Dreambooth

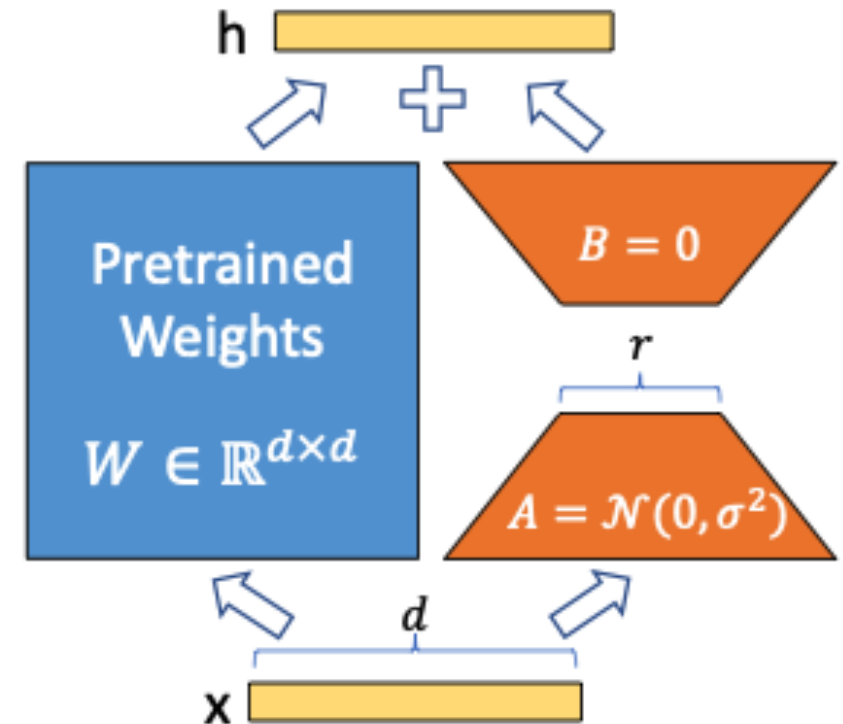
- Вводится новый token объекта и token его класса
- При обучении на малом количестве данных модель может забывать исходные знания.
- Prior Preservation Loss:
 - Перед обучением генерируются изображения по описанию класса.
 - Модель обязана воспроизводить их, что сохраняет общее качество генерации.





