IntentTuner: An Interactive Framework for Integrating Human Intentions in Fine-tuning Text-to-Image Generative Models

XINGCHEN ZENG, The Hong Kong University of Science and Technology (Guangzhou), China

ZIYAO GAO, The Hong Kong University of Science and Technology (Guangzhou), China

YILIN YE, The Hong Kong University of Science and Technology (Guangzhou), China and The Hong Kong University of Science and Technology, China

WEI ZENG*, The Hong Kong University of Science and Technology (Guangzhou), China and The Hong Kong University of Science and Technology, China

Fine-tuning facilitates the adaptation of text-to-image generative models to novel concepts (*e.g.*, styles and portraits), empowering users to forge creatively customized content. Recent efforts on fine-tuning focus on reducing training data and lightening computation overload but neglect alignment with user intentions, particularly in manual curation of multi-modal training data and intent-oriented evaluation. Informed by a formative study with fine-tuning practitioners for comprehending user intentions, we propose *IntentTuner*, an interactive framework that intelligently incorporates human intentions throughout each phase of the fine-tuning workflow. *IntentTuner* enables users to articulate training intentions with imagery exemplars and textual descriptions, automatically converting them into effective data augmentation strategies. Furthermore, *IntentTuner* introduces novel metrics to measure user intent alignment, allowing intent-aware monitoring and evaluation of model training. Application exemplars and user studies demonstrate that *IntentTuner* streamlines fine-tuning, reducing cognitive effort and yielding superior models compared to the common baseline tool.

CCS Concepts: • **Human-centered computing** → **User interface toolkits**.

Additional Key Words and Phrases: text-to-image generative model, user intent understanding, and data augmentation

1 INTRODUCTION

Recent advancements in pre-trained text-to-image generative models, such as Stable Diffusion [44] and DALL-E-2 [42], have facilitated the generation of high-quality images from natural language descriptions (*i.e.*, prompts) [47]. These technologies have shown promise in augmenting creative processes across various fields (*e.g.*, news illustrations [35], fashion design [59], and webtoon [26]). However, pre-trained models often fall short in catering to users' diverse and domain-specific demands, particularly when dealing with concepts not included in the training data. Consequently, users increasingly need to tailor text-to-image generation to their unique requirements on design concepts such as styles and clothing or facial features in portraits. Such demand has spurred the community of artificial intelligence (AI)

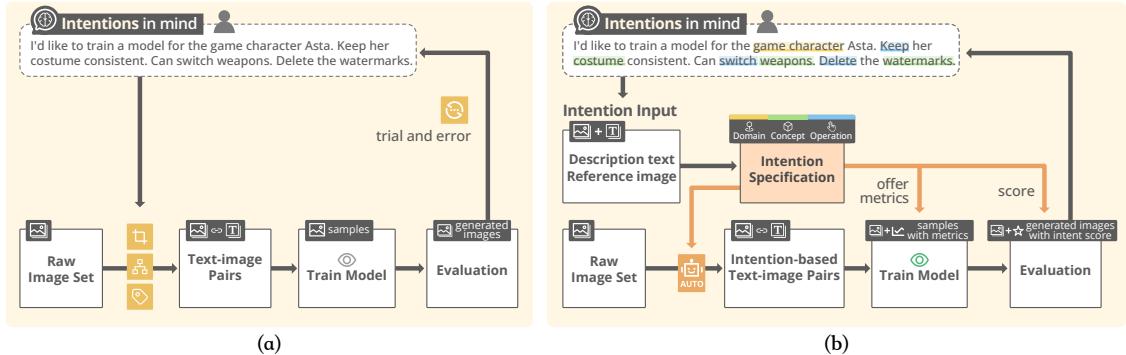*Wei Zeng is the corresponding author

Fig. 1. **Comparison of pipelines of general and our intent-aligned fine-tuning framework.** (a) General pipeline. Most users rely on a trial and error process to check whether the system properly understands their intents, where they manually preprocess the training images, such as cropping ⬚, categorizing ⬚ and tagging ⬚, and observe the generated images. (b) Our pipeline. *IntentTuner* allows users to efficiently articulate their intents to automatically steer important milestones of the fine-tuning, including data augmentation, training monitoring, and evaluation.

art and AI design users to learn and adopt ***fine-tuning***, a powerful technique that enables pre-trained models to learn new concepts with a small number of additional training examples reflecting users' desired outcomes. This practice has given rise to burgeoning online communities for sharing models fine-tuned by users, such as "Civitai" [1] and "Liblib AI" [3]. Furthermore, fine-tuning tutorials targeting non-expert users are receiving considerable attention, expanding the user base beyond AI experts to include artists, designers, and novice users.

Existing research has primarily focused on developing efficient fine-tuning methods [18, 21, 46] from the model's perspective, aiming to reduce the number of images required for a model to learn new concepts (*e.g.*, DreamBooth [46]) and to lighten computational resources demands (*e.g.*, Low-rank adaptation [21]). However, after selecting a specific fine-tuning method, effectively applying it remains challenging for AI novices and even for AI experts. When fine-tuning, users often have **conceptual intents** in mind, including what features the model should learn to keep and what features should be modified or deleted, such as shown in Fig. 1 ⊕. Difficulties arise when further aligning these intents with the technical steps involved in the fine-tuning process. Specifically, users are expected to translate their intents into concrete data strategies (*e.g.*, image augmentation and caption optimization) and performance assessments (*e.g.*, model evaluation and selection), which is currently a trial-and-error process (Fig. 1 (a)). Moreover, measuring the quality of generated images has been a challenging and ongoing research problem [49], especially evaluating the alignment between the fine-tuned model and user intents [61]. Recently, the community has developed interactive fine-tuning tools (*e.g.*, Koyhass [2]), offering detailed training settings, automated image captioning, and metrics visualization to aid in monitoring the training process. However, the tools do not address the challenges above for intent alignment and exhibit an "engineering mindset," only allowing users to control low-level settings and observe hard-to-interpret predefined metrics without an explicit connection to user intents. Moreover, these fine-tuning systems are not integrated with the generation interface, requiring users to load the fine-tuned models on a separate UI (*e.g.*, SD WebUI [4]) to try out generation. Thus, existing tools are insufficient for improving fine-tuning results, reducing the trial-and-error workload, and increasing the accessibility to a broad user community.

To tackle these challenges, we present *IntentTuner*, an interactive framework for integrating user intents into fine-tuning text-to-image generative models with threefold considerations: 1) understanding user intents via natural descriptions and interactions; 2) efficiently translating user intents into intent-aligned data strategies; and 3) monitoring and evaluating the training pipeline in an intent-aligned manner. The design of *IntentTuner* draws insights from a

preliminary study that delves into comprehending the practical fine-tuning workflow, the user intent structure, and the challenges users encounter (Sect. 3). As shown in Fig. 1 (b), our framework translates multi-modal user input into intent specifications to explicitly guide the data augmentation, training monitoring, and model evaluation (Sect. 4).

To facilitate our framework, we have developed an interactive system that enables users to articulate their intents effortlessly by focusing on specific concepts of target effect images with visual interaction and clarifying their intents with natural language (Sect. 5.1). These inputs are then mapped to clear intent specifications, including **Keep**, **Modify**, and **Delete** intents. Specifically, as shown in Fig. 1 (b), they are organized into a hierarchy of domain-concept-operation, linking associated domain (*e.g.*, *game character*) and concepts (*e.g.*, *costume-keep*, *weapons-modify*, and *watermarks-delete*). Based on the intent specifications and the characteristics of the training dataset, the system automatically generates strategies for data processing, basic hyperparameter setting, and model evaluation (Sect. 5.2). Our system subsequently supports intuitive training monitor and evaluation with intent-specific metrics combined with sample generation (Sect. 5.3). Overall, our framework and system aid users in efficiently and intuitively creating and evaluating fine-tuned models, reducing cognitive load during the process, enhancing the performance of the original dataset, and improving training and iteration efficiency.

In summary, our major contributions include:

- We introduce a novel framework to intelligently integrate human intents in fine-tuning text-to-image generative models by effectively decomposing user inputs into intent-aligned data augmentation, model monitoring, and evaluation strategies.
- We develop an integrated system that unifies the fine-tuning and generation into a holistic interface that enables both expert and novice users to flexibly customize text-to-image generation models based on their intents expressed in multi-modal natural inputs. It simultaneously supports user-friendly monitoring and evaluation, facilitating the intuitive selection of generation models.

Ethics statement. All the human portrait images used in this study do not concern any celebrities or other existing human beings. The human portrait fine-tuning data are synthetic and for demo purposes only. The models obtained by our experiment will not be used without consent to generate images imitating any real person.

2 RELATED WORKS

2.1 Text-to-Image Generative models and Fine-tuning

Text-to-image generative models have showcased impressive capabilities of translating natural language description to high-quality images [42, 44, 47], surpassing conventional mainstream generative adversarial networks (GANs) [19, 23] and auto-regressive models [43, 64]. Representatively, Imagen [47] discovers that pre-trained language models are surprisingly effective at encoding text for image synthesis. Rombach et al. [44] introduced latent diffusion models (*i.e.*, stable diffusion), where the forward and reverse diffusion processes happen on the latent space learned by an auto-encoder, remarkably reducing the computational cost. Following works bring further improvements on various down-stream tasks, such as image editing [11, 24], inpainting [37, 65], and style transferring [68], directly contributing to the explosive growth of the Artificial Intelligence Generated Content (AIGC) community [1, 3] and real-world applications. These advancements have also inspired HCI researchers to explore how human users can harness generative models for AI-supported creation or human-AI co-creation [22, 33, 53]. Particularly, many studies focus on developing tools [10, 17, 56] or guidelines [34] to help users adjust prompts to optimize the quality of generated images.

Nevertheless, the outputs of pre-trained diffusion models are inherently constrained by their training corpus, sometimes even directly copying the data [51]. Although techniques like prompt engineering, image in-painting [37] and editing [11, 19], and other image modifications [67] can help users iteratively prompt the models to refine the generation results. These methods still rely on concepts already learned by pre-trained models. Consequently, no matter how much effort is invested in prompt design, they still fall short when confronted with "unseen" concepts spanning from abstract styles to specific content types. To mitigate this issue, researchers have introduced various fine-tuning techniques aimed at instructing pre-trained models about novel concepts, such as Textual Inversion [18], DreamBooth [46], and Low-rank Adaptation (LoRA) [21]. Out of these approaches, LoRA has gained substantial traction within the community for significantly reducing the number of trainable parameters and thus minimizing the need for GPU memory. Our work is designed to support a wide range of users, thus choosing LoRA to maximize its accessibility.

