

دانشگاه تهران
پردیس دانشکده‌های فنی
دانشکده مهندسی برق و کامپیوتر



بررسی و مقایسه روش‌های افزودن شخصیت‌های بازتولید پذیر به مدل‌های مولد تصویر

گزارش پروژه برای دریافت درجهٔ کارشناسی در رشتهٔ مهندسی برق
گرایش کنترل

عرفان عسگری

استاد راهنما

دکتر مصطفی توسلی پور

استاد داور

دکتر رشاد حسینی

شهریور ۱۴۰۴

رَبِّ الْجَنَّاتِ وَالْجَمَارِ

چکیده

این پژوهش به بررسی چالش‌های افزودن شخصیت‌های بازتولیدپذیر در مدل‌های مولد تصویر، به ویژه مدل‌های انتشار نهفته، می‌پردازد. مسئله اصلی تحقیق، ناتوانی این مدل‌ها در حفظ ثبات و بازشناسی شخصیت‌ها در سناریوها و زمینه‌های متنوع است؛ مشکلی که کاربرد آن‌ها را در حوزه‌های مانند تصویرسازی و اینیمیشن محدود می‌سازد. در این راستا، چندین رویکرد شامل تنظیم دقیق، وارون‌سازی متن و رمزگذاری شخصیت مورد آزمون قرار گرفت.

برای ارزیابی عملکرد روش‌ها، شاخص‌هایی همچون وفاداری به بازتولید شخصیت، تطابق با متن ورودی، هزینه آموزش و استنتاج، و مقیاس‌پذیری مدنظر قرار گرفتند. سنجش کیفیت خروجی‌ها از طریق یک آزمون ترجیحی انجام شد که در آن افراد بین دو تصویر تولیدشده توسط دو روش مختلف، گزینه مطلوب‌تر را انتخاب می‌کردند. این آزمون توسط چند نفر انجام شد تا دیدگاه‌ها و سلیقه‌های مختلف در نتایج لحاظ گردد.

آزمایش‌ها در ابتدا روی چند شخصیت مختلف طراحی شدند، اما به دلیل هزینه محاسباتی بالا و ضرورت خلاصه‌سازی گزارش، نتایج نهایی بر یک شخصیت انسانی (یک بازیگر مرد) متمرکز شد. با این حال، مجموعه داده‌ای شامل تصاویر و توصیفات تصاویر ۹ شخصیت مختلف گردآوری و به صورت عمومی منتشر گردید. افزون بر این، کلیه مدل‌های آموزش‌دیده نیز برای استفاده عمومی در دسترس قرار دارند.

نتایج نشان داد که هیچ‌یک از رویکردها برتری مطلق در همه شاخص‌ها ندارند و هر کدام مزایا و محدودیت‌های خاص خود را دارا هستند. به عنوان نمونه، برخی روش‌ها وفاداری بالاتری به تصاویر مرجع داشتند اما هزینه محاسباتی و نگهداری بیشتری نیاز داشتند، در حالی که برخی دیگر توانایی بیشتری در تنوع بخشی داشتند اما ثبات کمتری ارائه کردند. یافته‌های این پژوهش می‌تواند راهنمایی برای انتخاب روش مناسب در کاربردهای خاص و نیز زمینه‌ساز پژوهش‌های آینده در توسعه شخصیت‌های بازتولیدپذیر در مدل‌های مولد تصویر باشد.

واژگان کلیدی هوش مصنوعی مولد، مدل‌های مولد عمیق، مدل‌های تولید تصویر، مدل‌های انتشار نهفته، شخصی سازی مدل‌های مولد

فهرست مطالب

۱	مقدمه	فصل ۱:
۴	مرور ادبیات	فصل ۲:
۴	۱.۲ مدل‌های مولد	
۵	۱.۱.۲ مدل‌های انتشار	
۶	۲.۱.۲ مدل‌های مولد تصویر شرطی	
۷	۳.۱.۲ مدل‌های انتشار نهفته	
۸	۲.۰.۲ افزودن مفاهیم نو به مدل‌های مولد انتشار	
۹	۱.۰.۲.۲ تنظیم دقیق کامل	
۱۰	۲.۰.۰.۲ تنظیم دقیق کم رتبه	
۱۱	۳.۰.۰.۲ وارونسازی متنی	
۱۲	۴.۰.۰.۲ رمزگذاری شخصیت	
۱۲	۱.۴.۰.۲ استخراج ویژگی‌های چهره با Buffalo-L	
۱۳	۲.۰.۰.۲ شرط تصویری با IP-Adapter	
۱۳	۳.۰.۰.۲ رمزگذاری چهره با IP-Adapter FaceID	
۱۵	روش تحقیق	فصل ۳:
۱۵	۱.۰.۳ جمع آوری داده	
۱۶	۲.۰.۳ پیاده سازی روش‌ها	
۱۶	۱.۰.۰.۳ تنظیم دقیق کامل	

۱۷	۲.۲.۳ تنظیم دقیق کمرتبه
۱۸	۳.۲.۳ وارونسازی متنه
۱۸	۴.۲.۳ رمزگذاری شخصیت
۱۹	۳.۳ معیارهای مقایسه
۱۹	۱.۳.۳ وفاداری در بازتولید شخصیت
۱۹	۲.۳.۳ تطابق با متن ورودی
۲۰	۳.۳.۳ هزینه آموزش و بهره برداری
۲۰	۴.۳.۳ مقیاس پذیری
۲۰	۴.۳ آزمون کیفیت سنجی
۲۱	۱.۴.۳ فاز اول: سنجش وفاداری به شخصیت مرجع
۲۱	۲.۴.۳ فاز دوم: سنجش تطابق با متن ورودی
۲۱	۳.۴.۳ رعایت اصول اجرایی و کنترل سوگیری

فصل ۴: نتایج

۲۲	۱.۴ انتخاب بهترین تنظیمات هر روش
۲۵	۲.۴ مقایسه کیفیت خروجی روش ها
۲۸	۳.۴ مقایسه هزینه آموزش و بهره برداری
۲۹	۴.۴ بررسی مقیاس پذیری

فصل ۵: جمع بندی

کتاب نامه

پیوست آ: تصاویر و توصیفات مورد استفاده

پیوست ب: متون ورودی مورد استفاده برای تولید تصاویر

فصل ۱

مقدمه

مدل‌های مولد تصویر در سال‌های اخیر به یکی از جریان‌سازترین فناوری‌های حوزه هوش مصنوعی تبدیل شده‌اند. این مدل‌ها با بهره‌گیری از معماری‌های پیشرفته، بهویژه مدل‌های انتشار نهفته، توانسته‌اند تصاویر با کیفیت بالا و جزئیات قابل توجه تولید کنند و بدین ترتیب جایگاه خود را در صنایع خلاق، طراحی بصری، تبلیغات، سرگرمی و حتی پژوهش‌های علمی ثبت نمایند. ظرفیت این مدل‌ها در تولید سریع و کم‌هزینه محتوای تصویری، آن‌ها را به ابزاری ارزشمند برای هنرمندان، طراحان و تولیدکنندگان محتوا بدل کرده است.

با وجود این پیشرفتهای محدودیت‌های بنیادینی وجود دارد که مانع از بهره‌برداری گسترده‌تر از این فناوری در تولیدات حرفه‌ای و آثار بلندمدت می‌شود. یکی از مهم‌ترین چالش‌ها، دشواری در معرفی شخصیت‌های دلخواه به مدل و بازتولید پایدار آن‌ها در شرایط وزمینه‌های گوناگون است. به بیان دیگر، اگرچه مدل‌های مولد می‌توانند بر اساس توضیحات متنی یا نمونه‌های تصویری، شخصیت‌هایی تازه خلق کنند، اما حفظ ثبات و یکنواختی این شخصیت‌ها در تصاویر متعدد همچنان با مشکل جدی مواجه است. این مسئله بهویژه در تولیداتی مانند انیمیشن‌ها یا کتاب‌های مصور، که نیازمند بازنمایی مکرر و قابل اعتماد یک شخصیت هستند، اهمیت دوچندان پیدا می‌کند.

علاوه بر مسئله ثبات، محدودیت‌های فرهنگی و منطقه‌ای نیز بر کارایی مدل‌های مولد تأثیر گذاشته است. اغلب این مدل‌ها بر داده‌هایی آموزش دیده‌اند که متعلق به تعداد محدودی از کشورها و فرهنگ‌های است. در نتیجه، توانایی آن‌ها در بازنمایی شخصیت‌ها، نمادها و عناصر فرهنگی سایر کشورها به‌طور محسوسی محدود است. این کاستی، کاربرد مدل‌های مولد را در زمینه‌هایی که به بازتاب هویت فرهنگی یا تولید محتوای بومی نیاز دارند،

با مشکل رو به رو می سازد.

بر همین اساس، پژوهش حاضر به بررسی مسئله بازتولید پایدار شخصیت ها در مدل های انتشار نهفته پرداخته و تلاش می کند راهکارهایی برای رفع این محدودیت ارائه دهد. هدف اصلی، توسعه و ارزیابی روش هایی است که امکان افزودن شخصیت های بازتولید پذیر به مدل های مولد را فراهم می کنند، به گونه ای که این شخصیت ها بتوانند در سناریوها، حالات و زمینه های مختلف با ثبات و وفاداری مناسب بازتولید شوند. دستیابی به چنین قابلیتی می تواند تحولی در کاربرد مدل های مولد در صنایع خلاق ایجاد کرده و آن ها را از ابزاری صرفاً آزمایشی به ابزاری عملیاتی و حرفه ای ارتقا دهد.



Amirkabir



Shahab Hosseini



Ali Daei

شکل ۱.۱: مثال هایی از عدم توانایی مدل پر استفاده SDXL در تولید شخصیت های شناخته شده ایرانی (از راست به چپ): علی دایی، شهاب حسینی، امیرکبیر

برای تحقق این هدف، چندین رویکرد موجود و پرکاربرد مورد بررسی قرار گرفته است. این رویکردها شامل تنظیم دقیق در دو حالت کم رتبه و رتبه کامل، وارون سازی متن، و رمزگذاری شخصیت هستند. برای ارزیابی دقیق تر، معیارهایی همچون وفاداری در بازتولید شخصیت، تطابق با متن ورودی، هزینه آموزش و بهره برداری و مقیاس پذیری مد نظر قرار گرفت. علاوه بر این، سنجش کیفیت خروجی ها نه صرفاً بر مبنای معیارهای محاسباتی، بلکه از طریق یک آزمون ترجیحی انسانی انجام شد. در این آزمون، چندین شرکت کننده تصاویر تولید شده توسط روش های مختلف را مقایسه کرده و بر اساس سلیقه و ادراک خود، تصویر برتر را انتخاب کردند. این طراحی به پژوهش کمک کرد تا کیفیت خروجی ها از زاویه دید کاربر نهایی نیز مورد سنجش قرار دهد.