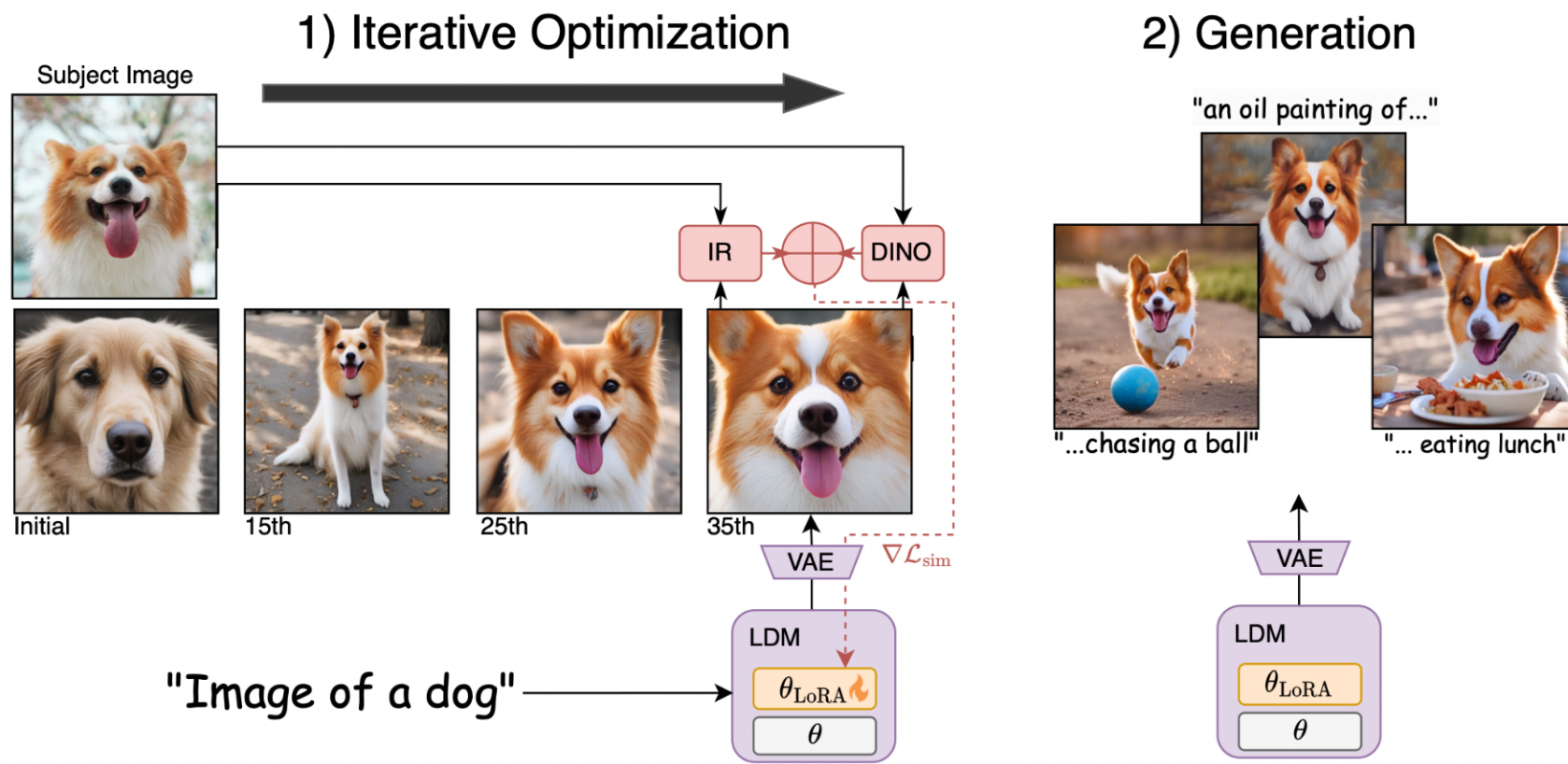
Dreambooth & LoRA

- Метод сочетает идею Dreambooth с лёгким адаптационным слоем LoRA.
- Вместо дообучения всех весов модели, корректирующие матрицы добавляются к выбранным слоям основной модели и текстового энкодера.
- LoRA снижает риск переобучения и делает процесс быстрее и дешевле, сохраняя при этом способность Dreambooth точно передавать особенности объекта.



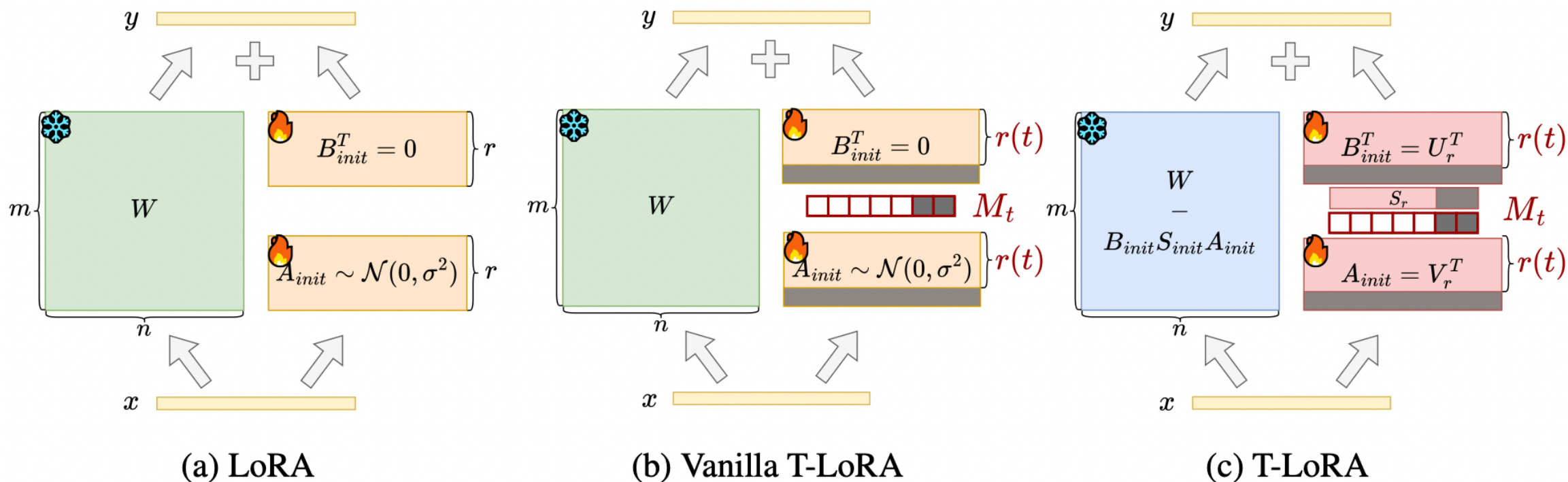


SISO: Single Image Iterative Subject-driven Generation and Editing





T-LoRA: Single Image Diffusion Model Customization Without Overfitting





Проблематика



Повысить устойчивость персонализации при малом количестве обучающих изображений, сохранив баланс между узнаваемостью объекта и гибкостью генерации.





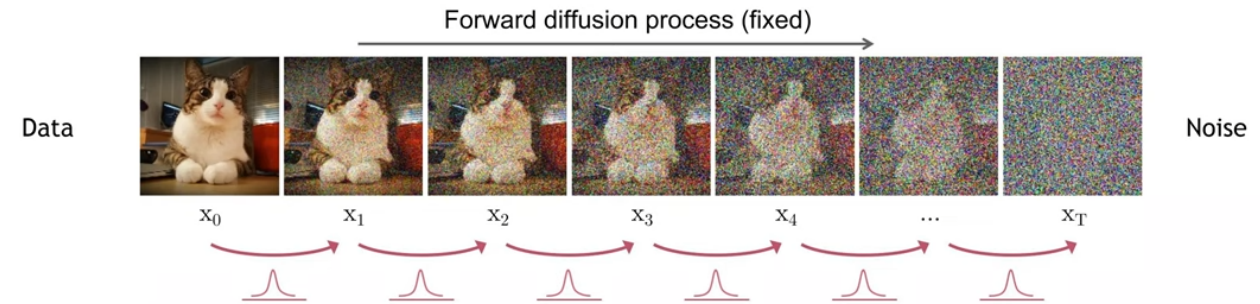
Метод





Гипотеза

- Мелкие детали ($t \approx 0$) важны для узнаваемости объекта.
- Крупные детали ($t > 500$) задают форму и структуру, но менее критичны.
- Стратегия: чаще сэмплировать важные шаги (низкий шум), реже — менее важные (высокий шум), но не игнорировать их.
- Метод: Importance Sampling с biased timestep sampling.