However, the effectiveness of fine-tuning does not solely depend on the chosen training technique but also significantly on other crucial factors, especially the availability of high-quality training images with target concepts embedded and textual descriptions that align well with user intents. Particularly, fine-tuning without well-aligned high-quality data leads to models lacking controllability and generating images excessively resembling patterns in the training data. Furthermore, many models are generated during the fine-tuning process, posing a challenge in efficiently selecting the model that balances controllability and alignment with the user's intent. Our work contributes to the automatic optimization of training data (*i.e.*, text-image pairs) by utilizing user intents as a guiding principle and helping users pick the intent-aligned model from multiple perspectives.

2.2 Evaluation of Text-to-Image Generation

Evaluating the quality of text-to-image generation has been a challenging and ongoing research problem due to the subjective nature of image evaluation and the inherent gap between text and image modalities [49]. Specifically for evaluating the generation after fine-tuning, there are two aspects to consider: the model's ability to replicate target concepts and its controllability in modifying concepts using different textual prompts [18]. Traditional image quality metrics like saliency scores [9] fall short in evaluating similarities between images and establishing connections with the text modality. Inception Score [48] and Fréchet Inception Distance [20] are commonly used to assess generative models by measuring distributional differences between generated images and training images. However, they cannot evaluate either single-image generations or text-image consistency. To evaluate individual generated images based on a prompt, previous studies [49, 64] often employ metrics based on Contrastive Language-Image Pre-Training (CLIP) [41], which compute text-image consistency by cosine similarity between text and image embeddings in the joint representation space. To better align with human preferences, researchers explored fine-tuned CLIP using datasets of human ratings on images created from identical prompts [16, 61, 62]. They further utilized scores predicted by the fine-tuned CLIP to approximate human assessment.

However, these works have focused on evaluating the overall images. Although those board metrics and other narrow metrics (*e.g.*, color harmonious [13]) are useful, the granularity of practical user intents is more moderate, *i.e.*, the alignment between the generated images and users' specific intended concepts (*e.g.*, the hair color of a portrait and the face similarity). Moreover, previous studies have not considered the controllability of the fine-tuned models in modifying fine-grained user-intended attributes, which is a critical factor in evaluating the overfitting of fine-tuning. In our work, we consider both the replication and modification of user-intended concepts at a moderate granularity, ensuring that the outcomes align with user intents while avoiding overfitting.

2.3 Intent Understanding with Large Vision-Language Models

Understanding user intent and incorporating it into interactive systems to make them more user-friendly and intelligent is a common topic of interest in the HCI community. Both language and vision are powerful communication channels. Recently, large pre-trained language models (LLM) have shown significant ability to comprehend fuzzy text inputs, stimulating the development of natural language interfaces for user interactions [12, 30, 45, 50, 52]. For example, Ross et al. [45] investigated the feasibility of using conversational interactions based on code and whether software engineers are open to conversing with LLM. HuggingGPT [50] helps users match their natural language inputs with AI-assisted task requirements and directs users toward the most suitable model published in machine learning communities (*e.g.*, Hugging Face). Regarding the vision channel, Vinker *et al.* [54] presented an approach for decomposing user-provided exemplar images into distinct visual elements. This resulted in a hierarchical structure of sub-concepts, which could be combined and explored through textual inversion to generate imaginative ideas.

In specific scenarios, language or vision alone may be insufficient to convey complex user intents, particularly in visual or cross-modal tasks such as interactive image segmentation [25, 28] and visual question answering [60]. To address this problem, researchers have been exploring using visual-language models [29] to support both language and vision channels. For example, the Segment Anything Model (SAM) [28] allows users to combine textual and visual prompts to precisely segment fine-grained elements within an image. SAM's visual prompts empower users to effortlessly click anywhere on the image or draw a straightforward bounding box to convey their intent for selecting semantic areas. With textual prompts, SAM can be guided to segment objects of interest based on their semantic properties. This innovative approach enhances understanding of user intents and holds substantial promise for HCI applications involving visual data, such as personalized image inpainting [65] and interactive image matting [63].

However, as user intentions for fine-tuning mainly involve infusing new semantic concepts and aligning with the desired visual features, intent comprehension in any single modality is insufficient. In addition, existing multi-modal prompting models like SAM are not tailored to fine-tuning and are unclear as to how to make the vision and language channels complement each other [66]. To fill the gap, our work leverages textual and visual channels to effectively enable users to articulate their fine-tuning intentions. By combining complementary cues from both language and vision channels, we adopt a cross-modal strategy to holistically translate user intents, leading to an intent-aligned optimization of the fine-tuning process.

3 PRELIMINARY STUDY

We conducted a preliminary study (Sect. 3.1) to comprehend the current practice of fine-tuning text-to-image models and the pain points. From the study, we summarized general workflow (Sect. 3.2.1), intentions (Sect. 3.2.2), and challenges users encounter (Sect. 3.2.3) in the fine-tuning process. Based on the findings, we formulated design goals to more effectively integrate human intentions into the fine-tuning of text-to-image models (Sect. 3.3).

3.1 Study Design

Participants. To ensure our tool is widely accessible to users within the model fine-tuning community, we conducted observational tasks and semi-structured interviews with experts from diverse backgrounds: an enthusiast model trainer majoring in e-commerce (P1, Male, Age: 23) that fine-tunes and shares models online but has limited knowledge in technical principles of fine-tuning, and struggles to obtain high-quality datasets and models consistently; two intermediate model trainers with majors in computer science and industrial design (P2-P3, Males, Ages: 23-27) that
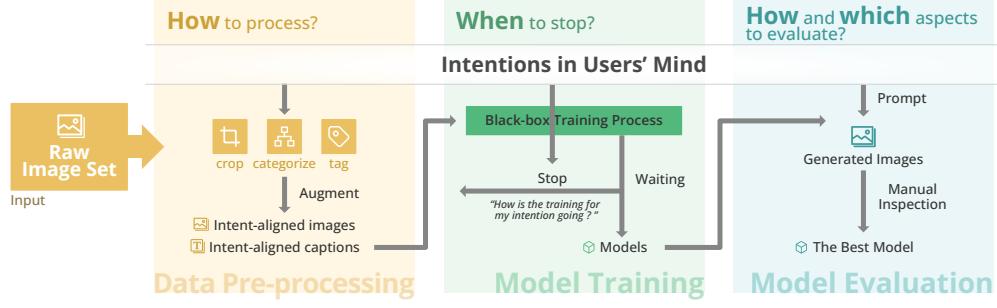
Fig. 2. **General workflow.** The input Raw Image Set is enhanced during the **Data Pre-processing** phase by cropping, categorizing, and tagging according to the intended requirements, producing images and captions that align with the intention. Next, using the processed image-caption pairs, the **Model Training** begins. Users can monitor the progress of model training to determine whether to continue or stop. Finally, users generate images to conduct **Model Evaluation.** Users manually input various prompts to test the model's performance from perspectives related to the intention. After manual inspection, the optimal model is selected.

fine-tune models for academic research and have studied the impact of various settings on training outcomes and possess high-quality datasets and models in specific domains; and a professional model trainer with a background in illustration (P4, Female, Age: 25) that fine-tunes models for commercial use, and has higher demands on result stability and controllability and acquired a comprehensive understanding of technical principles of model fine-tuning.

Procedure. We initially collected demographic information from the participants and requested them to present their previous work to confirm their expertise in fine-tuning. Next, each participant was asked to complete two fine-tuning tasks in two distinct domains, followed by an interview. In *Task 1*, participants were asked to perform a complete fine-tuning process in their familiar domain using the dataset they had prepared. In *Task 2*, they were assigned to fine-tune a model in an unfamiliar domain. Finally, we conducted interviews with the participants. They were asked about their workflows when fine-tuning the text-to-image models, including adjustments to specific strategies in workflow based on their training intentions and the challenges they encountered throughout the process. The entire process lasted from 90 to 120 minutes. After the interviews, we transcribed the audio recordings, and the findings are summarized below.

3.2 Findings

3.2.1 General Workflow. As depicted in Fig. 2, the general workflow of model fine-tuning encompasses three phases, including *Data Pre-processing*, *Model Training*, and *Model Evaluation.*

- **Data pre-processing.** In this phase, users categorize, crop, and tag the raw dataset, to enhance visual features that align with the intention meanwhile reducing undesired content. For example, P1 manually duplicated original images and then cropped the clothing while fine-tuning the 2D Character model in *Task 1*, aiming to "*emphasize character clothing features*." Tagging the training images can be assisted with automation tools. However, manual refinements are necessary because appropriate captions are required to align with the user's intention, whilst current solutions fail to capture.
- **Model training.** Based on the processed data, users set hyper-parameters to start the training process. The process will be terminated if anomalies arise (*e.g.*, non-converging loss values), and users will adjust the dataset or hyper-parameters to restart training. However, the training process is not transparent, as users can only monitor the progress through log data such as loss values that only indicate whether the training converges or not. Due to the subjective nature of the quality of image generation matching intentions, some users choose to
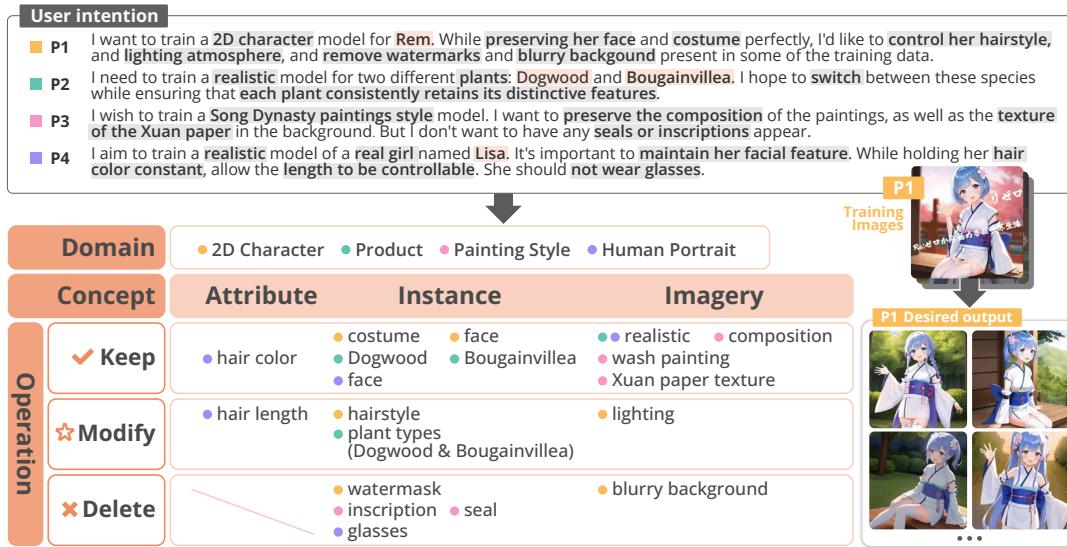
Fig. 3. User intention. We summarize user intentions in **Domain**, **Concept**, and **Operation**. **Domain** refers to the specialized area of creation, such as *"2D character"*. **Concept** defines the specific elements in the intentions, including three different granularities: Attribute (*e.g.*, *"hair color"*), Instance (*e.g.*, *"face"*, *"costume"*) and Imagery (*e.g.*, *"lighting"*, *"blurry background"*). **Operation** encompasses three different types of intended manipulations on specific concepts, including Keep, Modify, and Delete. For example, P1 wants to keep the costume, modify the hairstyle, and delete the watermarks.

periodically generate sample images and break the "black box" progress. They configure a set of intent-aligned prompts to check that current training settings are consistent with their intentions.