آزمایش های انجام شده در ابتدا برای چندین شخصیت برنامه ریزی شده بود. با این حال، به دلیل هزینه محاسباتی بالا و نیاز به تمرکز در ارائه نتایج، تحلیل های اصلی بر روی یک شخصیت انسانی (یک بازیگر مرد) انجام گرفت. با وجود این، مجموعه داده ای شامل تصاویر ۹ شخصیت مختلف گردآوری شد که به صورت عمومی منتشر گردیده است. همچنین مدل های آموزش دیده در دسترس قرار گرفته اند تا امکان استفاده و بازتولید نتایج

برای سایر پژوهشگران فراهم شود. [۲، ۱]

این پژوهش در پنج فصل تنظیم شده است. در فصل دوم، مروری بر ادبیات و پژوهش‌های مرتبط ارائه می‌شود. فصل سوم به تشریح روش تحقیق می‌پردازد که شامل معیارهای مقایسه، شیوه جمع‌آوری داده، طراحی آزمون کیفیت‌سنجدی و جزئیات پیاده‌سازی هر یک از روش‌هاست. فصل چهارم نتایج به‌دست‌آمده از جمله انتخاب بهترین تنظیمات برای هر روش، مقایسه کیفیت خروجی‌ها، هزینه محاسباتی و بررسی مقیاس‌پذیری را ارائه می‌کند. در فصل پنجم، جمع‌بندی نتایج، محدودیت‌ها و مسیرهای پژوهشی آینده مطرح خواهد شد.

فصل ۲

مرور ادبیات

۱.۲ مدل‌های مولد

مدل‌های مولد^۱ دسته‌ای از الگوریتم‌های یادگیری ماشین هستند که هدف آن‌ها یادگیری توزیع داده‌های واقعی و سپس نمونه‌برداری از آن توزیع برای تولید داده‌های جدید است. به بیان دیگر، اگر $p_{data}(x)$ توزیع داده‌های واقعی باشد، مدل مولد سعی می‌کند توزیعی به شکل $p_{\theta}(x)$ بیاموزد به‌گونه‌ای که:

$$p_{\theta}(x) \approx p_{data}(x).$$

به عبارت دیگر، این مدل‌ها می‌توانند داده‌های مصنوعی بسازند که از نظر آماری مشابه داده‌های واقعی هستند. در سال‌های اخیر، مدل‌های مولد به دلیل توانایی بالای آن‌ها در تولید تصویر، متن، صدا و ویدئو اهمیت ویژه‌ای در حوزه هوش مصنوعی پیدا کرده‌اند.

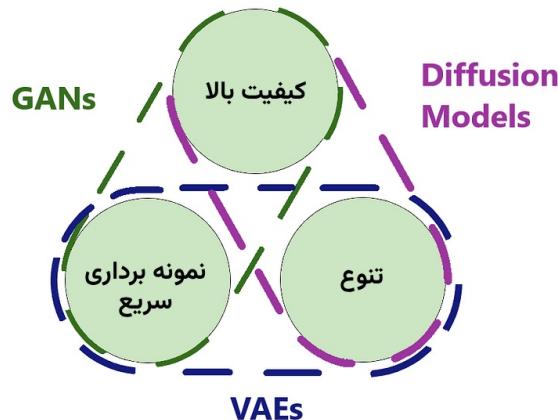
مدل‌های مولد تصویر به طور خاص برای تولید یا بازسازی تصاویر به کار می‌روند. نخستین تلاش‌های جدی در این زمینه با معرفی خودرمزنگذارهای واری (VAE)^۲ و سپس شبکه‌های مولد متخاصم (GAN)^۳ آغاز شد [۳، ۴]. این مدل‌ها توانستند تصاویر نسبتاً باکیفیتی تولید کنند، اما با محدودیت‌هایی از جمله ناپایداری در

¹Generative Models

²Variational Autoencoder

³Generative Adversarial Network

آموزش و عدم تنوع در نمونه های تولیدی (در GAN ها) و کیفیت پایین تر (در VAE ها) مواجه بودند. پیشرفت های اخیر، به ویژه با ظهور مدل های انتشار^۴، موجب جهشی قابل توجه در کیفیت تولید تصویر شده است. این مدل ها توانسته اند با بهره گیری از فرآیندهای تدریجی نویز افزایی و نویز زدایی، تصاویری با وضوح بالا و جزئیات دقیق تولید کنند [۵].



شکل ۱.۲: نقاط قوت و ضعف مدل های خود رمزگذار، متخاصل و انتشار [۶]

۱.۱.۲ مدل های انتشار

مدل های انتشار بر مبنای یک فرآیند دوبخشی تعریف می شوند: (۱) تبدیل تدریجی داده های اصلی به نویز سفید از طریق افروزن نویز گاووسی، و (۲) یادگیری یک مدل عصبی برای بازسازی داده ها از روی نویز. فرآیند انتشار رو به جلو به صورت زیر تعریف می شود:

$$q(x_t | x_{t-1}) = \mathcal{N}(x_t; \sqrt{1 - \beta_t} x_{t-1}, \beta_t I), \quad t = 1, \dots, T,$$

که در آن β_t شدت نویز در مرحله t است.

فرآیند معکوس توسط یک شبکه عصبی پارامتری شده ϵ_θ مدل سازی می شود. در عمل، این مدل سعی می کند

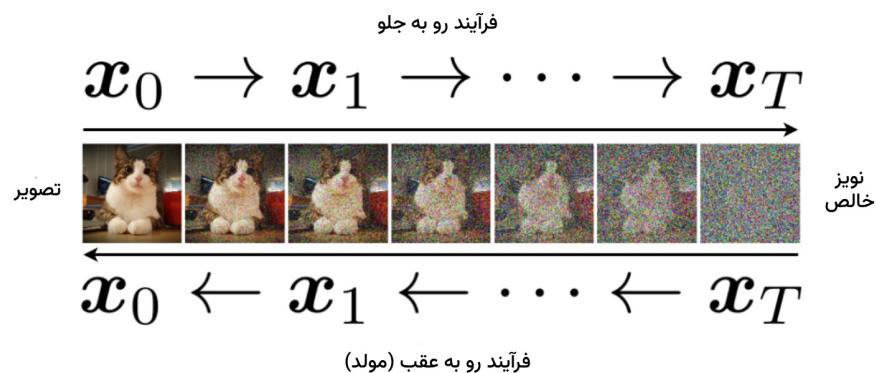
⁴Diffusion Models

نویز افزوده شده در هر مرحله را پیش بینی کند و با کمینه سازی تابع خط آموزش می بیند:

$$L_{simple} = \mathbb{E}_{x_0, \epsilon, t} \left[\|\epsilon - \epsilon_\theta(x_t, t)\|^2 \right].$$

مدل های انتشار اولیه تحت عنوان DDPM^۵ معرفی شدند [۵]. به منظور افزایش سرعت نمونه برداری، نسخه های بهبود یافته ای مانند DDIM^۶ ارائه شد که به جای فرآیند تصادفی از نگاشت های قطعی برای بازسازی استفاده می کنند. [۷]

یکی از مهم ترین مزایای مدل های انتشار، پایداری در آموزش و توانایی تولید تصاویر متعدد و سازگار با ورودی های متنی است [۸].



شکل ۲.۲: نمای کلی فرآیند نویز افزایی و نویز زدایی در مدل های انتشار [۹]

۲.۱.۲ مدل های مولد تصویر شرطی

یکی از توسعه های مهم در مدل های مولد، افزودن شرط^۷ به فرآیند تولید داده است. در این حالت، مدل علاوه بر نمونه برداری از توزیع داده، از یک سیگنال خارجی برای هدایت فرآیند تولید استفاده می کند. این رویکرد با عنوان مدل های مولد شرطی^۸ شناخته می شود.

⁵Denoising Diffusion Probabilistic Model

⁶Denoising Diffusion Implicit Model

⁷Condition

⁸Conditional Generative Models

به طور رسمی، اگر \mathcal{U} متغیر شرطی (مانند متن، برچسب کلاس یا یک تصویر دیگر) باشد، هدف مدل، تقریب توزیع شرطی $(x | y)$ است. برای مثال در حالت متن-به-تصویر:

$$x \sim p_{\theta}(x | y), \quad y = \text{prompt}$$

در این چارچوب، «متن ورودی»^۹ عبارتی است که کاربر به عنوان ورودی ارائه می‌دهد و مدل تصویری متناسب با آن تولید می‌کند.

۳.۱.۲ مدل های انتشار نهفته

مدل های انتشار نهفته (LDM)^{۱۰} رویکردی نوین در بهینه سازی مدل های انتشار هستند که با فشرده سازی داده ها در یک فضای نهفته^{۱۱} یادگیری را انجام می‌دهند. این مدل ها ابتدا یک خودر مزگذار واری آموزش می‌دهند تا نگاشت داده های تصویری $x \in \mathbb{R}^{H \times W \times 3}$ به یک بازنمایی فشرده $z \in \mathbb{R}^{h \times w \times c}$ و بازسازی معکوس آن را ممکن سازد:

$$z = E(x), \quad \hat{x} = D(z),$$

که در آن E رمزگذار^{۱۲} و D رمزگشا^{۱۳} هستند.

پس از یادگیری این نگاشت، فرآیند انتشار نه در فضای تصویر، بلکه در فضای نهفته z انجام می‌شود. بدین ترتیب، فرآیند افزودن نویز رو به جلو به شکل زیر تعریف می‌شود:

$$q(z_t | z_{t-1}) = \mathcal{N}(z_t; \sqrt{1 - \beta_t} z_{t-1}, \beta_t I), \quad t = 1, \dots, T$$

و مدل معکوس شرطی $(z_t | z_{t-1})$ در این فضای فشرده آموزش داده می‌شود.
مزیت اصلی این رویکرد آن است که ابعاد فضای نهفته به مراتب کوچکتر از فضای پیکسلی است ($h \ll$)

⁹prompt

¹⁰Latent Diffusion Models

¹¹Latent Space

¹²Encoder

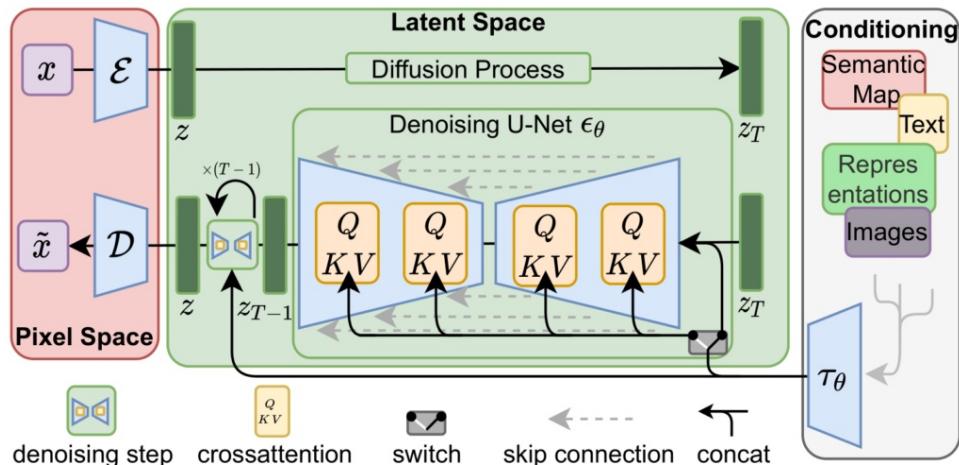
¹³Decoder

$w \ll H$) و بنابراین هزینه محاسباتی فرآیند انتشار و بازسازی بسیار کاهش می‌یابد، در حالی‌که کیفیت بازسازی با کمک رمزگشا حفظ می‌شود.