Importance Sampling

Importance Sampling – это метод Монте-Карло, который позволяет эффективнее оценивать математическое ожидание, выбирая более важные значения чаще.

$$\mathbb{E}_{x \sim p(x)}[f(x)] = \int f(x)p(x)dx$$

Идея Importance Sampling: вместо исходного распределения $p(x)$ использовать другое распределение $q(x)$, из которого удобнее или полезнее делать выборку, чтобы оценка осталась корректной, каждый пример домножается на вес:

$$\mathbb{E}_{x \sim p(x)}[f(x)] = \mathbb{E}_{x \sim q(x)} \left[\frac{p(x)}{q(x)} f(x) \right]$$





Importance Sampling in DDPM

Base Diffusion:

$$\mathcal{L} = \mathbb{E}_{t \sim U(1, \dots, T), x_0, \epsilon} \left[\left\| \epsilon - \epsilon_{\theta}(x_t, t) \right\|^2 \right]$$

P-LoRA:

$$\mathcal{L} = \mathbb{E}_{t \sim U(1, \dots, T), x_0, \epsilon} = \mathbb{E}_{t \sim q(t), x_0, \epsilon} \left[\frac{p(t)}{q(t)} \cdot \left\| \epsilon - \epsilon_{\theta}(x_t, t) \right\|^2 \right]$$





Left P-LoRA & Right P-LoRA

Left P-LoRA:

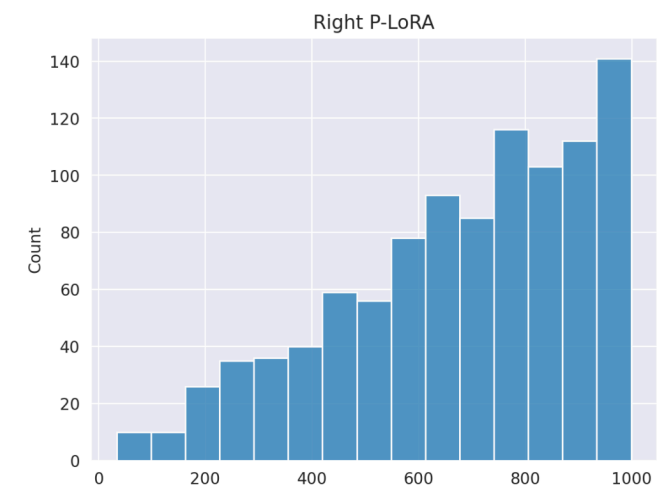
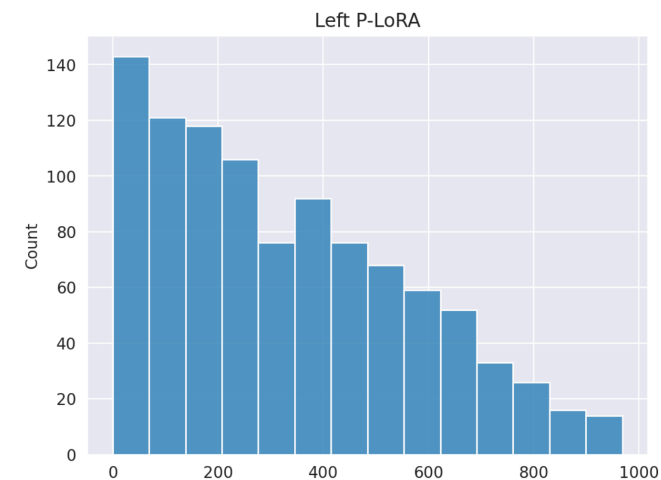
$$q_{\text{left}}(t) \propto (1 - \tau)^\alpha, \quad \alpha > 0$$

Распределение концентрирует вероятность вблизи $t = 0$, где латентное представление сохраняет детальную информацию о внешнем виде

Right P-LoRA:

$$q_{\text{right}}(t) \propto \tau^\alpha, \quad \alpha > 0$$

Это распределение концентрирует вероятность вблизи $t = T$. Акцент на глобальной структуре объекта.





U-Shape P-LoRA & Center P-LoRA

U-Shape P-LoRA:

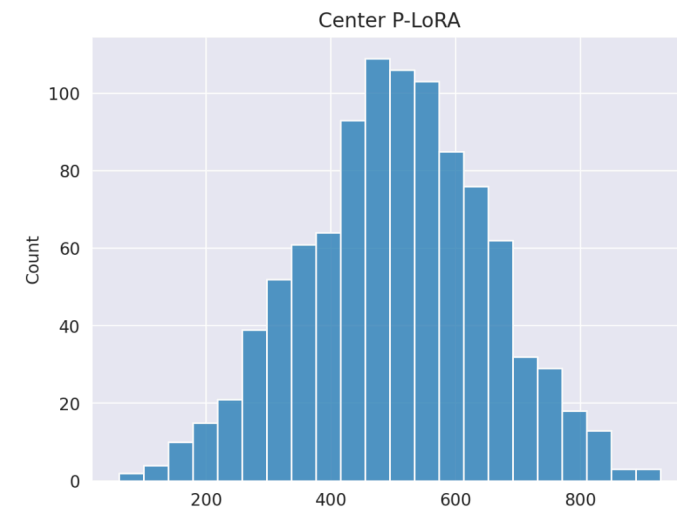
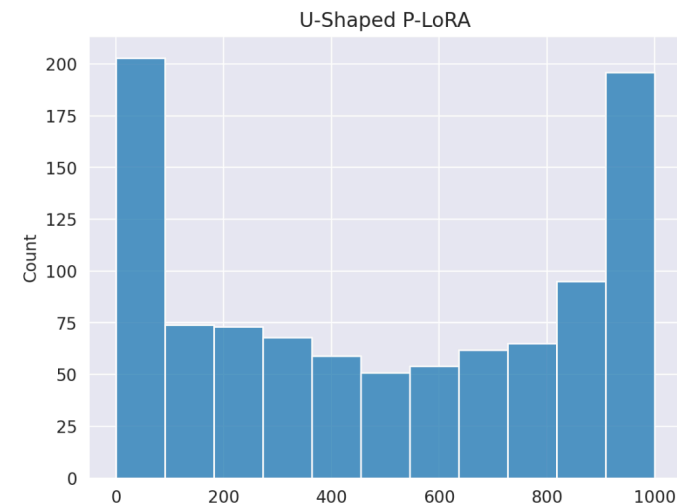
$$q_U(t) \propto \tau^{\alpha-1}(1 - \tau)^{\alpha-1}, \quad 0 < \alpha < 1$$

Распределение объединяет обучающий сигнал тонких деталей с сигналом грубой структуры и оставляя меньше массы на средних шумах.

Center P-LoRA:

$$q_U(t) \propto \tau^{\alpha-1}(1 - \tau)^{\alpha-1}, \quad \alpha > 1$$

Это распределение концентрирует вероятность вблизи $t = T/2$, где обучающий сигнал представляет компромисс между деталями и глобальной структурой.





Режимы обучений

One Shot



Few Shot

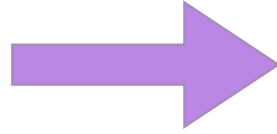




Метрики

Reconstruction score

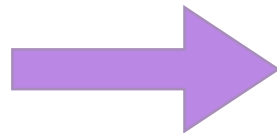
$$R = \frac{1}{nm} \sum_{i,j} \cos \left(\phi_{\text{img}} \left(x_i^{\text{gen}} \right), \phi_{\text{img}} \left(x_j^{\text{train}} \right) \right)$$



среднее попарное CLIP-сходство между
сгенерированными и обучающими
изображениями

Editability score

$$E = \cos \left(\bar{\phi}_{\text{img}}, \phi_{\text{text}}(p) \right), \quad \bar{\phi}_{\text{img}} = \frac{1}{n} \sum_{i=1}^n \phi_{\text{img}} \left(x_i^{\text{gen}} \right)$$



косинусное сходство между средним
эмбедингом изображений и
эмбедингом текста





Объем генераций

- Для оценки используются три группы промптов: base prompt, object_set и live_set
- Base prompt = "a photo of sks dog". Используется для reconstruction score
- В object_set (N=31) объект меняется по стилю, одежде, позе или действию: супергеройский костюм, балерина, скейтборд, прыжок с парашютом
- В live_set (N=31) объект сохраняется, но фон и контекст меняются: пляж, деревянный пол, Эйфелева башня и другие места
- На каждый промпт генерируется 15 изображений