- **Model evaluation.** Users select the best model from a series of checkpoints obtained during training. Some checkpoints are often not satisfactory, and users need to manually adjust prompt schemes to better match user intentions based on multiple evaluation metrics.

3.2.2 User Intention. We summarize user intentions in **Domain**, **Concept**, and **Operation**, as depicted in Fig. 3.

- **Domain** designates the specialized area of creation targeted by the fine-tuning (*e.g.*, 2D Character, human portrait, product design, and painting). Based on the general characteristics of the target domain, users can form an initial understanding of the training focus. For instance, P1 stated, "*The focus of a 2D character lies in its costume,*" while P4 believed, "*For real-human, facial features should be the primary concern.*" The pre-trained model selection and training parameters setting are affected by the domain.

- **Concept** is the elements specified in the intention, such as hair length, costume, and background. Depending on the scope of different concepts, we classify them in ascending order of granularity: *Attribute level* (*e.g.*, hair color, facial expression), *Instance level* (*e.g.*, hair, face), and *Imagery level* (*e.g.*, background, lighting atmosphere). Fine-tuning usually introduces new concepts or redefines and aligns existing concepts within the pre-trained model, which are called "trigger words" by users (*e.g.*, "Lisa," "Dogwood," and "Bougainvillea" in Fig. 3).

- **Operation** refers to the intended manipulations of concepts in the generated results, which can be categorized into *Keep*, *Modify*, and *Delete*. *"Keep"* signifies the desire to retain certain concepts from the training set, ensuring that they can be stably invoked by trigger words after fine-tuning. For instance, P1 wishes to preserve the "*character's costume,*" while P3 aims to retain the "*aesthetic composition of Song Dynasty paintings.*" Specifically,

keeping a concept at the instance level essentially means retaining all elements associated with that concept at the attribute level. *"Modify"* represents the capability to adjust certain concepts. Modifying a concept at the instance level can only manifest as switching to a different concept. For example, P2 needed to "*switch different types of plants.*" *"Delete"* removes undesirable concepts from the training data to mitigate any negative effects. Only the concepts at the instance and imagery level can be deleted. For instance, P3 tried to remove the "*seals and inscriptions in painting,*" and P4 preferred the character not to "*wear glasses.*"

3.2.3   Challenges in fine-tuning practice.

- **C1: The abstract intentions are challenging to be translated into clear data strategies**.
Regardless of user expertise, fine-tuning often undergoes a tedious trial-and-error process. One reason is that the training intentions are typically diverse and complex, involving varied operations for multiple levels of concepts. Moreover, the focus and operations in training intention vary significantly across different domains, making it challenging to adapt experience from one domain to another. For example, P1 is "*only familiar with 2D Character model fine-tuning*", and P3 specializes in "*painting style.*" As such, fine-tuning models in alignment with intentions entails substantial learning overhead and cognitive load.

- **C2: Insufficient visual samples and unreliable textual captions hinder intention alignment**.
C2.1: Insufficient visual samples. The quantity and quality of training images for fine-tuning tasks are usually insufficient or deficient, resulting in poor training. For example, during a 2D Character model fine-tuning task, P1 only obtained a limited dataset of 7 images. He had to crop the images and duplicate them to "*emphasize character clothing features*". However, the augmentation process requires repetitive manual operations, making it tedious and time-consuming. Moreover, removing unintended concepts from the raw dataset can be difficult. For example, P1 wanted to "*remove watermarks and blurry background.*" He manually cropped out watermarks to "*remove watermarks,*" but removing "*blurry background*" is challenging by editing the images manually. Adding the concepts into negative prompts can reduce their appearance in the generated images, yet the approach only suppresses the concept's appearance without actually deleting it.
C2.2: Unreliable auto-tagging and cumbersome manual tagging. Manual tagging is labor-intensive, yet auto-tagging methods indiscriminately describe image content that may produce inaccurate and non-intended tags. Users often experience confusion when manually adjusting the auto-tagged captions. They usually "*do not understand the relationship between the tagging strategy and the intended result*" (P1). All participants found the data processing phase "*tedious and challenging*" but believed that "*automated methods are likely to have difficulty replacing this process due to the system's inability to understand my intentions*" (P1).

- **C3: Intuitive training monitoring and effective evaluation are missing**.
C3.1: Unintuitive monitoring of the black-box training process. All participants were concerned about the black-box training process, They monitor anomalies through abstract parameters in the logs but can not establish an intuitive expectation of the outcome. Some training tools have a "sample image" feature. However, this feature "can only observe the effect in a preset perspective" (P3). Moreover, the model training process is volatile. P2 wishes to see "*the trend over time*" to form a clear expectation.
C3.2: Lack of intention-aligned evaluation metrics. After training, users must evaluate the model from multiple perspectives based on training intentions to find the direction for further iterations or select a satisfactory model. For evaluation metrics, current practice requires users to manually adjust prompts continuously to test the

model's alignment with intentions. For evaluation methods, participants believe that the assessment of model performance relies on human subjective feelings.

3.3 Design Goals

Based on the findings, we established three design goals to guide the development of *IntentTuner*:

- **G1: Understand user intentions via natural descriptions and interactions.** The system should automatically extract intention structures from natural user input and align them with concepts in the training dataset for understanding, facilitating the translation of abstract intentions into specific system commands (C1).
- **G2: Provide efficient intent-aligned data augmentation.** Regarding C2.1, the system should support efficient and effective augmentation of image training samples while mitigating potential risks of overfitting. To address C2.2, the system should incorporate intelligent caption optimizations to highlight user intentions in fine-tuning.
- **G3: Offer intention-aware intuitive monitoring and evaluation of model performance.** The system should provide intent-aware metrics (C3) and support intuitive monitoring with trend visualization of the metrics (C3.1) along with an easy-to-use generation panel for swiftly and comprehensively evaluating model checkpoints (C3.2).

4 FRAMEWORK OVERVIEW

Based on the identified design goals, we propose a novel framework to integrate human intent into the fine-tuning workflow. The framework consists of three stages: 1) understanding user intents and transforming them into structured intent specifications (Sect. 4.1); 2) enhancing training datasets with intent-guided image augmentation and caption optimization (Sect. 4.2); and 3) monitoring and evaluating generated images with intent-aligned metrics (Sect. 4.3).

4.1 Language-Vision Intent Alignment and Transformation

Supporting users to express their training intent accurately with a low burden is a non-trivial task due to the multi-modal nature of fine-tuning text-to-image generation. While natural language serves as an intuitive and widely used channel for user input, it alone cannot sufficiently describe user intents due to the complexity of training images. This issue becomes more pronounced when some concepts in the training images share similar semantics. For example, as shown in Fig. 4, when the user intends to teach the model a new human concept while keeping his clothing, they may initially describe their intents in simple keywords (*e.g.*, *"learn the black jacket."*). Language-vision models can not fully clarify users' intents because there are two specific types of black jackets, namely *black leather jacket* and *black striped jacket*. Conveying such detailed intentions requires more concrete textual descriptions, and general vision models still struggle to discriminate such fine-grained categories directly [32]. Therefore, explicit correspondences between text and fine-grained visual elements are needed. To fulfill the requirement, we construct a language-and-vision input alignment stage to assist users in articulating their intents robustly and accurately.

As shown in Fig. 4, users can articulate their intentions by providing text descriptions and reference images. Specifically, users can choose reference images from their image set and use bounding boxes to select specific visual concepts. We set up a unique grammar to help users refer to those visual concepts in the text, indicated by numbers in brackets (*e.g.*, "[1]"). For example, the user inputs "*I want to train a model for a man named Vincent. Ensure his facial features remain consistent. He should be able to switch between a black leather jacket [1] and a black striped jacket [2]. His hair color should be adjustable, and don't let him wear a necklace.*" We target to extract a concrete intent hierarchy
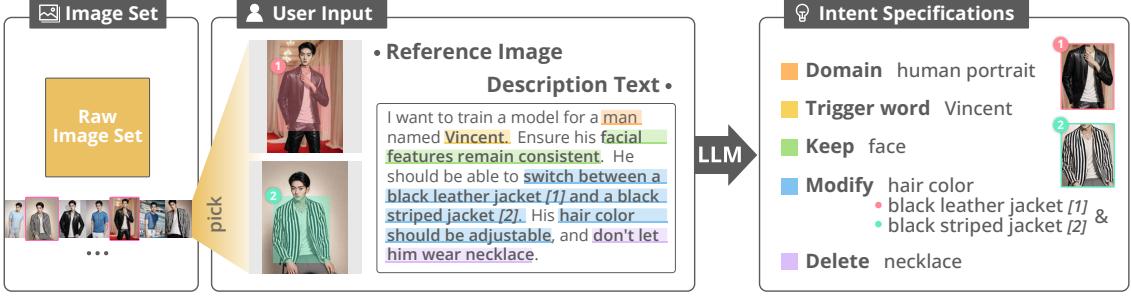
Fig. 4. **Language-vision intent input and transformation.** We allow users to provide detailed multi-modal input to clarify their intents, including the description text and reference images. Powered by the language model, the user input will be transformed into intent specifications, including trigger words, domain, concepts, and operations.

(*i.e.*, domain-concept-operation, see details in Sect. 3.2.2) from the multi-modal input, where operation includes **Keep**, **Modify**, and **Delete**. Among the text descriptions, "face," "hair color," "black leather jacket," "black striped jacket," and "necklace" are detected as visual concepts. Moreover, their associated domain (*e.g.*, "human portrait") and trigger words (*e.g.*, "Vincent") are detected according to the context, and image references will be attached to concepts if any number reference is provided. These intent specifications (*e.g.*, Fig. 4 💡) will guide subsequent parts of our framework.