پس از اتمام فرآیند نمونه‌برداری در فضای نهفته، تصویر نهایی از طریق رمزگشا بازسازی می‌شود:

$$\hat{x}_o = D(z_o), \quad z_o \sim p_\theta(z)$$

یکی از شناخته‌شده‌ترین نمونه‌های این دسته، Stable Diffusion است که با بهره‌گیری از مدل انتشار در فضای نهفته توانست فرآیند تولید تصویر متن‌به‌تصویر را به شکل کارآمد و مقیاس‌پذیر در دسترس عموم قرار دهد .[۱۰]



شکل ۳.۲: نمای کلی معماری یک مدل انتشار نهفته (LDM)

۲.۲ افزودن مفاهیم نو به مدل‌های مولد انتشار

یکی از چالش‌های اصلی در مدل‌های مولد تصویر، امکان افزودن سریع و هدفمند مفاهیم جدید بدون آموزش مجدد کل مدل است. در طول سال‌های اخیر، روش‌های متعددی برای این منظور معرفی شده‌اند که می‌توان آن‌ها را به دو دسته اصلی تقسیم کرد:

۱. روش های نیازمند آموزش^{۱۴}۲. روش های بدون نیاز به آموزش^{۱۵}

در ادامه مثال هایی از روش های نیازمند آموزش (۱ تا ۳) و سپس مثالی از روش های بدون نیاز به آموزش آورده شده است.

۱۰۲۰۲ تنظیم دقیق کامل

تنظیم دقیق کامل به معنی ادامه دادن آموزش تمامی وزن های یک مدل به وسیله داده های جدید است. یکی از روش های تنظیم دقیق کامل در کاربردهای متن-به-تصویر تحت عنوان DreamBooth شناخته می شود و روشی برای افزودن مفاهیم یا شخصیت های جدید به یک مدل مولد می باشد [۱۱].

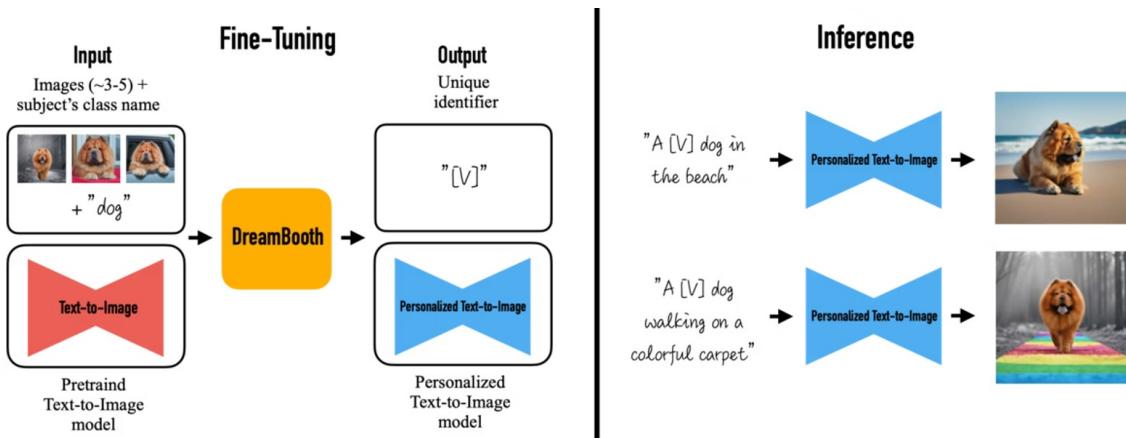
ایده اصلی این روش آن است که یک بردار نهفته جدید $v_{trigger}$ در فضای متنی مدل ایجاد شود که نشان دهنده مفهوم یا شخصیت جدید باشد. سپس این بردار با یک کلمه کلاس پایه c ترکیب می شود تا مدل بتواند مفهوم جدید را در زمینه های مختلف تولید تصویر بازشناسی کند.

مدل سپس با نمونه های محدود از تصاویر واقعی x_i آموزش داده می شود تا خروجی های تولیدی x'_i بیشینه شباهت بصری با نمونه های واقعی داشته باشند:

$$\theta^* = \arg \min_{\theta} \sum_{i=1}^N \mathcal{L}_{recon}(x_i, G_{\theta}(z_{text,i})),$$

که در آن G_{θ} مدل مولد اصلی و \mathcal{L}_{recon} تابع خطای بازسازی یا loss تصویری است. مزیت اصلی DreamBooth آن است که می توان با تعداد کمی تصویر نمونه، مفاهیم خاص و منحصر به فرد را به مدل اضافه کرد و در عین حال سبک و کیفیت تصویر اصلی حفظ می شود. این روش به طور گسترده برای شخصی سازی شخصیت ها، افزودن سبک های هنری یا اشیاء خاص به مدل های متن-به-تصویر مانند Stable Diffusion استفاده می شود.

¹⁴Training-Based Methods¹⁵Training-Free Methods



شکل ۴.۲: نمای شماتیک روش DreamBooth: واژه محرك^{۱۷} جدید به همراه واژه کلاس پایه به مدل معروفی می‌شود و مدل با نمونه‌های محدود تصاویر مفهوم جدید را یاد می‌گیرد [۱۱].

این روش بیشترین انعطاف‌پذیری را دارد و می‌تواند مفاهیم پیچیده و متفاوت را به خوبی یاد بگیرد، اما هزینه محاسباتی بسیار بالایی دارد و پس از آموزش، یک مدل کاملاً جدید بدست می‌آید که نگهداری از آن، نیازمند فضایی معادل مدل اصلی می‌باشد.

۲.۲.۲ تنظیم دقیق کم‌رتبه

روش تطبیق کم‌رتبه (LoRA)^{۱۸} به منظور افزودن مفاهیم جدید به مدل‌های مولد تصویر بدون نیاز به آموزش مجدد تمامی پارامترهای مدل، معرفی شده است [۱۲].

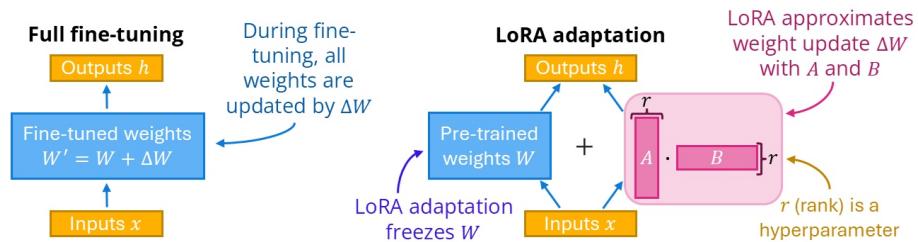
در این روش، تغییرات وزن‌ها به صورت ترکیب دو ماتریس کم‌رتبه A و B مدل می‌شود:

$$W' = W + \alpha \cdot A \cdot B^T$$

$$W' = W + \Delta W$$

که در آن W وزن‌های اصلی مدل، α ضریب مقیاس، و A و B ماتریس‌های کم‌رتبه هستند. این رویکرد باعث کاهش تعداد پارامترهای قابل آموزش می‌شود و در نتیجه، سرعت یادگیری و کارایی مدل افزایش می‌یابد. یکی از روش‌های پیشرفته در این زمینه، DiffLoRA است که از مدل‌های انتشار به عنوان یک هایپرنت

¹⁸Low-Rank Adaptation



شکل ۵.۲: نمای شماتیک روش LoRA در مقایسه با تنظیم دقیق کامل: تغییرات وزن‌ها به صورت ترکیب دو ماتریس کم‌رتبه مدل می‌شود [۱۳].

برای پیش‌بینی وزن‌های LoRA شخصی‌سازی‌شده استفاده می‌کند. در این روش، به جای آموزش مجدد مدل، وزن‌های LoRA در زمان استنتاج تولید می‌شوند، که باعث افزایش کارایی و حفظ وفاداری هویت در تصاویر تولیدی می‌شود [۱۴].

این روش‌ها به‌ویژه در کاربردهایی مانند تولید تصاویر با هویت‌های مشخص یا سبک‌های هنری خاص، بدون نیاز به آموزش مجدد مدل‌های بزرگ، مورد استفاده قرار می‌گیرند.

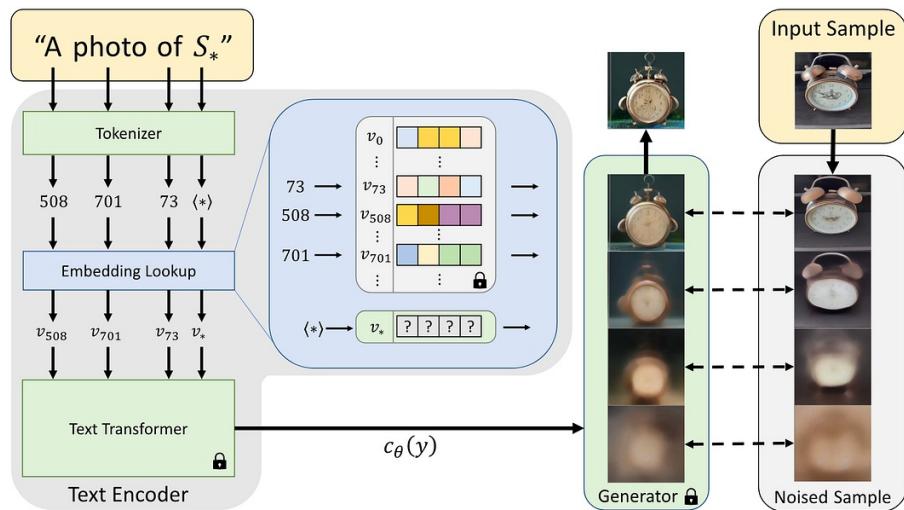
مزیت اصلی این روش، تعداد پارامترهای قابل آموزش کمتر و همچنین فضای کمتر برای ذخیره سازی مدل آموزش دیده شده جدا از مدل پایه می‌باشد.

۳۰.۲.۲ وارونسازی متنی

وارونسازی متنی^{۱۹} روشی است برای افروzen مفاهیم جدید به مدل‌های متن-به-تصویر بدون نیاز به آموزش مجدد وزن‌های شبکه اصلی. ایده اصلی این روش، یادگیری بردارهایی برای یک توکن مفهومی جدید $v_{concept}$ در فضای نهفته متن است که نشان‌دهنده مفهوم، شخصیت یا شیء خاصی می‌باشد. بنابراین تمامی وزن‌های شبکه در این روش دست نخورده باقی می‌مانند و تغییر نتیجه خروجی توسط یادگیری ورودی مناسب حاصل می‌گردد [۱۵].