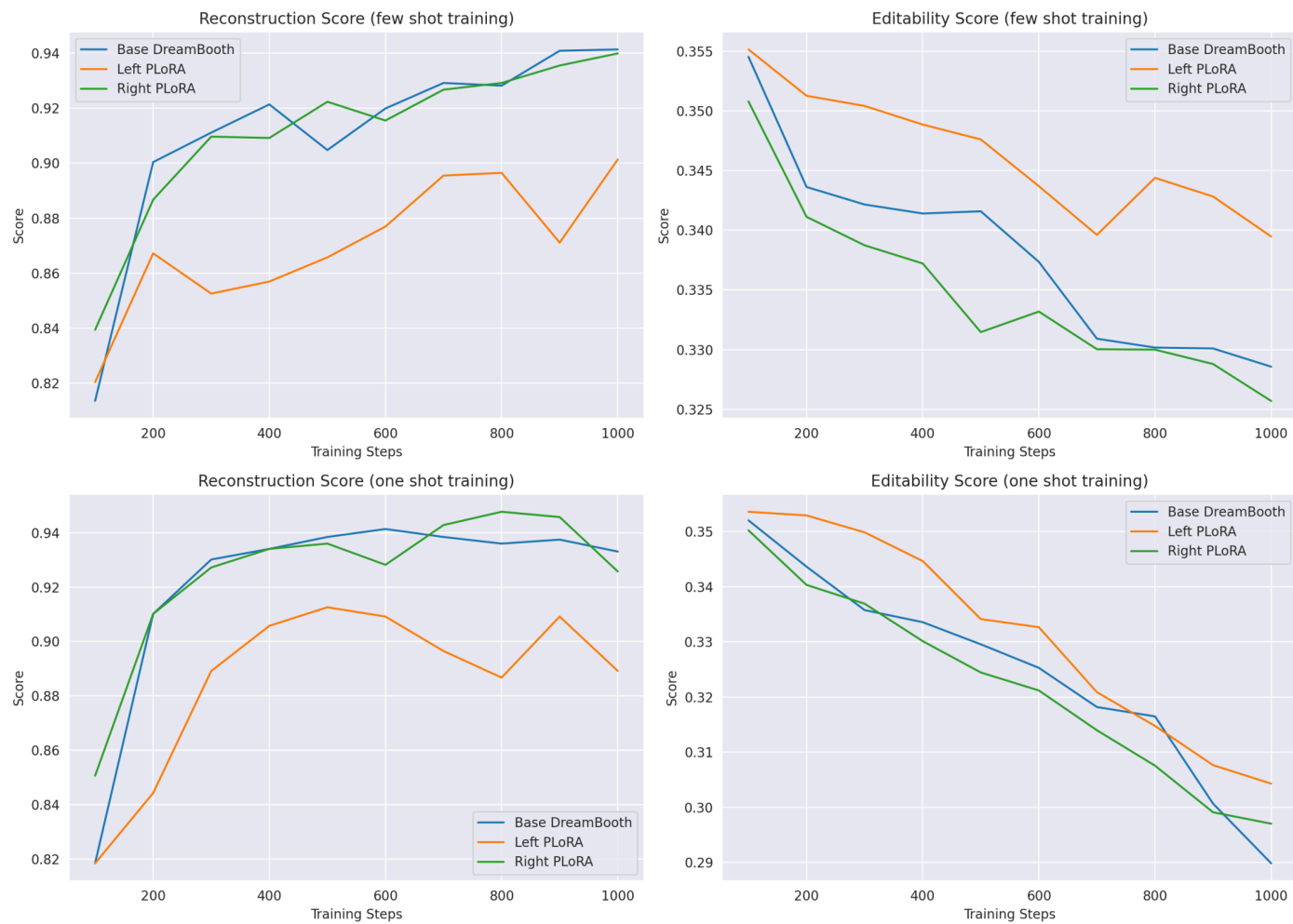


Результаты экспериментов





Результаты: Left P-LoRA и Right P-LoRA

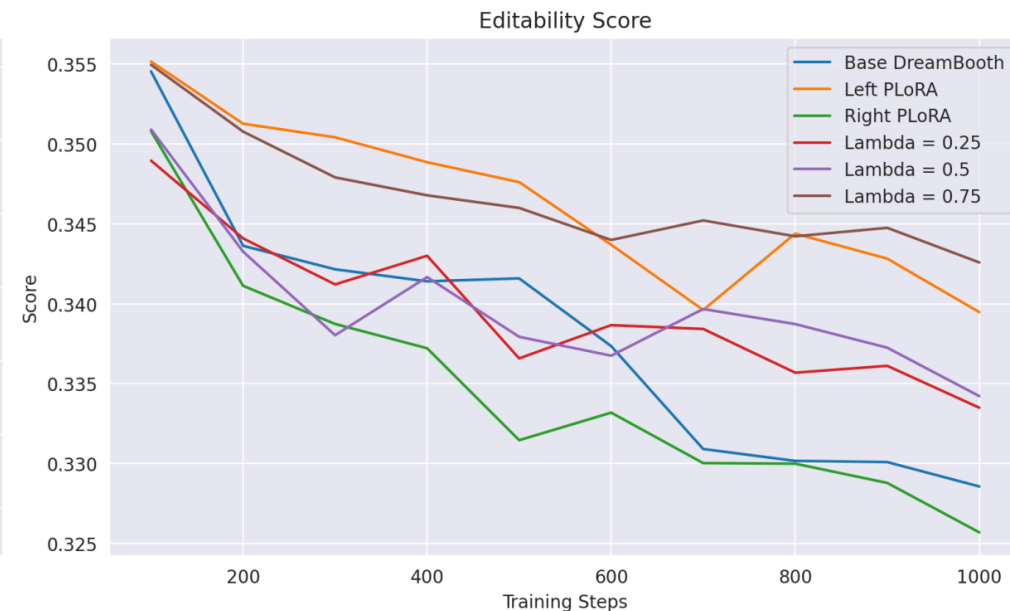




Mixture P-LoRA

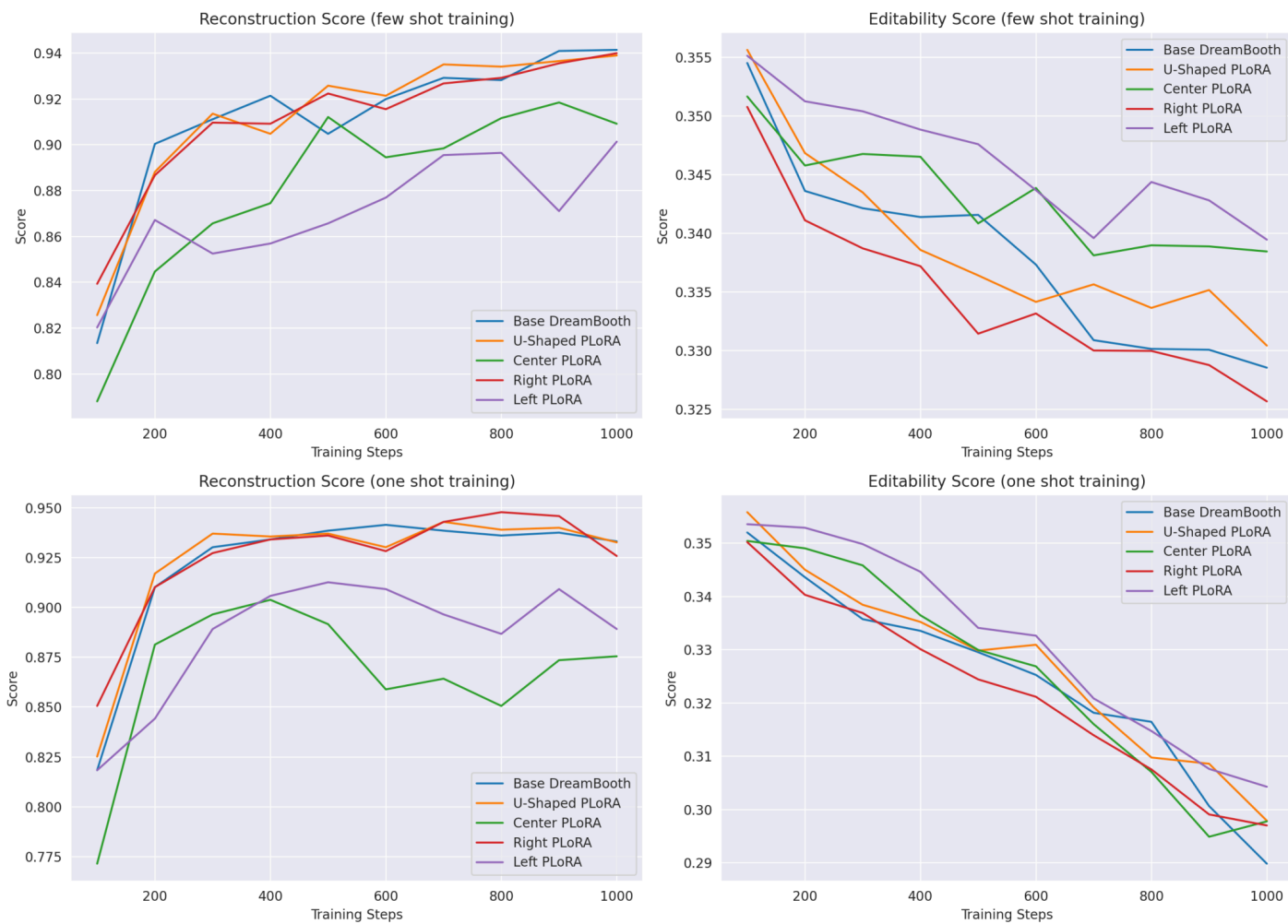
- Рассматриваем смесь двух смещений, где при $\lambda = 1$ восстанавливается Left P-LoRA, при $\lambda = 0$ - Right P-LoRA.
- Эксперименты только в режиме few-shot ($m = 5$) для трёх значений $\lambda \in \{0.25, 0.5, 0.75\}$.

$$q_{\text{mix}}(t) = \lambda q_{\text{left}}(t) + (1 - \lambda)q_{\text{right}}(t), \quad \lambda \in [0,1]$$