To implement this transformation, we exploit the in-context learning capability of LLM [38], which empowers the model to conduct novel tasks guided by minimal examples without specialized training for each task. Specifically, we enhance the robustness of intent transformation by structuring few-shot examples in the "chain-of-thought" manner [57], encouraging the LLM to follow a step-by-step reasoning process. In particular, chain-of-thought prompting not only clarifies the questions and requirements about the intent specifications but also provides exemplary intermediate human-like reasoning steps (*e.g.*, the rationale behind classifying the training domain to "human portrait") for the LLM to imitate, making complex inferences more transparent and interpretable. Detailed few-shot examples for instructing the LLM are provided in the Supplementary Materials.

4.2 Intent-guided Data Augmentation

Guided by the operation-concept pairs in intent specifications, we further conduct data augmentation on all raw images for intent-aligned data representations. Below, we introduce the image augmentation methods corresponding to different operations and present the caption generation and optimization strategies.

4.2.1 Image Augmentation. As shown in Fig. 5, image augmentation encompasses two primary stages: 1) detection and filtering of intent-related concepts; 2) augmentation based on different operations. Here, we first employ GroundingDino [32], a pre-trained vision model that excels at identifying image regions corresponding to textual descriptions. GroundingDino takes intent-related concepts as text input and detects the bounding boxes of their associated visual elements, as shown in Fig. 5. The detection results then form the basis for the subsequent augmentation process. To address the semantic ambiguity inherent in GroundingDino's detection, we utilize the earlier reference images as a "filter." Specifically, we calculate the similarity between the detected concepts and the reference image and filter out the visual concepts with low similarities. This approach effectively mitigates the issue of incorrectly detected visual concepts that share similar semantic meanings.

Based on the detected bounding boxes, different augmentation strategies will be performed based on different intents.
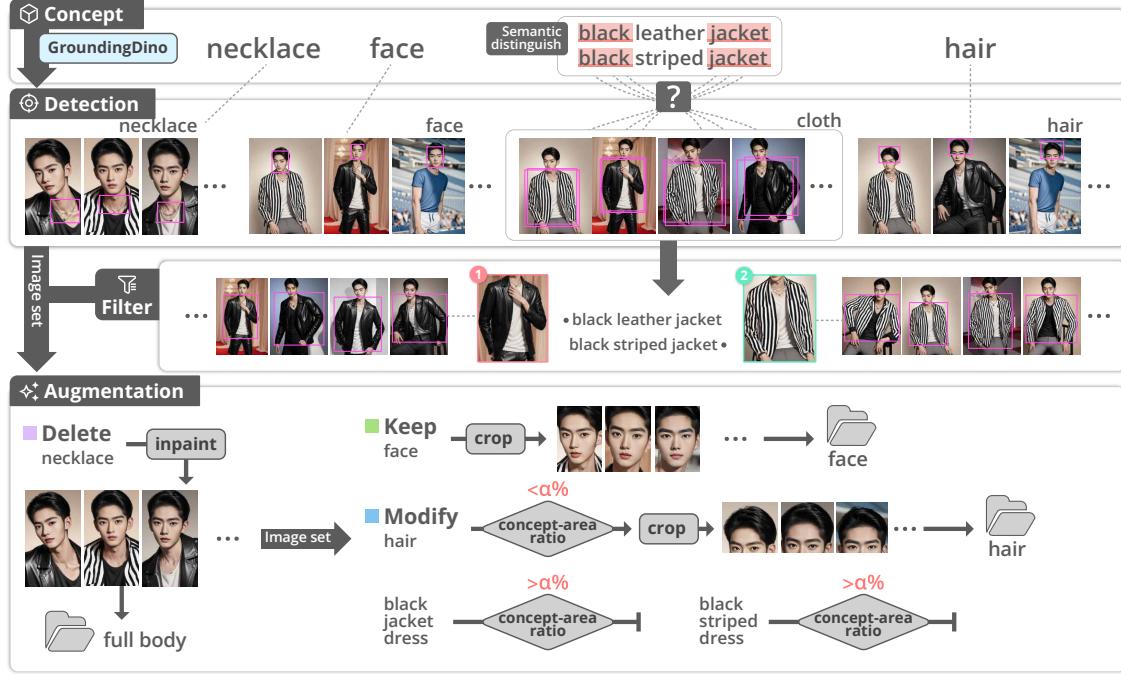
Fig. 5. **Image augmentation.** Based on the intent specifications shown in Fig. 4, we introduce a language-vision intent filter to transfer users' precise intentions to achieve intent-guided data augmentation. Specifically, the fine-grained concepts are passed to a cross-modal Detection module, which can disambiguate the intended concepts and locate the corresponding visual concepts. Then, in the Filter module, users can accurately retrieve samples with the specified concepts with the help of the reference images. Finally, the concept-aligned samples are augmented based on different intended operations to provide more intent-aligned fine-tuning data.

- *Delete* intent will require further inpainting. Cropping an image directly from the bounding box of a visual concept is an intuitive approach. However, this method may also remove visual and semantic information embedded in the cropping regions of the original image. To lighten such an issue, we employ inpainting techniques [44] to remove visual concepts while redrawing the removed area based on the surrounding area, thus preserving as much of the original semantic and visual information as possible.

- *Keep* intent will be translated as cropping the images inside the bounding boxes and adding the cropped parts as independent images in the training dataset. This will allow the model to pay more attention to the semantic and visual information of the concept in the following training phase.

- *Modify* intent is similar to *Keep*, which should be concerned about whether it takes an appropriate proportion of the training dataset. Differently, we set a trigger threshold (*i.e.*, the area percentage of the concept in the original image, default set to 40%) for the image cropping operation based on two intuitive principles from practice. Firstly, too many new independent images will significantly lead to an increase in training time. Secondly, repeating concepts that are inherently salient in original images easily leads to overfitting.

The resulting images will constitute a new image dataset that consists of multiple sub-folders, such as *face, hair*, and *full body* in Fig. 5, which will be used for subsequent training instead of using the original dataset.

4.2.2 Caption Optimization. Text-to-image generation is inherently a multi-modal process that requires each training image to be coupled with an intent-aligned text caption. Commonly used caption generation strategies leverage the
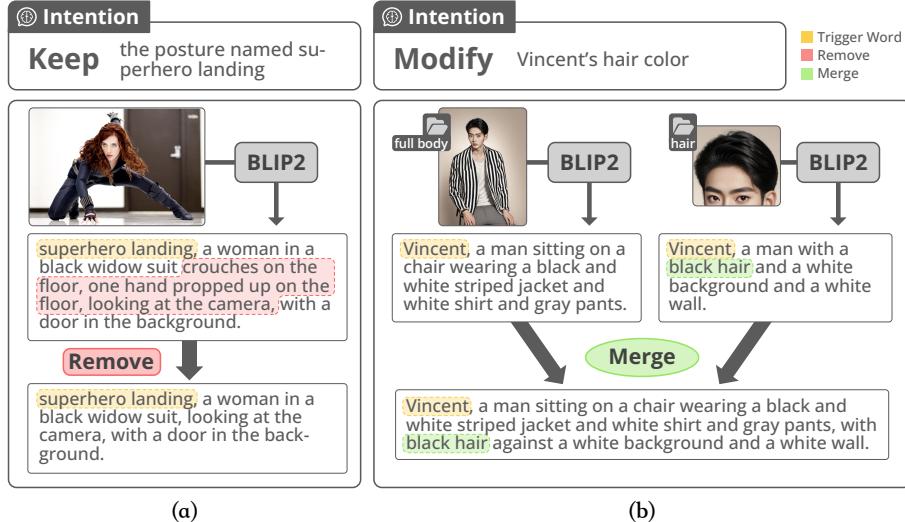
Fig. 6. **Caption optimization.** Caption optimization can intelligently enhance auto-generated captions based on intent specifications. (a) For *Keep* intent, the optimization detects and deletes the redundant concept description to maintain an unambiguous mapping to the trigger word. (b) For *Modify* intent, the optimization focuses on the target concept and generates a detailed description that complements the initial caption, to avoid the visual concept being bound to other words.

CLIP series model (*e.g.*, BLIP2 [29] and CLIP interrogator [40]) to caption images. However, an image can correspond to vastly different captions, and such processes do not consider user intent and require further manual adjustment of generated content that significantly increases user workload. To take advantage of the intention specifications obtained in Sect. 4.1, we design intelligent intent-aligned caption optimization. We first utilize the state-of-art captioning model BLIP2 [29] to automatically generate the initial caption of the image and add the trigger word at its beginning. Next, the caption optimization strategy will help users clarify their *Keep* and *Modify* operations. *Delete* operation has no caption optimization as the concept has been deleted during the image augmentation stage.

Keep means that users hope the concept can be bound to the trigger word (*"superhero landing"* in Fig. 6 (a)), so they can always generate an image containing the referred concept after fine-tuning. For this purpose, the caption optimization needs to identify and remove contents in the initial caption that relate to the referred concepts so that the model can learn an explicit mapping between the trigger word and the referred visual concept. Specifically, as in Fig. 6 (a), users want to keep a posture called "superhero landing" and use this single phrase to prompt the fine-tuned model to generate such posture. The initial caption automatically generated is *"superhero landing, a woman in a black widow suit crouches on the floor, one hand propped up on the floor, looking at the camera, with a door in the background"*, containing detailed descriptions of the posture (*"crouches on the floor, one hand propped up on the floor, looking at the camera"*), which can confuse the generative model as to what the trigger word refers to. Guided by the *Keep* operation in the intent specifications, we prompt the LLM to locate the relevant descriptions and remove them, yielding the optimized caption: *"superhero landing, a woman in a black widow suit, looking at the camera, with a door in the background"*. As a result, the special pose will be bound to the trigger word and not be confused with other descriptions.

Modify is incorporated to control and change certain concepts in the fine-tuned model. For this purpose, the caption should include a detailed description of the concept to avoid its visual features being bound to the trigger word or other parts of the caption. However, in some cases, the generated caption may overlook the concept and not provide a

detailed description. For example, as shown in Fig. 6 (b), users want to control the portrait's hair color. However, the automatically generated caption completely overlooked the visual concept. To describe the hair attribute in detail, we leverage the previous concept detection results to focus on the hair and prompt BLIP2 to generate a detailed caption. Next, we prompt the LLM to merge the initial caption with the detailed caption to produce the final optimized caption, successfully incorporating the user's intent into the captions.

4.3 Intent-aligned Image Evaluation

We aim to allow users to assess the alignment between model output and their intents quantitatively rather than merely observing images subjectively. Specifically, we evaluate the effect of text-to-image fine-tuning by measuring the generated images from two complementary aspects: stability and controllability. Stability measures the model's ability to replicate the target concept [18], *i.e.*, whether the target concepts in the generated images are visually similar to users' expectations. However, only considering stability may cause overfitting, as the model may only remember and repeat the training data in extreme cases and lose diversity in other visual properties [51]. For instance, for an overfitted model, regardless of how users set the positive and negative prompts, users cannot obtain specific properties (*e.g.*, "*long hair*"), and all the generated images end up being similar, such as having short hair. In other words, users lose control of other properties like hair via prompts because the overfitting leads to mode collapse. To mitigate this issue, we further develop controllability to measure the model's ability to modify the concepts using textual prompts. For instance, hair length controllability refers to whether users can manipulate the output images to have varying hair lengths by changing the prompts.