مزیت اصلی این روش آن است که با چند تصویر نمونه محدود، می‌توان مفاهیم خاص را به مدل اضافه کرد و در عین حال تمام پارامترهای شبکه اصلی ثابت می‌مانند. این امر باعث می‌شود یادگیری سریع، کم‌هزینه و سازگار با مدل‌های بزرگ متن-به-تصویر مانند Stable Diffusion باشد. همچنین فضای ذخیره سازی برای مفهوم جدید بسیار کم و به اندازه ابعاد بردارهای آموزش دیده است.

¹⁹Textual Inversion



شکل ۶.۲: نمای شماتیک روش وارونسازی متنه: تعدادی بردار برای یک توکن مفهومی جدید در فضای نهفته متن آموزش داده می‌شود و مدل مولد تصویر متناسب با آن تولید می‌کند [۱۵].

۴.۰.۲ رمزگذاری شخصیت

رمزگذاری شخصیت یکی از روش‌های کلیدی برای افزودن یا حفظ هویت شخصیت‌ها در مدل‌های مولد تصویر است. در این رویکرد که بدون نیاز به آموزش است، شخصیت‌ها در زمان استنتاج رمزگذاری شده و به عنوان شرط ورودی به مدل داده می‌شوند. یکی از مدل‌هایی که این روش را پیاده‌سازی کرده است، IP-Adapter می‌باشد که در این بخش با دو مقدمه توضیح داده شده است.

۱.۴.۰.۲ استخراج ویژگی‌های چهره با Buffalo-L

مدل L Buffalo یک شبکه عصبی پیشرفته است که برای شناسایی و استخراج ویژگی‌های هویتی چهره طراحی شده است. این مدل با استفاده از مجموعه داده‌های متنوع چهره و هدف یادگیری نمایندگی‌های نهفته آموزش دیده است تا بتواند ویژگی‌های هویتی را در شرایط مختلف نور، زاویه دید و حالات چهره استخراج کند. فرآیند معمولاً شامل یادگیری نگاشت تصویر چهره x_{face} به یک بردار نهفته $z_{id} \in \mathbb{R}^d$ است:

$$z_{id} = F_B(x_{face}),$$

که در آن F_B شبکه آموزش دیده مدل است. بردار نهفته z_{id} نماینده ویژگی های هویتی چهره بوده و می تواند در کاربردهایی مانند تشخیص چهره به کار گرفته شود.

۲.۴.۲.۲ شرط تصویری با IP-Adapter

مدل IP-Adapter^{۲۰} یک افزونه سبک برای مدل های مولد تصویر متن-به-تصویر است که امکان استفاده از ورودی های تصویری را به مدل می دهد [۱۶]. ایده اصلی این روش آن است که ویژگی های استخراج شده از تصویر ورودی x_{img} به فضای نهفته مدل اضافه شود و با ورودی متنی z_{text} ترکیب گردد تا تولید تصویر هدایت شود:

$$z_{text}^{adapted} = f_{IP}(z_{text}, E_{img}(x_{img}))$$

که در آن E_{img} رمزگذار تصویر و f_{IP} نگاشت تطبیق یافته IP-Adapter است. این مدل از یک مکانیزم توجه متقاطع جداگانه^{۲۱} استفاده می کند تا ویژگی های متنی و تصویری به طور همزمان در مدل مولد اعمال شوند. با این روش، نیازی به آموزش مجدد وزن های کل مدل اصلی نیست و تنها افزونه-IP Adapter آموزش می بیند که دارای حدود ۲۲ میلیون پارامتر است.

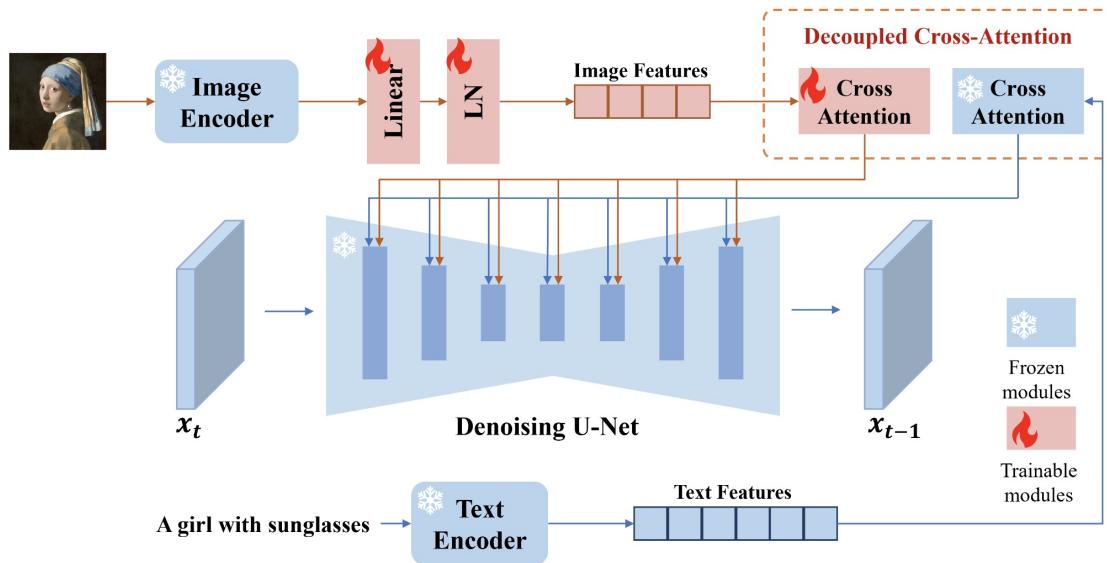
استفاده از IP-Adapter به ویژه برای کاربردهایی که نیاز به ترکیب اطلاعات تصویری و متنی دارند مناسب است و امکان انتقال مفاهیم از تصاویری که توصیف آنها به صورت دقیق مشکل است به تصویر تولید شده را فراهم می کند.

۳.۴.۲.۲ رمزگذاری چهره با IP-Adapter FaceID

مدل IP-Adapter FaceID نسخه ای پیشرفته از IP-Adapter است که مستقیماً از ویژگی های استخراج شده توسط Buffalo-L استفاده می کند و آن ها را در زمان استنتاج به مدل مولد تصویر وارد می کند [۱۷]. در این روش، بردار ویژگی هویتی z_{id} که نماینده چهره است، با پرامپت متنی z_{text} ترکیب می شود تا تصویر خروجی هم

²⁰Image Prompt Adapter

²¹Decoupled Cross-Attention



[۱۶] نمای شماتیک مدل IP-Adapter

هویت واقعی شخصیت و هم سبک یا محتواهای متن ورودی را حفظ کند:

$$x \sim G_\theta(z_{text}, z_{id})$$

که در آن G_θ مدل مولد اصلی است و x تصویر تولید شده با ویژگی های هویتی z_{id} می باشد.
مزیت اصلی IP-Adapter FaceID آن است که بدون نیاز به آموزش مجدد مدل مولد تصویر، می توان شخصیت های جدید را به مدل اضافه کرد و هویت آنها را در تصاویر تولیدی حفظ نمود. این رویکرد به ویژه برای کاربردهای شخصی سازی شخصیت ها، تولید تصاویر واقعی از افراد با سبک های مختلف و انتقال ویژگی های دقیق چهره مفید است.

فصل ۳

روش تحقیق

۱.۳ جمع آوری داده

برای آموزش روش هایی که نیازمند آموزش هستند و همچنین روش هایی که نیاز به تصاویر مرجع برای بازتولید شخصیت دارند، مجموعه داده ای از چندین شخصیت حقیقی (مدل ها و افراد مشهور) و شخصیت های غیرحقیقی (انیمیشنی و عروسکی) جمع آوری شد. تعداد تصاویر هر یک از این شخصیت ها متفاوت می باشد و جزئیات آن در جدول ۱.۳ آمده است [۱].

جدول ۱.۳: مشخصات مجموعه داده

تعداد تصاویر	شخصیت	دسته بندی
۱۴	داریوش فرضیایی	بازیگر مرد
۳۸	شهاب حسینی	بازیگر مرد
۱۶	میمون چی توز	شخصیت برند
۱۸	جناب خان	شخصیت عروسکی
۱۴	پسر دلفینی	شخصیت انیمیشنی مذکور (سه بعدی)
۱۸	وینگی	شخصیت کارتونی مونت
۲۱	رستم	شخصیت اساطیری
۱۶	لیتو	مدل مرد
۱۵	مگان	مدل مرد

پس از جمع آوری داده‌ها، از هوش مصنوعی Gemini برای نوشتن توصیف تصاویر استفاده شد. همچنین برای هر شخصیت تعدادی متن ورودی توسط Gemini تولید شد تا در مرحله تولید نمونه‌ها استفاده گردد. این متون و دستورات مورد استفاده برای تولید آنها در پیوست آ و ب آمده است.

در ادامه ۲ مجموعه دیگر تحت عنوان تصاویر منظم سازی جمع آوری شد که شرح آن در جدول ۲.۳ آمده است.

جدول ۲.۳: مشخصات مجموعه داده

عنوان	کلمه کلیدی	تعداد تصاویر
تصاویر مردان	man	۱۰۰
شخصیت های انیمیشنی	character	۴۱

۲.۳ پیاده سازی روش ها

سه روش نیازمند به آموزش با کمک ریپازیتوری Kohya_ss آموزش داده شدند که یک محیط جامع برای آموزش این سه روش می‌باشد [۱۸]. در ادامه جزئیات و پارامترهای در نظر گرفته شده برای هر روش آمده است و در بخش آخر نحوه خروجی گرفتن از روش های بدون نیاز به آموزش توضیح داده شده است.

۱.۲.۳ تنظیم دقیق کامل

برای آموزش این روش ۲ تنظیمات مختلف را در نظر گرفتیم تا تاثیر حضور تصاویر منظم سازی را بررسی کنیم. در آموزش این روش، فقط پارامترهای Unet آموزش دیدند و بخش Text encoders به صورت فریز شده در نظر گرفته شد.

جدول ۳.۳: مشخصات مدل های آموزش داده شد در روش تنظیم دقیق کامل

نام مدل	تصاویر منظم سازی
db1	✓
db2	✗

این روش با استفاده از واحد محاسبات گرافیکی Nvidia A40 آموزش داده شد و واژه ohwx به عنوان واژه محرك و man به عنوان واژه کلاس پایه در نظر گرفته شد. سایر پارامترهای مهم آموزش در جدول ۴.۳ آمده است.

جدول ۴.۳: پارامترهای بکار رفته در روش تنظیم دقیق کامل

پارامتر	قدار	توضیحات
base_model	Stable Diffusion XL [19]	
unet_learning_rate	1e-5	نرخ یادگیری
optimizer	adafactor	
gradient_checkpointing	false	جهت تسریع آموزش
cross_attention	xformers	جهت کاهش VRAM

۲.۲.۳ تنظیم دقیق کم رتبه

برای آموزش این روش ۴ تنظیمات مختلف را در نظر گرفتیم تا تاثیر رتبه و حضور تصاویر منظم سازی را بررسی کنیم.