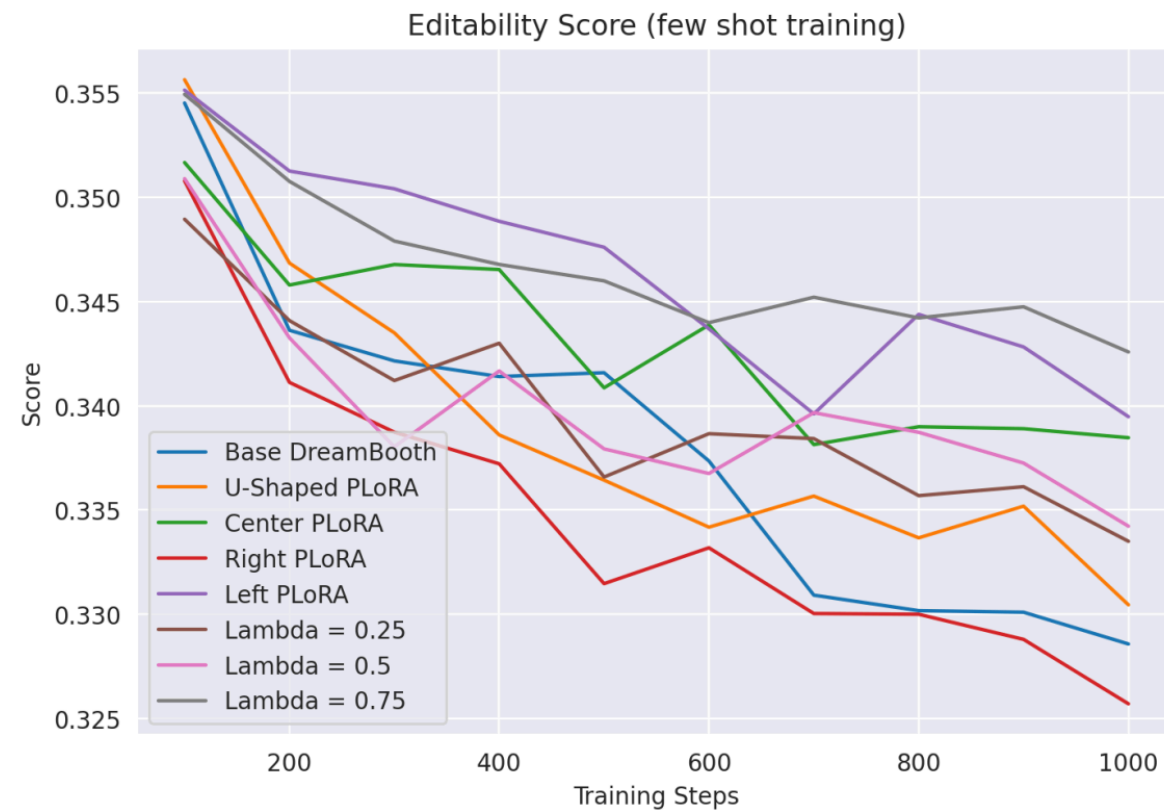
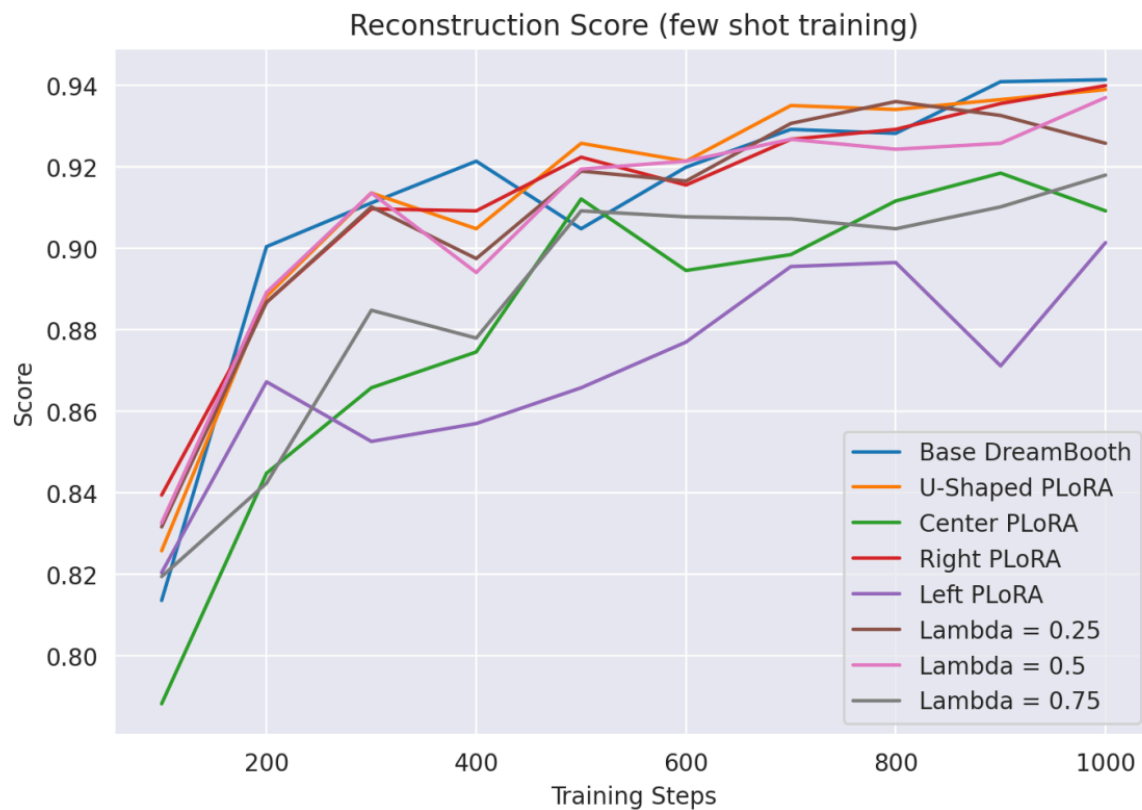