Stability. We evaluate model stability at the granularity of user-desired concepts, not limited to the overall image. This allows users to evaluate multiple intents independently. Specifically, we crop intent-related objects and compute the visual similarity between them and each generated image. Conventional methods that transform images into high-dimensional vectors and compute vector similarity based on a pre-trained feature extractor (*e.g.*, CLIP-Vit and CLIP-ResNet [41]) neglect the perception of human preferences [62]. Inspired by [61], we leverage the human-preference classifier to mitigate the issue, which is fine-tuned on a large-scale human-labeled dataset that focuses on discriminating the common drawbacks of the generated images compared to the real ones. Specifically,

$$\text{Stability} = \frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \text{sim}(I_i, R_j), \tag{1}$$

where I refers to the sampling image set during training and M is its batch size. R refers to the intent-related object set, N is its total number, while $sim(\cdot)$ refers to the similarity calculation based on the human preference classifier [61].

Controllability. To evaluate whether we can modify specific concepts as expected, we first use prompts containing opposite attributes of the specified concept (*e.g.*, short hair *vs.* long hair) to generate sampling images. Then, we measure the standard semantic alignment metric in a high-dimensional, non-linear neural embedding space like CLIP to represent text-image similarity. Inspired by previous research [17], we leverage the two opposing attribute keywords to transform the image evaluation task to a binary classification, which can effectively reduce the ambiguity that arises from using a single attribute [55]. We then calculate the latent space cosine similarity of each image with the two opposing text keywords of user-intended control attributes (sim_1 and sim_2, where sim_1 denotes the similarity to the correct control keyword and sim_2 denotes the similarity to the opposing keyword). Next, we compute the controllability score as the
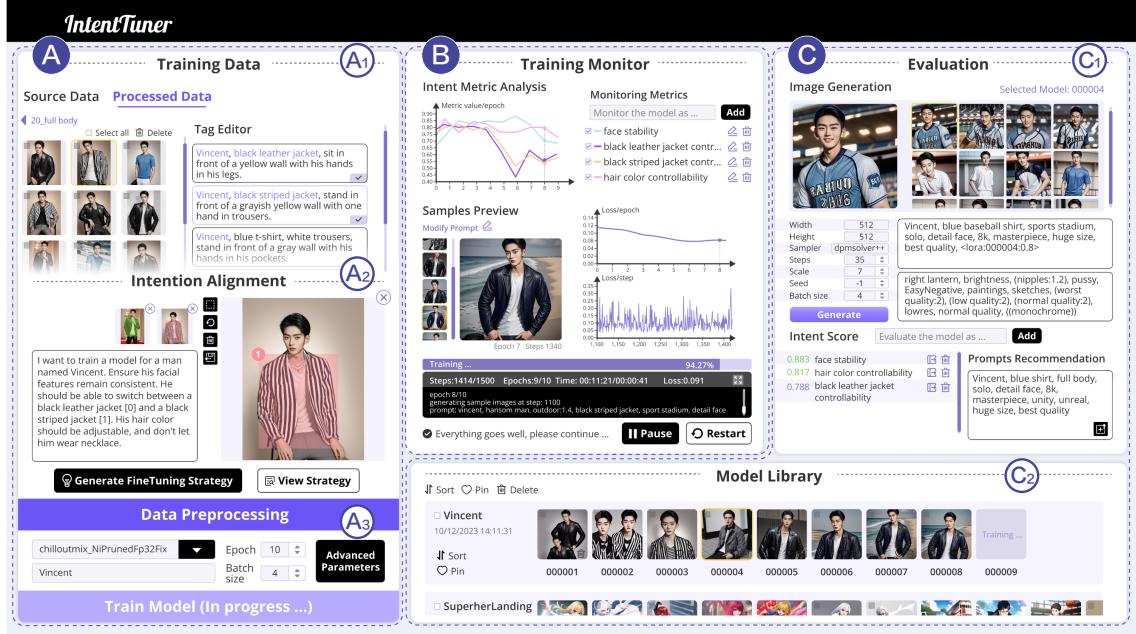
Fig. 7. **User interface of *IntentTuner*.** The *Intention-Data Alignment Module* (A) allows the user to input the model's fine-tuning training data and intentions, conduct pre-processing, and configure other training settings and hyperparameters. The *Training Monitor Module* (B) monitors and visualizes training progress based on intentions. The *Model Evaluation Module* (C) helps users evaluate models based on multiple metrics from intentions.

normalized Softmax similarity toward the intended attributes:

$$\text{Controllability} = \frac{1}{M} \sum_{i=1}^{M} \frac{e_i^{sim_1}}{e_i^{sim_1} + e_i^{sim_2}}, \tag{2}$$

where *M* is the batch size of sampling images, and *sim* refers to the text-image similarity.

5 SYSTEM DESIGN

We develop an interactive system to support user-intended fine-tuning of text-to-image models. This section introduces the system's interface and interaction design to accomplish: 1) user intent understanding (Sect. 5.1), 2) automated data preparation (Sect. 5.2), and 3) intent-aligned monitoring and evaluation (Sect. 5.3), as described above. We associate them with the design goals in Sect. 3.3 to underscore the design rationale.

5.1 User Intent Understanding

The *Intention Alignment* panel (Fig. 7 A2) allows users to input their intentions naturally, assisting them in materializing the abstract training intention (G1). First, users upload their training image set in the Source Data page in the *Training Data* panel (Fig. 7 A1). Subsequently, they can describe training intentions using natural language in the text area of the *Intention Alignment* panel (Fig. 7 A2). After that, to accurately refer to some visual concepts in the text description, users can drag images containing relevant concepts from the training set into the image canvas. The image canvas offers a selection tool ▦ to highlight concepts, and it will automatically number the user's selections, distinguishing them with

different colors. Other functions for canvas annotation are provided, such as *undo* ⟳, *clear all* 🗑, and *save* 💾. Users can drag in multiple images for annotation, facilitating the training task with multiple concepts. Then, users can reference the annotated concept in the textual description using square brackets "[]", enabling references between text and specific visual concepts. Finally, users can click the *Generate FineTuning Strategy* button 💡 to automatically translate the intention input into intent specifications (*e.g.*, domain, trigger word, and operation-concept pairs) and recommend prompts needed in subsequent image generation testing accordingly. The intent specifications can be viewed in a JSON format via the "View Strategy" button 🖺, and users can edit them flexibly to increase the controllability of the system's behavior and enhance user trust.

5.2 Automated Data Preparation

When the user's intention aligns with the system's understanding, they can preprocess the entire dataset through the *Data Preprocessing* button. Guided by the intent specifications, the system will automatically perform cropping, inpainting, and classification of the image set and image captioning (G2). The newly processed dataset will be displayed on the Processed Data page in the *Training Data* panel (Fig. 7 A1). Users can view the system's image processing results in this panel, including classified folders and cropped images, and modify the captions. However, the amount of captions in a training set is enormous. To save users' workload in caption modification, the *Tag Editor* supports the flexible propagation of modifications to a specific caption to other captions. In addition, content related to intentions will be highlighted in the captions, helping users efficiently locate key focus information and enhancing their understanding of the relationship between captions and intentions. The design principle behind the *Tag Editor* is providing appropriate user feedback, ensuring the system fully understands user intent before fine-tuning.

The *Settings* panel (Fig. 7 A3) displays the necessary settings for model training (*e.g.*, base pre-trained model and trigger word). The system presets training hyperparameters based on the fine-tuning domain identified from user intents. The system hides hyperparameters other than batch size and epoch in the Advanced Parameters button, which users can click to view and adjust more detailed parameters. For more details and illustrations about default hyperparameters (*e.g.*, optimizer, scheduler), please refer to Appendix A, as different domains correspond to different settings.

5.3 Intent-aligned Monitoring and Evaluation

After preparing the training data, users can click the *Train Model* button to start training. Subsequently, the *Training Monitor* module (Fig. 7 B) is activated. Specifically, the *Intent Metric Analysis* panel visualizes monitoring metrics based on intentions. These metrics are automatically generated based on the system's understanding of intentions, which users can also continue to add or modify. The trend of all metrics is displayed on the same coordinate using line graphs of different colors, helping users form overall expectations of the progressive and fluctuating progress of model training. In addition, the *Samples Preview* panel continuously displays generated images more intuitively to show the training progress. The initial image prompt is automatically generated based on intentions, but users can also manually change the prompt at any time for more flexible observation.

Meanwhile, the checkpoint models generated during training are stored in real-time in the *Model Library* (Fig. 7 C2), and the sample image corresponding to the model is automatically set as its cover to help users browse. Users can select models from the *Model Library* and perform image generation testing in the *Evaluation* panel (Fig. 7 C). The *Intent Score* panel automatically provides intention-related metrics and their performance scores. Users can also continue to add, delete, or modify metrics. All metrics are sorted from high to low based on their scores, making it easy for users to learn
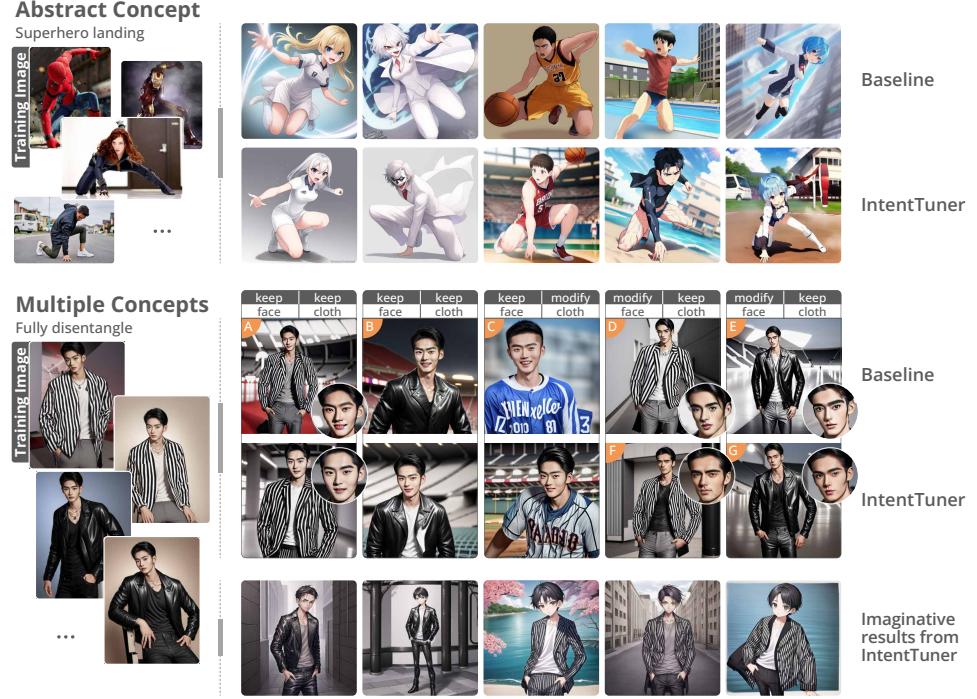
Fig. 8. **Two general fine-tuning usage scenarios.** In each case, the first line presents results from our system, while the second line presents results without an intent-image alignment module. Interesting application results are added in the third line for the multiple concept scenario.

about the strengths and weaknesses of the current model. Based on the metric content, the system can recommend prompts for image generation, and then models can be tested through the *Image Generation* panel (Fig. 7 C1).