جدول ۵.۳: مشخصات مدل های آموزش داده شد در روش تنظیم دقیق کم رتبه

نام مدل	رتبه	تصاویر منظم سازی
lora0	۶۴	X
lora1	۱۲۸	X
lora2	۶۴	✓
lora3	۱۲۸	✓

این روش با استفاده از واحد محاسبات گرافیکی Nvidia RTX A5000 آموزش داده شد و واژه ohwx به عنوان واژه محرك و واژه man به عنوان واژه کلاس پایه در نظر گرفته شد. سایر پارامترهای مهم آموزش در جدول ۶.۳ آمده است.

جدول ۶.۳: پارامترهای بکار رفته در روش تنظیم دقیق کمرته

توضیحات	مقدار	پارامتر
	Stable Diffusion XL	base_model
نرخ یادگیری	4e-5	unet_learning_rate
	adafactor	optimizer
جهت کاهش VRAM	true	gradient_checkpointing

۳.۲.۳ وارونسازی متنی

برای آموزش این روش از توصیفات متنی و تصاویر منظم سازی استفاده نشده است. همچنین بعد از تلاش های ناموفق برای آموزش این روش در مدل SDXL، از مدل Stable Diffusion v1.5 استفاده گردید. این روش با استفاده از واحد محاسبات گرافیکی Nvidia RTX A5000 آموزش داده شد. بردارهای مورد آموزش توسط کلمه man مقداردهی اولیه شدند.

سایر پارامترهای مهم آموزش در جدول ۷.۳ آمده است.

جدول ۷.۳: پارامترهای بکار رفته در روش وارونسازی متنی

توضیحات	مقدار	پارامتر
	Stable Diffusion v1.5	base_model
نرخ یادگیری	5e-5	learning_rate
تعداد بردارهای مورد آموزش	6	vectors_per_token

۴.۲.۳ رمزگذاری شخصیت

برای پیاده سازی این روش از مدل IP-Adapter FaceID SDXL استفاده گردید. در این روش برای استخراج ویژگی های چهره از مدل Buffalo-L استفاده می گردد. در پیاده سازی این روش دو تنظیمات مختلف به کار گرفته شد. در حالت اول تنها یک تصویر مرجع به عنوان ورودی استخراج کننده ویژگی در نظر گرفته شد و بردار ویژگی آن به مدل IP-Adapter داده شد. در حالت دوم بردار ویژگی حاصل از ۵ تصویر مرجع میانگین گرفته شد و سپس به مدل IP-Adapter وارد گردید.

جدول ۸.۳: حالت های آزمایش شده در روش رمزگذاری شخصیت

تصاویر مرجع به کار رفته	نام حالت
7	ref7
9	ref9
7, 9, 12, 4, 5	mean

۳.۳ معیارهای مقایسه

در این پژوهش به منظور ارزیابی و مقایسه روش های مورد بررسی، چهار معیار اصلی در نظر گرفته شده است. این معیارها جنبه های مختلف کیفیت بازتولید شخصیت، میزان انطباق با ورودی متنی، هزینه های محاسباتی و نیز مقیاس پذیری روش ها را پوشش می دهند. در ادامه، هر یک از این معیارها معرفی شده اند.

۱.۳.۳ وفاداری در بازتولید شخصیت

یکی از مهم ترین معیارها، میزان وفاداری^۱ روش به تصاویر مرجع و حفظ ویژگی های ظاهری شخصیت در بازتولید آن است. برای کمی سازی این معیار، در آزمون کیفیت سنجی از داوطلبان خواسته شد تا میزان شباهت تصاویر ساخته شده با تصاویر مرجع شخصیت ها را ارزیابی و رتبه بندی کنند.

۲.۳.۳ تطابق با متن ورودی

دومین معیار، تطابق با متن ورودی^۲ یا توانایی مدل در بازتولید شخصیت در شرایط، زمینه ها و سبک هایی است که در تصاویر مرجع وجود نداشته است. برای اندازه گیری این معیار، در آزمون کیفیت سنجی از داوطلبان خواسته شد تا میزان انطباق تصاویر تولید شده با پرامپت های ورودی مربوط به آن ها را ارزیابی کنند.

¹Reproduction Fidelity

²Prompt Adherence

۳.۳.۳ هزینه آموزش و بهره برداری

معیار سوم، هزینه آموزش و بهره برداری است که به مقایسه هزینه های زمانی و محاسباتی روش های مختلف می پردازد. برای این منظور، بسته به ماهیت هر روش، شاخص های زیر در فرآیند آزمایش ثبت و مقایسه شدند:

- حجم VRAM اشغالی در زمان آموزش
- مدت زمان تقریبی آموزش
- حافظه دیسک مورد نیاز برای نگهداری مدل های آموزش دیده

۴.۳.۳ مقیاس پذیری

چهارمین معیار، مقیاس پذیری^۳ روش ها است. این معیار ظرفیت روش ها در مواجهه با افزایش تعداد شخصیت ها را بررسی می کند. برای ارزیابی آن، تغییرات هزینه های محاسباتی و نیز افت احتمالی در میزان وفاداری با افزایش تعداد شخصیت های آموزش داده شده مورد توجه قرار گرفت.

۴.۳ آزمون کیفیت سنجی

برای ارزیابی کیفی خروجی های روش های مورد بررسی، یک آزمون ترجیحی زوجی طراحی و اجرا شد. نکته کلیدی در طراحی این آزمون آن است که در هر تکرار دو تصویر خروجی که هر یک متعلق به یک روش متفاوت هستند به صورت زوجی به شرکت کننده نشان داده می شوند و از او خواسته می شود بین آن دو یکی را انتخاب کند یا در صورت عدم تطابق، گزینه «هیچ کدام» را برگزیند. این فرمت در هر دو فاز آزمون به کار رفت تا مقایسه مستقیم و عینی بین روش ها میسر شود.

³Scalability

۱.۴.۳ فاز اول: سنجش وفاداری به شخصیت مرجع

در این فاز، هدف سنجش توانایی مدل‌ها در بازتولید هویت بصری شخصیت مرجع بود. برای هر پرسش یک مرجع (نام یا تصویر مرجع شخصیت) و دو تصویر تولیدشده توسط دو روش مختلف نشان داده شد. سپس از شرکت‌کننده پرسیده شد کدامیک از دو تصویر شباهت یا وفاداری بیشتری به شخصیت مرجع دارد؛ در صورت عدم تشخیص یا عدم تطابق، گزینه «هیچ‌کدام» قابل انتخاب است. ترتیب نمایش تصاویر به صورت تصادفی بود و تنها انتخاب شرکت‌کننده ذخیره می‌شد.

۲.۴.۳ فاز دوم: سنجش تطابق با متن ورودی

در این فاز، هدف اندازه‌گیری میزان پایبندی تصویر خروجی به دستور متنی است. برای هر پرسش، متن ورودی همراه با دو تصویر (از دو روش مختلف) به شرکت‌کننده نمایش داده شد. از شرکت‌کننده خواسته شد تعیین کند کدامیک از دو تصویر بیشترین تطابق را با متن ورودی دارد؛ در صورت عدم تطابق قابل قبول، می‌تواند «هیچ‌کدام» را انتخاب کند. پاسخ ثبت‌شده برای هر پرسش صرفاً انتخاب شرکت‌کننده (اول/دوم/هیچ‌کدام) است.

۳.۴.۳ رعایت اصول اجرایی و کنترل سوگیری

برای کاهش سوگیری و افزایش اعتبار نتایج، رویه‌های زیر اعمال شد:

- نام یا مشخصه روش‌ها از دید شرکت‌کننگان پنهان شد.
- ترتیب نمایش تصاویر برای هر پرسش تصادفی سازی شد.
- پیش از آزمون نمونه‌های تمرینی و دستورالعمل روش‌ن ارائه شد.

فصل ۴

نتایج

مدل های آموزش داده شده در تنظیمات مختلف با استفاده از مشاهده خروجی ها مقایسه گردیدند تا بهترین مدل برای هر روش انتخاب گردد. در ادامه بهترین مدل از هر روش با روش های دیگر به رقابت گذاشته شده اند.

۱۰.۴ انتخاب بهترین تنظیمات هر روش

در این بخش از متون ورودی زیر برای نمایش عملکرد مدل های آموزش دیده استفاده شده است که کامل آن در پیوست آمده است:

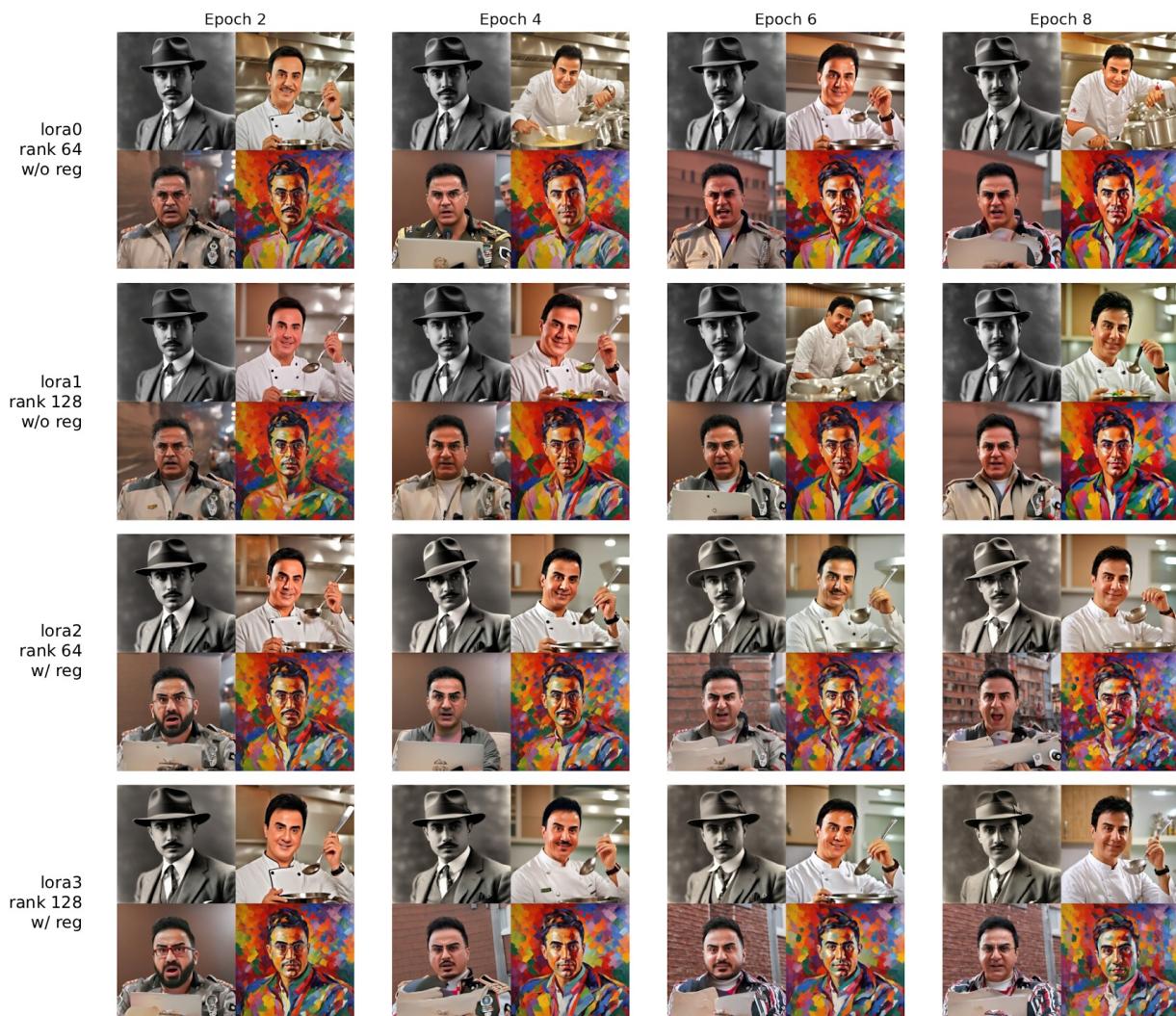
07	02
19	21



شکل ۱.۴: مقایسه خروجی تنظیمات مختلف روش تنظیم دقیق کامل در ایپاک های مختلف

با مشاهده نتایج بدست آمده در روش تنظیم دقیق کامل، مدل db2-epoch4 انتخاب گردید که بدون استفاده از تصاویر منظم سازی آموزش دیده است.

فصل ۴: نتایج



شکل ۲.۴: مقایسه خروجی تنظیمات مختلف روش تنظیم دقیق کم رتبه در ایپاک های مختلف

با مشاهده نتایج بدست آمده در روش تنظیم دقیق کم رتبه، مدل lora0-epoch8 انتخاب گردید که دارای رتبه ۶۴ بوده و بدون استفاده از تصاویر منظم سازی آموزش دیده است.

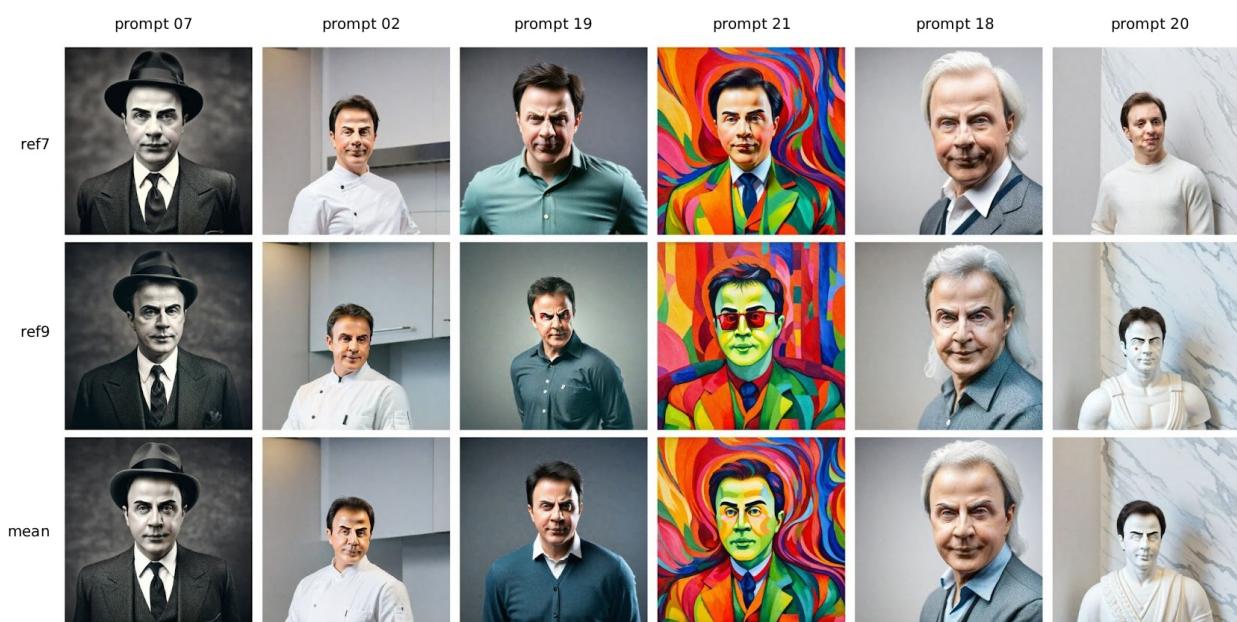
فصل ۴: نتایج

بررسی و مقایسه روش های افروzen شخصیت های بازتولید پذیر به مدل های مولد تصویر



شکل ۴.۳: مقایسه خروجی روش وارونسازی متنی در قدم های مختلف

با مشاهده نتایج بدست آمده در روش وارونسازی متن، مدل ti0-step14000 انتخاب گردید.



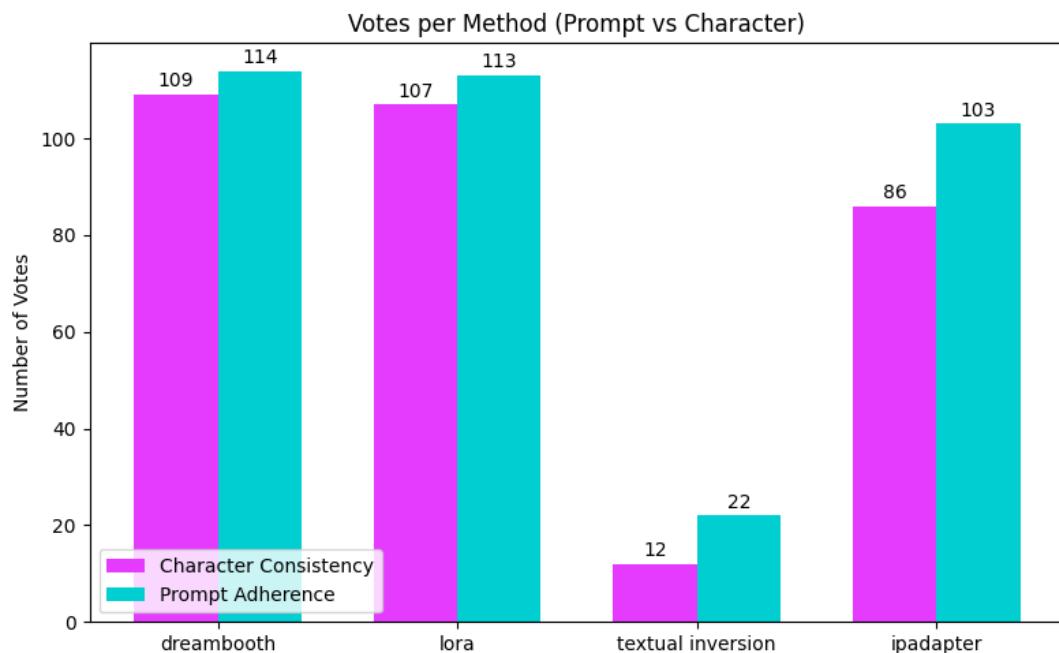
شکل ۴.۴: مقایسه خروجی روش رمزگذاری شخصیت در سه تنظیمات مختلف برای تصویر مرجع

با مشاهده نتایج بدست آمده در روش رمزگذاری شخصیت، بنظر می‌رسد بهره گیری از میانگین بردار تعییه چندین تصویر مرجع (mean) عملکرد بهتری داشته است.

۲.۴ مقایسه کیفیت خروجی روش‌ها

در شکل ۶.۴ مقایسه‌ای بین تصاویر تولید شده توسط ۴ روش بررسی شده در این پژوهش با ۷ متن ورودی مختلف آمده است. در آزمون کیفیت سنجی، خروجی های مربوط به ۲۳ متن ورودی مختلف برای هر روش

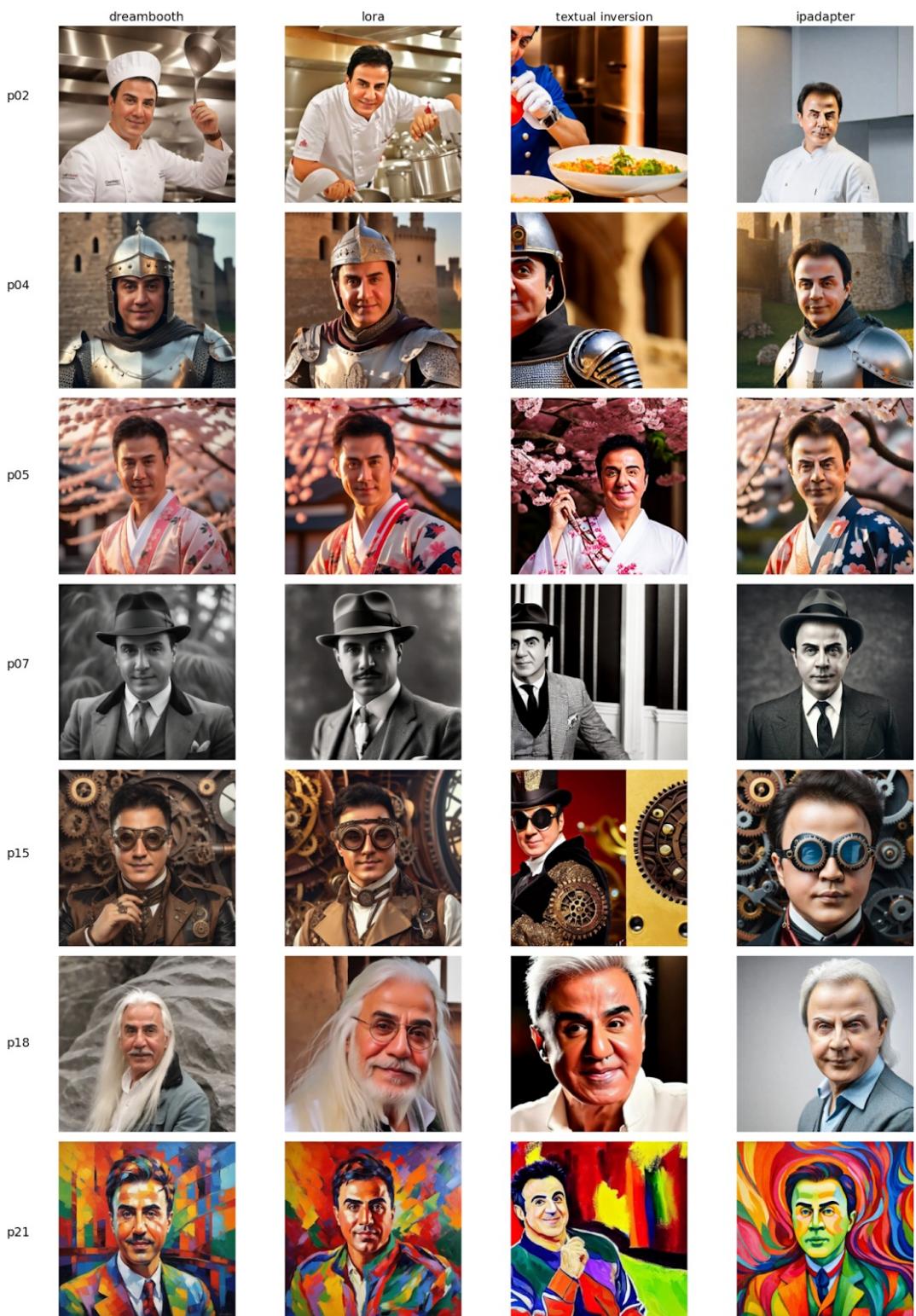
مورد رقابت قرار گرفته اند.