Результаты: U-Shape P-LoRA и Center P-LoRA



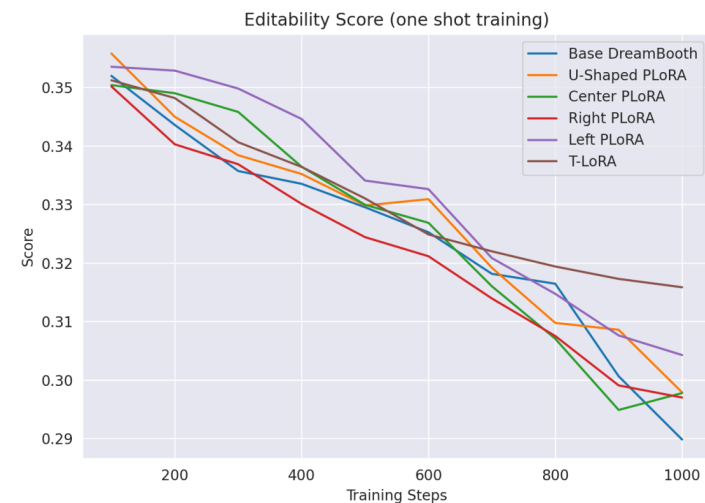
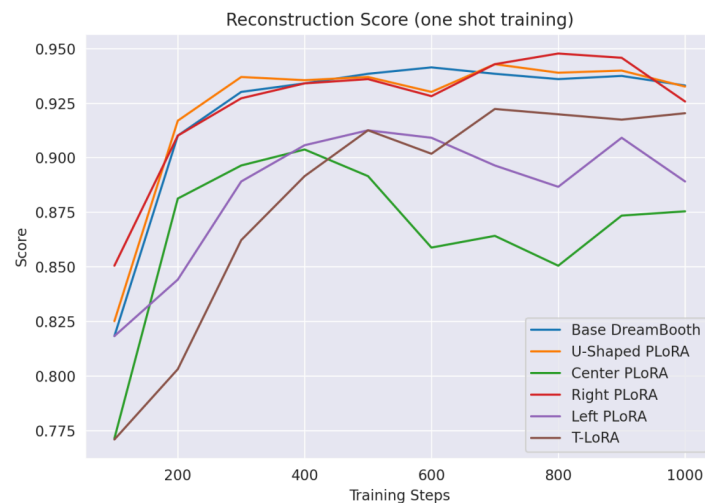
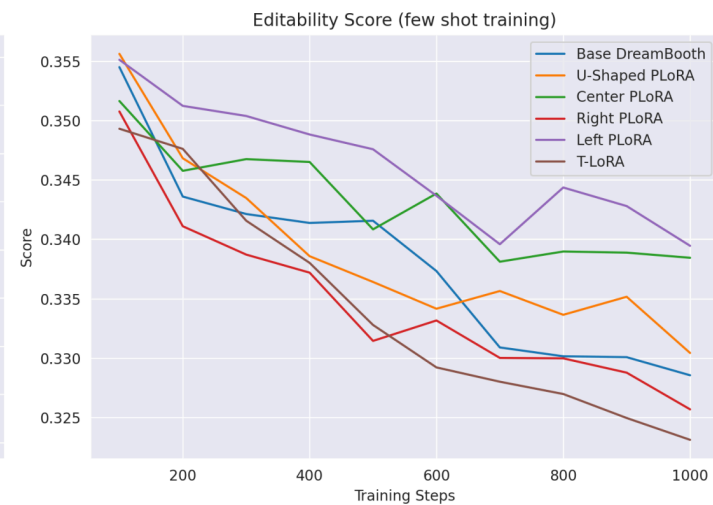


Результаты: общее сравнение





Результаты: сравнение с T-LoRA





Размытие & Насыщенность

- P-LoRA генерирует более чистый и стабильный фон.
- У DreamBooth появляются шумные текстуры и артефакты размытия.
- DreamBooth часто делает цвета слишком насыщенными.
- P-LoRA лучше сохраняет реалистичную цветовую гамму.
- Визуально P-LoRA лучше отделяет объект от фона.



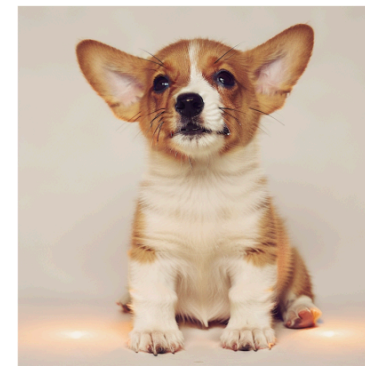
DreamBooth
One shot



Right P-LoRA
One shot



DreamBooth
One shot



Right P-LoRA
One shot





Вывод

- Перераспределение timestep является простым и гибким способом управлять компромиссом между сохранением идентичности и редактируемостью
- Низкошумовые шаги помогают лучше сохранять редактируемость и способность модели следовать новым текстовым запросам
- Высокошумовые шаги улучшают reconstruction score, так как сильнее отвечают за общую структуру объекта и сцены





Достижения

IEEE Xplore[®]

[Browse](#) ▾

[My Settings](#) ▾

[Help](#) ▾

Institutional Sign In

All ▾

Q

ADVANCED SEARCH

[Journals & Magazines](#) > [IEEE Access](#) > [Volume: 14](#) ?

Focusing the Diffusion: Importance Sampling for One-Shot Subject-Driven Image Generation

Publisher: IEEE

Cite This

PDF

More Like This

[Concealed Backdoor Attack on Diffusion Models for Smart Devices With Non-Standard](#)





Достижения



ОКН





Литература

- Ho, Jonathan; Jain, Ajay; Abbeel, Pieter. *Denoising Diffusion Probabilistic Models*. arXiv:2006.11239 (2020).
- Song, Jiaming; Meng, Chenlin; Ermon, Stefano. *Denoising Diffusion Implicit Models*. arXiv:2010.02502 (2020).
- Rombach, Robin et al. *High-Resolution Image Synthesis with Latent Diffusion Models*. arXiv:2112.10752 (2021).
- Zhang, Lvmin; Rao, Anyi; Agrawala, Maneesh. *Adding Conditional Control to Text-to-Image Diffusion Models*. arXiv:2302.05543 (2023).
- Ye, Hu et al. *IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models*. arXiv:2308.06721 (2023).
- Hu, Edward J. et al. *LoRA: Low-Rank Adaptation of Large Language Models*. arXiv:2106.09685 (2021).
- Ruiz, Nataniel et al. *DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation*. arXiv:2208.12242 (2022).
- Shpitzer, Yair; Chechik, Gal; Schwartz, Idan. *Single Image Iterative Subject-driven Generation and Editing (SISO)*. arXiv:2503.16025 (2025).
- Soboleva, Vera; Alanov, Aibek; Kuznetsov, Andrey; Sobolev, Konstantin. *T-LoRA: Single Image Diffusion Model Customization Without Overfitting*. arXiv:2507.05964 (2025).
- Gu, Yuchao et al. *Mix-of-Show: Decentralized Low-Rank Adaptation for Multi-Concept Customization of Diffusion Models*. arXiv:2305.18292 (2023).