6 EVALUATION

The key innovations of *IntentTuner* fall into two categories: 1) a fine-tuning framework that intelligently transforms user intentions into intent-aligned data strategies and 2) an integrated system that unifies fine-tuning and generation, supporting both expert and novice users to customize text-to-image generation models flexibly.

To verify them, we conducted evaluations through two studies, respectively. In **Study 1**, we evaluated the framework across two general fine-tuning application scenarios: 1) abstract concept preserving and 2) multiple concepts augment and modification. We compared the outcomes our intent-aligned fine-tuning technique produced for each scenario with those from baseline systems. In **Study 2**, we conducted a user study, engaging in comparative tasks with a widely-used fine-tuning baseline pipeline which requires users to combine Koyhass [2] and Stable Diffusion Web UI [4]. By analyzing participant questionnaires and interview feedback, we measure the functionality, overall effectiveness, and usability of the system and baseline, affirming its enhancement to the user experience.

6.1 Application Examples

Abstract concept. As shown in Fig 8 (top), the original training dataset includes several portraits with a common pose, namely "superhero landing." Users intend to teach the model such an abstract concept, enabling it to synthesize images

16

of different people in that pose in interesting scenarios We observe that the baseline resulted in images that can learn the "landing" concept, but could not accurately mimic the essence of the pose, *i.e.*, placing one hand on the ground and positioning legs in front and back. In contrast, our approach can apply it stably to objects through our intent-aligned caption optimization strategy. The reason behind the success is to delete all information related to the human pose in the captions (see Sec. 4.2.2 for details). Other examples of abstract style include painting styles, photo filters, *etc*, which often share the same optimization strategy and training parameters.

Multiple concepts. Fig 8 (bottom) shows a complex but, in practice, a rigidly necessary scenario that requires the model to learn several key concepts from a diverse dataset and to distinguish between them. Practical applications, such as e-shopping models and product graphic design [31], often fall into this category. Here, we present a complex example with the intent of "*To learn the model's looks, as well as the product features of the black leather jacket and the black striped jacket, and to be able to support switching between and combining these new concepts without being uncontrollably bound.*" That means the user can prompt to generate images that contain these originally bound concepts, *i.e.*, (portrait face, black leather jacket) and (portrait face, black striped jacket), independently and rebind them on demand, which is also referred to as "disentanglement" [69] in deep learning.

In the results, first, we observe that our intent-aware data augmentation can improve the training quality of individual concepts. For example, in Fig. 8 (A) and (B), where the user tries to generate the original portrait and clothes with baseline, the quality of the face is worse than ours. Second, the problem of conceptual entanglement between the original portrait's face and clothes can be alleviated with our tool. When users intend to modify one of the concepts, such as changing the face or the clothes while keeping the other concept, the features of the original portrait will still affect the generation, directly leading to the collapse of the other portrait. For example, as shown in Fig. 8 (C), when the user attempts to replicate the face but to change the clothes to a sports shirt, the face in the baseline also becomes less similar to the desired original portrait compared to our result. Moreover, we can observe that the facial features of the portraits in the baseline crumble significantly when the user tries to make the other portraits wear target clothes, as shown in Fig. 8 (D) and (E) in the baseline. In contrast, with our tool, the two costumes can be more effectively combined with other portraits, including completely different live-action portraits, as shown in Fig. 8 (F) and (G) where the user successfully controls the face to look like "European." This case demonstrates the strong controllability of the fine-tuned model, which benefits from *IntentTuner*'s powerful data augmentation strategies. In addition, the fine-tuned model can generate learned clothes and synthesize them with imaginative concepts (see the third row of Fig. 8 (bottom)), thus taking full advantage of the generative model.

6.2   User Study

We designed a qualitative study and recruited participants to assess their user experiences and evaluate the quality of the fine-tuned model results compared to the text-to-image model fine-tuning tool widely adopted within the model community. Specifically, we aim to evaluate and compare with the baseline system 1) the usability and usefulness of the intention-aware fine-tuning framework and 2) the overall system's support for fine-tuning practices.

6.2.1   Participants. We recruited 12 post-graduate students (5 females, 7 males, Average age 23.0) through recruitment messages posted on social media platforms. The participants are from various disciplines, including art, design, and engineering. To showcase the versatility of our framework among various user levels, we enrolled six experts, including three accustomed to using training interfaces, three AI developers familiar with code-level operations, and six novices

with no fine-tuning experience but willing to fine-tune in the future. All participants have used the text-to-image generative model.

6.2.2 *Baseline System.* To establish a baseline for comparison against *IntentTuner*, because there is no integrated system that allows users to perform fine-tuning and generation at one stop, we combine two systems widely used in the AIGC community: Kohya-ss [2] for the training phase and Stable Diffusion Web UI [4] for the evaluation phase of our system.

Kohya-ss is a popular open-source project designed for Stable Diffusion trainers and has received 5.7k stars on GitHub. It provides an interface for model training, which comprises panels of pre-trained model selection, path setting, and training parameters. Stable Diffusion Web UI is an open-source user interface for image generation and has received 102k stars on GitHub. It features a suite of image generation functionalities, including Text to Image, Image to Image, and InPainting, *etc.* In this study, we leverage its Text-to-Image feature, which encompasses modules for prompt input, parameter setting, and image generation. To ensure a fair comparison, both *IntentTuner* and the baseline were set up using the same dataset and pre-trained model, and they carried out training tasks within the same domain.

6.2.3 *Procedure and Task.*

- **Introduction.** We first provided a brief introduction of the research background. Next, we gathered the demographic information from the participants and asked for their consent to record their operations and results for further analysis. We then introduced the interface of both our system and the baseline system and demonstrated their usage through a mock dataset. If expert users were already acquainted with the baseline system, they could ask to skip its introduction. Then, we presented the training dataset required for the training tasks to the participants and explained the training objective.
- **Task Design.** To delve deeper into the evaluation of *IntentTuner*'s performance, we designed a model fine-tuning task for users that combines various intentions, containing multiple requirements from the general application scenarios mentioned in Section 6.2.
 We provided participants with the following training requirements: *Please train a model for the real-life girl Sophia. Ensure that her facial features remain consistent and allow for hairstyle adjustments, but she shouldn't wear a hat. Additionally, Sophia has two outfits, including Outfit 1 and Outfit 2, and the model should be able to switch between them.*
 They were asked to organize the training intention based on the given requirements and complete the fine-tuning task using the *IntentTuner* and baseline systems, respectively. They followed the think-aloud protocol during tasks and were free to ask questions at any point. To measure the task efficiency of users with different expertise levels across the two systems and to ensure sufficient exploration for participants to allow them to assess each system in terms of required cognitive effort subsequently, we did not set a time limit.
- **Questionnaire and Interview.** Upon finishing the task, each participant completed a 7-point Likert scale questionnaire regarding the system's functionalities and overall performance. In the system features evaluation, we discussed from the perspectives of intention understanding, data augmentation, and intuitive metrics. For the overall system usability evaluation, we asked for ratings on four aspects: easiness of use, usefulness, flexibility, and engagement. While answering the questionnaire, participants were asked to continue with the "think-aloud" protocol to explain the rationale behind their ratings. Subsequently, a semi-structured interview was conducted, discussing their satisfaction with the training results, their impressions of the system, and expectations and
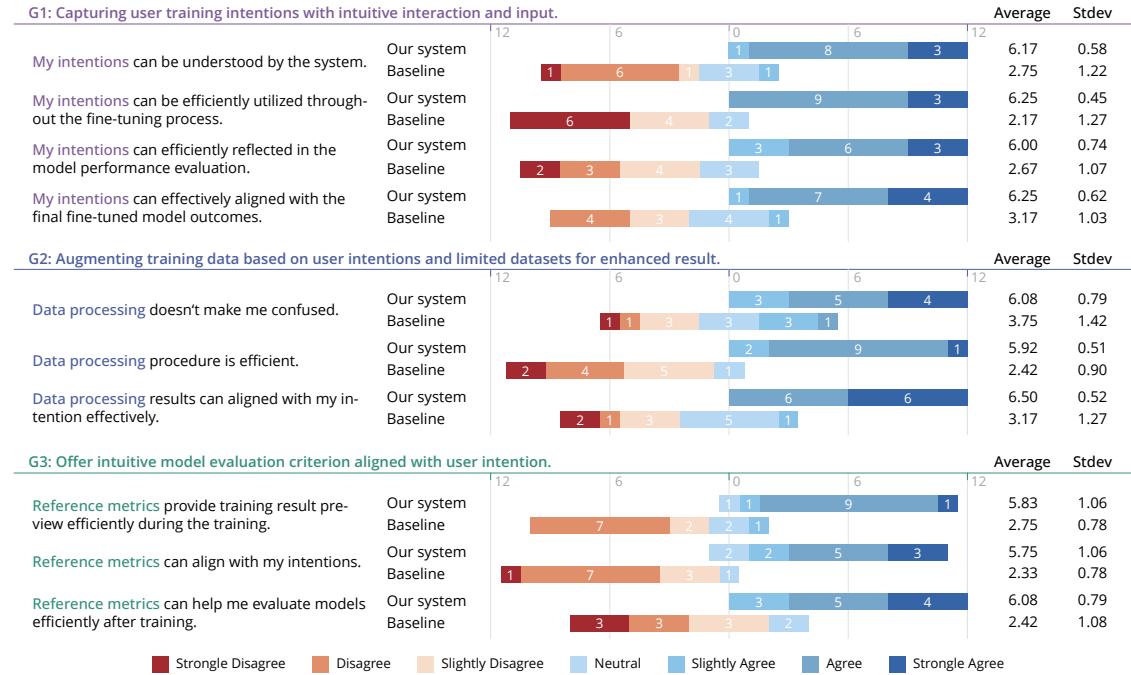
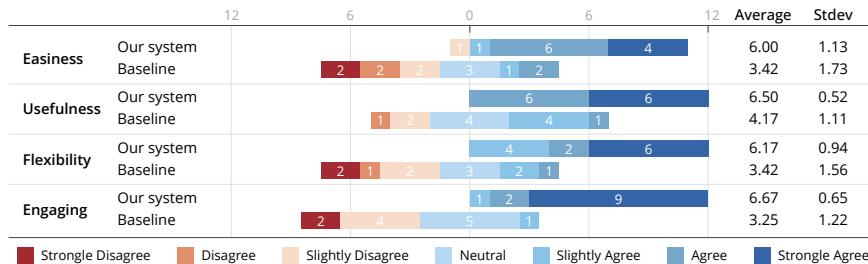Fig. 9. **The results of the questionnaire regarding the features experience of our system and baseline system.**

Fig. 10. **The results of the questionnaire regarding the system impressions of our system and baseline system.**

suggestions to further gauge the user experience with *IntentTuner*. On average, each participant spent 70 minutes on the entire study.