شکل ۴.۵: نتایج آزمون کیفیت سنجی از ۳ شرکت کننده: مقایسه دو معیار وفاداری به بازتولید شخصیت (آبی) و تبعیت از متن ورودی (صورتی) در میان روش های مختلف

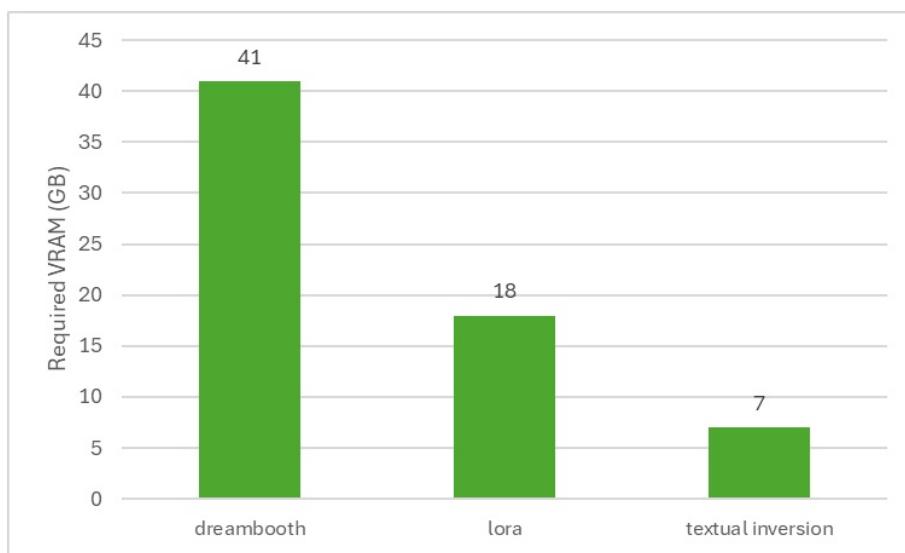
با توجه به نمودار ۴.۵، عملکرد روش های تنظیم دقیق کامل و کم رتبه شبیه به هم می باشد و از سایر روش ها قوی تر است. با وجود اینکه روش وارونسازی متی عملکرد بسیار ضعیفی از خود نشان داده است، نتایج روش رمزگذاری شخصیت با دو روش برتر دیگر قابل رقابت است.

فصل ۴: نتایج

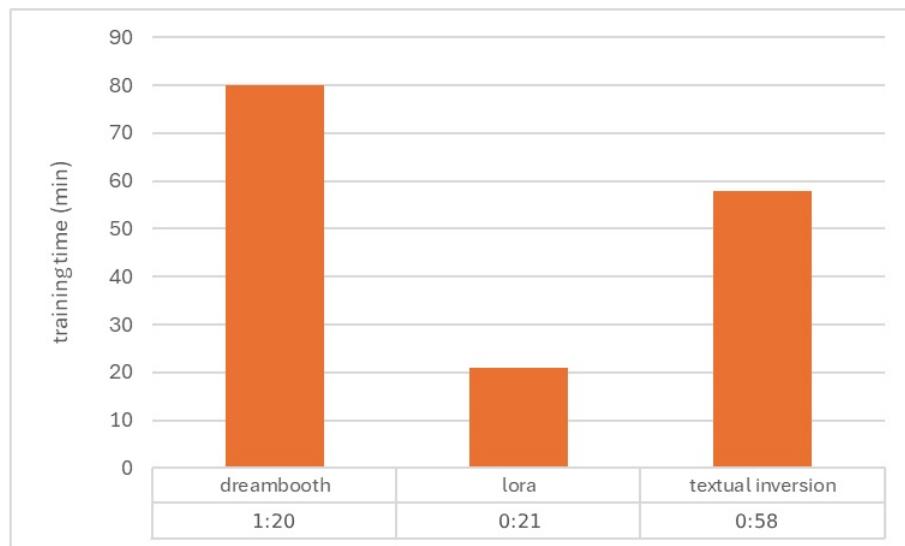


شکل ۶.۴: مقایسه خروجی روش های مختلف

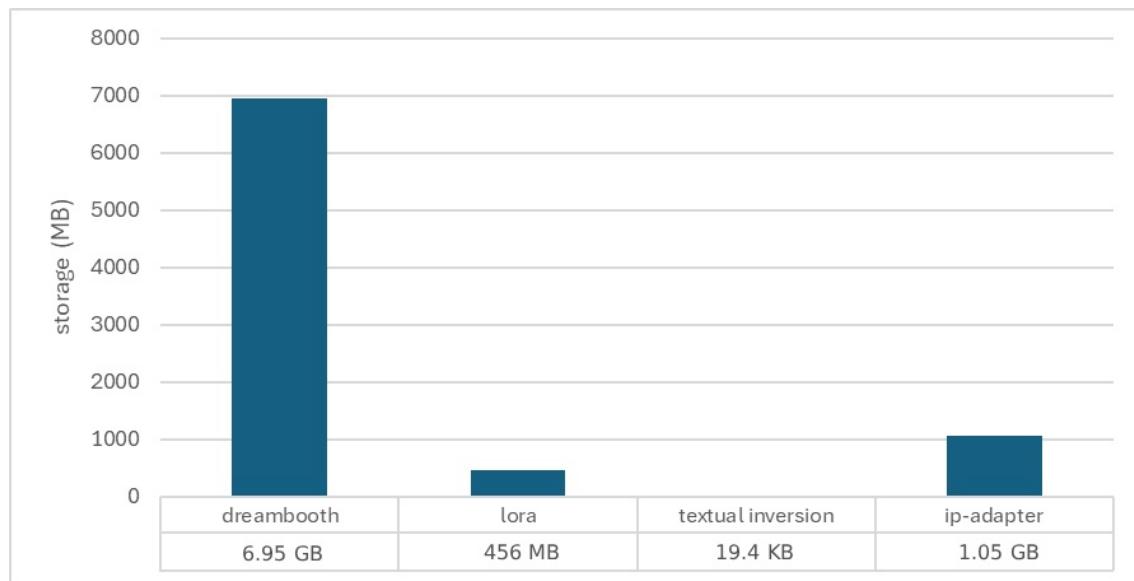
۳.۴ مقایسه هزینه آموزش و بهره برداری



شکل ۷.۴: مقایسه حجم VRAM اشغالی توسط هر روش در زمان آموزش



شکل ۸.۴: مقایسه مدت زمان آموزش برای رسیدن به مدل منتخب از هر روش



شکل ۹.۴: مقایسه حجم دیسک مورد نیاز (صرف نظر از مدل پایه) برای نگهداری از یک شخصیت در هر روش

۴.۴ بررسی مقیاس پذیری

در این بخش مقیاس پذیری را با یک هدف کاربردی مورد توجه قرار می‌دهیم. برای اینکه بتوانیم سامانه‌ای طراحی کنیم که در آن کاربر بتواند شخصیت‌های خود را تعریف کرده و از آن تصاویری تولید کند، نیاز است مباحثی مثل هزینه آموزش و بهره برداری برای ارائه دهنده این سامانه از نظر زیرساختی به صرفه باشد.

چنانچه بخواهیم از روش تنظیم دقیق کامل استفاده کنیم یا باید تمامی شخصیت‌هایی که توسط کاربران تعریف می‌شوند را روی یک مدل آموزش دهیم که موجب افت کیفیت در بازتولید هر شخصیت می‌گردد و یا برای هر شخصیت مدل جدگانه‌ای اختصاص دهیم. مزیت روش دوم دقت بالا و اشکال آن نیاز به فضای بزرگ برای نگهداری از مدل‌های اختصاصی می‌باشد. توجه شود که برای اینکار نیاز است طبق نمودار ۹.۴ فضایی چندگیگابایتی را به هر شخصیت اختصاص دهیم. همچنین استقرار و پیاده کردن مدلی با این حجم روی پردازنده گرافیکی برای تغییر بین شخصیت‌ها سربار زمانی ایجاد می‌کند.

در روش تنظیم دقیق کم‌رتبه، مسئله فضای ذخیره سازی کمترگ تر است و می‌توان برای هر شخصیت یک مدل جدگانه آموزش داد که فضایی معادل با آنچه در نمودار ۹.۴ مشخص شده برای هر شخصیت اشغال می‌کند. همچنین استقرار و خارج کردن این مدل از پردازنده گرافیکی سربار زمانی کمتری دارد. بر اساس نمودار

۵.۴ کیفیت این دو روش قابل مقایسه هستند و بر اساس نمودار ۸.۴ آموزش این روش نیاز به زمان کمتری نسبت به روش های دیگر دارد. بنابراین روش تنظیم دقیق کمرتبه دارای مزایای بیشتری نسبت به تنظیم دقیق کامل می باشد. در روش وارونسازی متغیر، حجم اشغال شده توسط مدل و سرباری که به حافظه پردازنده گرافیکی اضافه می کند، همچنین زمان استقرار و خارج کردن آن از پردازنده ناچیز است. برای نگهداری از این مدل ها تنها به چند کیلو بایت فضای نیاز است اما بر اساس نمودار ۵.۴ و همچنین مشاهده خروجی ها در شکل ۴.۴ می توان گفت کیفیت این روش در کاربردهای مطرح شده نسبت به سایر روش ها ضعیف است.

در روش رمزگذاری شخصیت، مشکلات روش های دیگر وجود ندارد. در این روش به ازای هر شخصیت، تنها لازم است تصاویر آن شخصیت ذخیره گردند تا در زمان استنتاج برای رمزگذاری استفاده شوند و یا حتی می توان بردار رمزگذاری شده تصاویر را نگهداری کرد که حجم کمتری دارد. البته این روش به یک تطبیق دهنده نیاز دارد که حجم آن در نمودار ۹.۴ آمده است. این تطبیق دهنده یک بار ذخیره می گردد و روی پردازنده گرافیکی مستقر می شود و به ازای هر شخصیت سرباری اضافه نمی کند. مزیت دیگر این روش عدم نیاز به آموزش است که از نظر زمان تاخیر در کاربرد حائز اهمیت است.