6.2.4 Results. All participants completed both tasks, the following questionnaire and interview. We first report the participants' responses to the functionalities of the system and our observations on their fine-tuning practices. We then discuss the system's overall usability and the limitations we learned from the participants.

- **Feedback on User Intention Capture.** Participants found that our system can effectively comprehend their training intentions and manifest throughout the fine-tuning process, enhancing the final model quality aligned with intention (Fig 9, top). Using natural language and reference images to express intentions is perceived as an intuitive and natural way of communication. This aligns with our first design goal (G1), which is to allow users

to input their intentions effortlessly. This approach "directly reflects the high-level intentions during fine-tuning" (P1), especially when users are fine-tuning models for creative and hobbyist purposes, enabling them to "feel more free without being overly concerned about its feasibility" (P3).

Participants also noted that the system's accurate understanding of their input intentions contributed to building their confidence in the current fine-tuning task. P1, during her initial attempt to input intentions, had some confusion due to uncertainties about the system's granularity in understanding natural language inputs. However, after reviewing the examples we provided and giving it a try, she felt the intention input was "flexible and robust." After inputting their intentions, participants would use the View Strategy feature to assess how their intentions were understood. This feature offered them a "window to ensure the correct comprehension of intentions," reinforcing their understanding and trust in the system's behavior. Participants thought the system was "simpler and more intuitive" (P3) in fine-tuning text-to-image models than the previous workflow.

- **Feedback on Data Augmentation.** Participants believed that our automatic data augmentation could easily, efficiently, and accurately integrate the embedded user intentions into the training dataset (Fig 9, middle). In comparison, even the AI expert users found the baseline system's data processing inefficient. They believed that "manually processed results might not fully align with training intentions" (P7). The intent-aligned data augmentation enabled them to "execute intentions more effectively," addressing our second design objective (G2). The automated process marked a significant improvement over the generic data processing workflow conducted with the baseline system. Participants estimated that manual data processing using generic workflows would require substantial mental and manual effort, "possibly taking 1.5-2 hours" (P1). For novice users, they "might not know how to align intentions with data processing methods without enough training" (P8).

  After reviewing our automated data processing results, expert users believed the outcomes to be "accurate and comprehensive" (P10). P9 stated, "I now feel confident about the automated data augmentation." Additionally, they deemed the caption modification feature as "essential," and the automatic propagation of the modified caption content "saved a lot of tedious steps" (P6). For novice users, they didn't delve much into the data-processed results but focused more on evaluating the outcomes of the training results. After comparing the training results between the baseline system and ours, all participants agreed that our training outcomes "better reflected the training intentions" (P7).

- **Feedback on Metrics Reference.** Multiple participants (7/12) mentioned that our system provided intention-aligned reference metrics during both the training process and the evaluation of training results (Fig 9, bottom), addressing our third design goal (G3).

  Participants believed that our reference metrics could visualize the training progress trend in a user-friendly manner. "The visualization of multiple intent-aligned metrics in the same coordinate provides an intriguing reference, allowing me to easily estimate the model's overall situation" (P7). P7 further mentioned, "Based on the 'pink dress' prompt, I could observe the color transitioning from a vibrant pink to a hue closer to that in the training set, giving me a general understanding of the training progress." The Model Library and Evaluation panel also integrated the model training and image generation processes. P5 felt that "not having to move models and switch tools manually is very convenient." Participants enjoyed testing images on the fine-tuned models after some training time, especially novice users who haven't yet experienced selecting models based on training epochs and other parameters. "This makes full use of the waiting time during training," (P1) noted. Specifically, P1 pointed out that the feature to view image effects during training "is not only suitable for text-to-image model fine-tuning but also applicable to all AI models that target image generation, such as image-to-image."

- **Experience of Overall System.** Overall, participants appreciate the usability of our system more than the baseline system. In terms of the easiness of use, our system ($mean = 6.00, std = 1.13$) scores higher than the baseline ($mean = 3.42, std = 1.73$). As P9 commented, *"even though I'm familiar with prototype AI model applications built using Gradio, I think the Gradio style is more suitable for demonstrating simple model features with fewer control settings. But in the fine-tuning tasks, it shows too many setting parameters, requiring me to scroll up and down the interface to set different parameters, adding to my burden."* Concerning the usefulness, our system ($mean = 6.50, std = 0.52$) also outperforms the baseline ($mean = 4.17, std = 1.11$). Regarding the flexibility, our system ($mean = 6.17, std = 0.94$) is also better than baseline ($mean = 3.42, std = 1.56$). *"The IntentTuner system enables me to flexibly adjust my fine-tuning in more diverse and intuitive textual and visual interactions, while in the other system, I can only provide ready-made training data and adjust the caption because I don't understand the other complex settings,"* P6 remarked. Furthermore, participants also rate our system ($mean = 6.67, std = 0.65$) as more engaging than the baseline ($mean = 3.25, std = 1.22$). P10 said, *"IntentTuner system allows me to intuitively see samples and how much they match my fine-tuning goal during the training, making the process more engaging."* Moreover, P5 noted, *"In IntentTuner, I can directly see the model gallery and instantly use different fine-tuned models to try out the generation, making the experience much more engaging than the other system, which requires me to manually select and load individual model in another UI to test the generation."*
- **System Limitations**. Even though *IntentTuner* can support most fine-tuning requirements, when the user intentions become too complex, such as involving "multiple operations on multiple concepts" (P2), the training time and the uncertainty of results will increase, which is a common problem for existing fine-tuning practice. In addition, some participants (3/12) think we should further improve the interactions in the system to increase the sense of control. For example, P9 noted, *"After I clicked the training button, I did not have much to do except wait for the results and see some intermediate samples. The monitoring is good, but I think you could try to improve this process by adding more interactions that can be performed in parallel."*

6.2.5 Summary. All participants were excited about the capacity of *IntentTuner* to improve their workflows when fine-tuning text-to-image models. They agreed that *IntentTuner* enables them to quickly realize their training intentions, achieve complex composite training objectives, and span various fine-tuning domains. Specifically, the intention-oriented fine-tuning framework enables participants to be objective-driven, achieving rapid and flexible fine-tuning of text-to-image models. The study also revealed that *IntentTuner* still has limitations in dealing with highly complex intentions and providing more user interactions during the training.

7 DISCUSSION

7.1 Customized AIGC: Ethical Risks and Responsive Strategies

While the benefits of advanced generative capabilities are undeniable, their impressive power, coupled with enhanced customization options, also introduces significant ethical concerns, particularly concerning misinformation and intellectual property rights. For instance, the ease of fine-tuning models could lead to a surge in deepfake creations [58], enabling users to generate unauthorized synthetic representations of real individuals, especially celebrities. These representations risk being misused to disseminate false information, potentially causing harm to those depicted. Additionally, the artistic community raises alarms over copyright infringement, as their creations are increasingly harvested without consent for model fine-tuning purposes [27]. These issues may also, in the long term, exacerbate biases in AI-generated

content, such as the gender discrimination caused by excessive use of female figures and biased art generation due to the abuse of popular contemporary artists' styles, limiting the diversity of creation.

These ethical risks require more transparent and regulated data usage in the AIGC era. On one hand, legislation should keep up with the rapid development of generative AI. Particularly, the legal definition of what constitutes a misuse of personal or intellectual property data needs to adapt to AI's increasing capacity for learning and generation. On the other hand, advanced data security technology needs to be developed as a countermeasure. In this regard, both preventive and reactive measures need to be developed. For example, one preventive method [49] is to add toxic noises to the original images, designed to be visually negligible yet sufficient to mislead generative models with a substantial divergence in the semantic interpretation of the content. Regarding reactive measures, researchers are expected to develop more robust models to dissect the distribution differences between generative and real data, enabling discrimination and effective data governance [8, 14].

7.2 Bridging AI and Human Creativity in Commercial and Artistic Domains

Implications on commercial applications. While many users happily adopt easy-to-use platforms like Midjourney [6], our comprehensive tool innovatively merges fine-tuning and generation into an intuitive workflow, offering enhanced customization for both creative and commercial endeavors. For example, in the illustration industry, artists often receive commissions to create characters for animations or video games. Simply producing a single depiction of a character is insufficient. There's a need for multiple renderings of the same character from varied perspectives, in different attire and accessories, and set against diverse backdrops. Creating more images of the same character from different angles, in other clothes and accessories, or even different environments is also important. However, it is time-consuming for artists to finish all the drawings manually. Our tool allows them to draw only a few images of their character and fine-tune their character model in subsequent AI generation for quick ideation and prototyping. Furthermore, our tool adeptly assists in fine-tuning specific product models in commercial contexts like product advertising. This enables generating a wide range of product images, enhancing their appeal on online shopping platforms. In addition, with the person's authorization, real human images can be used with our tool to support applications like virtual try-ons of fashion products or even make cross-generation photos or avatars showing elderly parents at their young age to facilitate intergenerational communication.

Bringing AI to the front of creative tools. Traditionally, many users of generative AI treat the AI as a ready-made tool behind the scenes and only expect to use prompting on a simple interface to get the results, such as in popular tools like Midjourney [6] and Adobe Firefly [5]. However, such a cooperation paradigm gives humans the illusion that they are controlling the AI even though AI is at the helm of the creative process. Particularly, the AI algorithms are responsible for generating the content based on patterns and data they have learned from. Users are superficially guiding the AI, but the AI is doing the heavy lifting. Our work has made an early step towards bringing generative AI to the foreground, stressing the importance of allowing users to control and customize the model. We achieve this goal by enabling users to intuitively teach AI new concepts and align them to user intents.