در این پژوهش روش رمزگذاری شخصیت با مدلی بررسی شد که مختص به رمزگذاری چهره انسان حقیقی بود. این در حالی است که سه روشی که بالاتر توضیح داده شد، نه تنها قادر به یادگیری شخصیت های غیر واقع گرایانه مانند شخصیت های انیمیشنی هستند، بلکه می توانند ویژگی های شخصیت های غیر انسانی را نیز بیاموزند. بنظر می رسد یک روش جامع که می تواند تمامی نیازهای این کاربرد را بر طرف نماید روشی برای رمزگذاری شخصیت است که انواع شخصیت ها را پوشش دهد.

فصل ۵

جمع بندی

در این پژوهش چهار رویکرد اصلی برای مقیاس‌پذیری بررسی شدند.

روش تنظیم دقیق کامل با وجود کیفیت بالا، به دلیل نیاز به فضای ذخیره‌سازی زیاد و سربار زمانی در استقرار، در کاربردهایی با تعداد زیاد شخصیت عملیاتی و مقرنون به صرفه نیست.

روش تنظیم دقیق کم رتبه نسبت به روش قبل بهبود قابل توجهی در زمان آموزش، حجم ذخیره‌سازی و سربار پیاده‌سازی دارد، در حالی که کیفیت خروجی آن همچنان قابل رقابت باقی می‌ماند. بنابراین این روش در شرایط کنونی بهترین گزینه برای پیاده‌سازی کاربردی است.

روش وارونسازی متنی از نظر نیاز به فضای سرعت استقرار بسیار سبک و مناسب است، اما کیفیت تولیدات آن نسبت به سایر روش‌ها پایین‌تر است. این روش بیشتر برای کاربردهای سبک یا نیازهای کم‌کیفیت می‌تواند مفید باشد.

روش رمزگذاری شخصیت نویدبخش‌ترین رویکرد به نظر می‌رسد. این روش نیاز به آموزش ندارد، سربار ذخیره‌سازی شخصیت‌ها تقریباً صفر است و تنها یک تطبیق‌دهنده کوچک بارگذاری می‌شود. کیفیت تولیدی این روش در آزمایش‌ها امیدوارکننده بوده و با توسعه بیشتر می‌تواند جایگزین برتری برای روش‌های مبتنی بر آموزش گردد.

با توجه به این نتایج، در حال حاضر تنظیم دقیق کم رتبه به دلیل تعادل میان کیفیت، سرعت و هزینه به عنوان بهترین روش برای کاربرد مطرح شده انتخاب می‌شود؛ اما به نظر می‌رسد آینده متعلق به روش‌هایی همچون رمزگذاری شخصیت و رویکردهای بدون نیاز به آموزش است.

در پایان لازم به ذکر است که نگاهی به روندهای اخیر در صنعت نشان می دهد که مدل های زبانی بزرگ چندوجهی^۱ به سرعت در حال پیشرفت هستند و شرکت های بزرگ فناوری سرمایه گذاری گسترده ای روی آن ها انجام داده اند. این مدل ها با ترکیب توانایی درک متی و تصویری و بهره گیری از یادگیری درون بافتی^۲ قادرند نه تنها تصاویر جدید با کیفیت بالا تولید کنند، بلکه ویرایش های هوشمندانه و مبتنی بر زمینه نیز انجام دهند. چنین رویکردهایی به عنوان یکی از مهم ترین ترند های جاری، نوید بخش شکل گیری نسل بعدی سامانه های مقیاس پذیر و خلاق در تولید و پردازش تصویر بر اساس شخصیت های متنوع هستند.

¹Multi Modal Large Language Models

²In-Context Learning

کتاب نامه

- [1] Asgari, Erfan. Character consistency project dataset. <https://huggingface.co/datasets/erfanasgari21/character-consistency>, 2025.
- [2] Asgari, Erfan. Character consistency project models. <https://huggingface.co/erfanasgari21/character-consistency-models>, 2025.
- [3] Kingma, Diederik P and Welling, Max. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [4] Goodfellow, Ian J., Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, and Bengio, Yoshua. Generative adversarial networks. *arXiv preprint arXiv:1406.2661*, 2014.
- [5] Ho, Jonathan, Jain, Ajay, and Abbeel, Pieter. Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*, 2020.
- [6] Gainetdinov, Ainur. Diffusion models vs. gans vs. vaes: Comparison of deep generative models, 2023. Accessed: 2025-09-16.
- [7] Song, Jiaming, Meng, Chenlin, and Ermon, Stefano. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2021.
- [8] Dhariwal, Prafulla and Nichol, Alex. Diffusion models beat gans on image synthesis. *arXiv preprint arXiv:2105.05233*, 2021.
- [9] Ghojogh, Benyamin and Ghodsi, Ali. Diffusion models: Tutorial and survey. 07 2024.
- [10] Rombach, Robin, Blattmann, Andreas, Lorenz, Dominik, Esser, Patrick, and Ommer, Björn. High-resolution image synthesis with latent diffusion models. *arXiv preprint arXiv:2112.10752*, 2021.
- [11] Ruiz, Nataniel, Li, Yuanzhen, Jampani, Varun, Pritch, Yael, Rubinstein, Michael, and Aberman, Kfir. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022.

- [12] Hu, Edward J., Shen, Yelong, Wallis, Phillip, Allen-Zhu, Zeyuan, Li, Yuanzhi, Wang, Shean, Wang, Lu, and Chen, Weizhu. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.
- [13] Herrmannova, Dasha. Low-rank adaptation (lora). <https://oneminutenlp.com/p/low-rank-adaptation>, 2024. Accessed: 2025-09-16.
- [14] Wu, Yujia, Shi, Yiming, Wei, Jiwei, Sun, Chengwei, Yang, Yang, and Shen, Heng Tao. Difflora: Generating personalized low-rank adaptation weights with diffusion. *arXiv preprint arXiv:2408.06740*, 2024.
- [15] Gal, Rinon, Alaluf, Yuval, Atzmon, Yuval, Patashnik, Or, Bermano, Amit H., Chechik, Gal, and Cohen-Or, Daniel. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022.
- [16] Ye, Hu, Zhang, Jun, Liu, Sibo, Han, Xiao, and Yang, Wei. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.
- [17] h94. Ip-adapter-faceid. <https://huggingface.co/h94/IP-Adapter-FaceID>, 2024. Accessed: 2025-09-15.
- [18] bmaltais. kohya_ss: Gui and cli tools for training stable diffusion (including lora, dreambooth, textual inversion, etc.). https://github.com/bmaltais/kohya_ss, 2025. Commit: master (or specify tag/commit hash if desired); accessed: 2025-09-16.
- [19] Podell, Dustin, English, Zion, Lacey, Kyle, Blattmann, Andreas, Dockhorn, Tim, Müller, Jonas, Penna, Joe, and Rombach, Robin. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

آپیوست

تصاویر و توصیفات مورد استفاده

در این پیوست تصاویر و توصیفات شخصیت مورد استفاده برای آموزش مدل‌ها آمده است که در دسترس عموم قرار دارد.



شکل آ.۱: تصاویر شخصیت مورد استفاده برای آموزش

این تصاویر به همراه دستور ورودی زیر به هوش مصنوعی Gemini داده شدند تا برای آنها توصیف متنی تولید کند:

I want you to caption some images for LoRA training. Use [V] to refer to the man. Keep captions short, only concentrating on differences, like facial hair, pose, camera distance, clothing, age hairstyle etc.

بررسی و مقایسه روش های افزودن شخصیت های بازتولید پذیر به مدل های مولد تصویبیوست آ: تصاویر و توصیفات مورد استفاده

تصویر	توصیف
1	[V], wearing a yellow and black jacket, smiling, close-up
2	[V], wearing a yellow shirt, smiling
3	[V], wearing a red shirt, striking a playful pose
4	[V], wearing a white sweater, leaning on his hand, close-up
5	[V], wearing a brown shirt and denim vest, smiling
6	[V], wearing a yellow t-shirt, lying in grass, outdoor setting
7	[V], wearing a white and red shirt, waving
8	[V], wearing a black t-shirt, posing with hand to chin, close-up
9	[V], wearing a dark blue shirt, arms crossed
10	[V], wearing a blue jacket, arms crossed, outdoor setting
11	[V], looking serious, close-up
12	[V], wearing a blue t-shirt, sitting
13	[V], smiling, standing with a group of people, full body
14	[V] with sunglasses, wearing a light-colored shirt, smiling, close-up

جدول آ.1: توصیفات مربوط به تصاویر مورد استفاده برای آموزش در شکل آ.

پیوست ب

متن ورودی مورد استفاده برای تولید تصاویر

در این پیوست، متن ورودی که برای تولید تصاویر شخصیت مورد استفاده قرار گرفته‌اند آمده است.

شماره	ورودی متن
01	[V] man in a rainy street at night, soaked trench coat, neon reflections, graphic illustration, comic art, graphic novel art, vibrant
02	[V] man wearing a white chef uniform in a professional kitchen, holding a ladle, close-up, shallow depth of field, food photography lighting
03	close-up of [V] man in cyberpunk setting, glowing tattoos on face, futuristic city backdrop, blue and purple hues, hyper-realistic render
04	[V] man in a medieval knight costume, close-up of his face with helmet off, soft lighting, castle ruins in the background, high detail skin texture
05	close-up of [V] man in traditional Japanese yukata, cherry blossoms behind, soft sunset lighting, fine art photography style
06	medium shot of [V] man sitting in a cozy cabin with warm lighting, wearing flannel and beanie, holding a mug of coffee, hygge vibe, photorealism

بررسی و مقایسه روش های افزودن شخصیت های بازتولید پذیر به مدل پیوشهای قوای بحث‌محبوبین ورودی مورد استفاده برای تولید تصاویر

07	black and white portrait of [V] man in a 1920s suit and fedora, old film grain effect, serious expression, noir aesthetic
08	[V] man in astronaut suit inside space station, glowing screens around, focused face expression, cinematic realism, medium shot
09	[V] man as a plastic figurine, miniature product photography
10	[V] man crying in a leather biker jacket, leaning on a vintage motorcycle, dusty desert backdrop, golden hour lighting, tears on face, stylized realism
11	line art drawing of [V] man holding a camera, urban rooftop, casual modern outfit, sleek, minimalist, graphic, line art, vector graphics
12	[V] man in traditional African attire, outdoor market background, medium shot, rich fabric textures, cultural and realistic photography
13	[V] man in a hazmat suit in a laboratory, reflections on the visor, glowing green elements, scientific thriller aesthetic
14	medium shot of [V] man in denim overalls, standing in sunflower field, rustic summer vibe, golden sunlight, country life realism
15	close-up portrait of [V] man in steampunk goggles and Victorian outfit, mechanical gears in background, stylized realism
16	photo of 2 persons, [V] man and Shah Rukh Khan
17	pixel-art of [V] man, low-res, blocky, pixel art style, 8-bit graphics
18	Old [V] man with long white hair
19	very angry [V] man
20	marble sculpture of [V] man
21	Fauvism painting of [V] man
22	watercolor painting of [V] man
23	[V] man in Ghibli anime style