However, more work must advance a transparent and balanced co-creation paradigm between humans and AI. Many previous explainable AI systems are expert-oriented, but for generative AI, which has attracted a much broader population of users, including artists, designers, and other AI novices, this poses new challenges to strike a balance between easy-to-use prompt interface and complex training and evaluation system.

7.3 Limitations and Future Work

Considering extensibility. Some expert users commented that although the baseline tool is designed with an engineering mindset and includes too many complex settings, it is built upon the Gradio [7] library, which is widely used for the quick implementation of rudimentary interfaces for models on Hugging Face, as P11 suggested. A notable advantage of such a tool is that it allows for easy extension with plug-ins developed by the community. In comparison, our *IntentTuner* is a specialized system that has not considered the easy addition of new functions by community users. As the number of generative AI users continues to grow, so do their customization demands. This includes modifications not only to the fine-tuning process but also to the fine-tuning interface. In future work, we plan to incorporate extensible modules into our system that will allow users to add their own plug-in functions easily.

Increasing the trust in LLM-assisted task. We leverage a large language model to parse user intents. However, despite the powerful reasoning ability of LLM, users sometimes feel they cannot fully trust its interpretation of their intentions. The system should more intuitively display the intermediate results of LLM to allow for user adjustments instead of treating the LLM as a fail-proof panacea. Users commented that our multi-modal intention input, which allows users to connect the language to visual concepts, can increase their trust to some degree because they feel the language interpretation is grounded by visual information. In future work, we should develop more multi-modal visualization and interactions to allow users to view and refine AI interpretation of their intentions in fine-tuning.

Achieving more sophisticated evaluation. The diffusion models still entail high uncertainty, as different random seed inputs can result in diverse generations. Although our work incorporates human intentions into the evaluation, we have not addressed the uncertainty issue. Moreover, domain expert users, like artists and designers, may have more sophisticated requirements for aligning aesthetic preferences. For example, artists may want to preserve or change the color palette of the whole image instead of the hair color alone, or they may wish to control certain composition features. These complex aesthetic features are challenging to describe in language and evaluate in the semantic space. Particularly, the human preference metric used in our evaluation neglects the specific aesthetic aspects like color, composition, and stroke. Future work needs to develop metrics and visualization methods that account for the uncertainty and concrete aesthetic aspects to improve the evaluation further. For example, we could leverage some public datasets of fine-grained human-annotated aesthetic scores [39] to train more comprehensive aesthetic evaluation metrics.

8 CONCLUSION

This study presents *IntentTuner*, an intelligent framework to integrate human intentions in a novel but burgeoning AI-augmented creation task, which is customizing the text-to-image generative model with fine-tuning. The framework first allows users to articulate their fine-tuning intentions in natural multi-modal input, which is translated into structured intention specifications. Then, the intention specifications guide the data augmentation to optimize the fine-tuning data in alignment with users' intentions. Finally, *IntentTuner* incorporates intent-aligned evaluation to help users evaluate the fine-tuning results based on their specific intentions instead of using generic metrics. Based on the framework, we develop an integrated system that seamlessly combines the fine-tuning and generation functionalities to support a holistic and flexible workflow for text-to-image generation. Application examples and a user study show that our framework and system can effectively help users reduce the trial and error workload and increase the intention alignment in fine-tuning. These show great potential in expanding the steerability and accessibility of generative AI to a broader group of users.

ACKNOWLEDGMENTS

REFERENCES

[1] 2022. Civitai. https://civitai.com.

[2] 2022. Kohya's Stable Diffusion trainers. https://github.com/bmaltais/kohya_ss.

[3] 2022. LibLibAI. https://www.liblibai.com/.

[4] 2022. Stable Diffusion Web UI. https://github.com/AUTOMATIC1111/stable-diffusion-webui.

[5] 2023. Adobe Firefly. https://www.adobe.com/sensei/generative-ai/firefly.html

[6] 2023. Midjourney. https://www.midjourney.com/

[7] Abubakar Abid, Ali Abdalla, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. 2019. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569* (2019).

[8] Agil Aghasanli, Dmitry Kangin, and Plamen Angelov. 2023. Interpretable-through-prototypes deepfake detection for diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 467–474.

[9] Ali Borji, Ming-Ming Cheng, Qibin Hou, Huaizu Jiang, and Jia Li. 2019. Salient object detection: A survey. *Computational Visual Media* 5 (2019), 117–150.

[10] Stephen Brade, Bryan Wang, Mauricio Sousa, Sageev Oore, and Tovi Grossman. 2023. Promptify: Text-to-Image Generation through Interactive Prompt Exploration with Large Language Models. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–14.

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

[12] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems* 33 (2020), 1877–1901.

[13] Daniel Cohen-Or, Olga Sorkine, Ran Gal, Tommer Leyvand, and Ying-Qing Xu. 2006. Color harmonization. In *ACM SIGGRAPH 2006 Papers*. 624–630.

[14] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. 2023. On the detection of synthetic images generated by diffusion models. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[15] Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2021. 8-bit optimizers via block-wise quantization. *arXiv preprint arXiv:2110.02861* (2021).

[16] Tan M Dinh, Rang Nguyen, and Binh-Son Hua. 2022. TISE: Bag of metrics for text-to-image synthesis evaluation. In *European Conference on Computer Vision*. Springer, 594–609.

[17] Yingchaojie Feng, Xingbo Wang, Kam Kwai Wong, Sijia Wang, Yuhong Lu, Minfeng Zhu, Baicheng Wang, and Wei Chen. 2023. PromptMagician: Interactive Prompt Engineering for Text-to-Image Creation. *IEEE Transactions on Visualization and Computer Graphics* (2023), 1–11.

[18] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618* (2022).

[19] Rinon Gal, Or Patashnik, Haggai Maron, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. 2022. StyleGAN-NADA: CLIP-guided domain adaptation of image generators. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–13.

[20] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems* 30 (2017).

[21] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

[22] Nanna Inie, Jeanette Falk, and Steve Tanimoto. 2023. Designing Participatory AI: Creative Professionals' Worries and Expectations about Generative AI. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.

[23] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4401–4410.

[24] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6007–6017.

[25] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643* (2023).

[26] Hyung-Kwon Ko, Subin An, Gwanmo Park, Seung Kwon Kim, Daesik Kim, Bohyoung Kim, Jaemin Jo, and Jinwook Seo. 2022. We-toon: A Communication Support System between Writers and Artists in Collaborative Webtoon Sketch Revision. In *Proceedings of the Annual ACM Symposium on User Interface Software and Technology*. 1–14.

[27] Zihang Lan, Shuhan Yang, Rui Fan, Bo Zhao, and Yanru Yan. 2023. Innovation or Piracy? Empirically Demarcating AI Painting Copyright Infringement Boundary. In *2023 3rd International Conference on Public Management and Intelligent Society (PMIS 2023)*. Atlantis Press, 1328–1341.

[28] Feng Li, Hao Zhang, Peize Sun, Xueyan Zou, Shilong Liu, Jianwei Yang, Chunyuan Li, Lei Zhang, and Jianfeng Gao. 2023. Semantic-SAM: Segment and Recognize Anything at Any Granularity. *arXiv preprint arXiv:2307.04767* (2023).

[29] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597* (2023).

[30] Yaobo Liang, Chenfei Wu, Ting Song, Wenshan Wu, Yan Xia, Yu Liu, Yang Ou, Shuai Lu, Lei Ji, Shaoguang Mao, et al. 2023. Taskmatrix. AI: Completing tasks by connecting foundation models with millions of apis. *arXiv preprint arXiv:2303.16434* (2023).

[31] Miao Liu and Yifei Hu. 2023. Application potential of stable diffusion in different stages of industrial design. In *International Conference on Human-Computer Interaction*. Springer, 590–609.

[32] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. 2023. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499* (2023).

[33] Vivian Liu. 2023. Beyond Text-to-Image: Multimodal Prompts to Explore Generative AI. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–6.

[34] Vivian Liu and Lydia B Chilton. 2022. Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–23.

[35] Vivian Liu, Han Qiao, and Lydia Chilton. 2022. Opal: Multimodal image generation for news illustration. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–17.

[36] Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983* (2016).

[37] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. 2022. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11461–11471.

[38] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the Role of Demonstrations: What makes In-context Learning Work?. In *EMNLP*.

[39] Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. AVA: A large-scale database for aesthetic visual analysis. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2408–2415.

[40] pharmapsychotic. 2023. GitHub. https://github.com/pharmapsychotic/clip-interrogator.

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

[42] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *Arxiv* abs/2204.06125 (2022).

[43] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*. 8821–8831.

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10684–10695.

[45] Steven I Ross, Fernando Martinez, Stephanie Houde, Michael Muller, and Justin D Weisz. 2023. The programmer's assistant: Conversational interaction with a large language model for software development. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 491–514.

[46] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. 2023. 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*. 22500–22510.

[47] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems* 35 (2022), 36479–36494.

[48] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *Advances in Neural Information Processing systems* 29 (2016).

[49] Shawn Shan, Jenna Cryan, Emily Wenger, Haitao Zheng, Rana Hanocka, and Ben Y Zhao. 2023. Glaze: Protecting artists from style mimicry by text-to-image models. *arXiv preprint arXiv:2302.04222* (2023).

[50] Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving ai tasks with chatgpt and its friends in huggingface. *arXiv preprint arXiv:2303.17580* (2023).

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

[52] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[53] Mathias Peter Verheijden and Mathias Funk. 2023. Collaborative Diffusion: Boosting Designerly Co-Creation with Generative AI. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–8.

[54] Yael Vinker, Andrey Voynov, Daniel Cohen-Or, and Ariel Shamir. 2023. Concept Decomposition for Visual Exploration and Inspiration. *arXiv preprint arXiv:2305.18203* (2023).

[55] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. 2023. Exploring clip for assessing the look and feel of images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 2555–2563.

[56] Yunlong Wang, Shuyuan Shen, and Brian Y Lim. 2023. RePrompt: Automatic Prompt Editing to Refine AI-Generative Art Towards Precise Expressions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–29.

[57] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems* 35 (2022), 24824–24837.

[58] Mika Westerlund. 2019. The emergence of deepfake technology: A review. *Technology innovation management review* 9, 11 (2019).

[59] Di Wu, Zhiwang Yu, Nan Ma, Jianan Jiang, Yuetian Wang, Guixiang Zhou, Hanhui Deng, and Yi Li. 2023. StyleMe: Towards Intelligent Fashion Generation with Designer Style. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. 1–16.

[60] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding* 163 (2017), 21–40.

[61] Xiaoshi Wu, Keqiang Sun, Feng Zhu, Rui Zhao, and Hongsheng Li. 2023. Better Aligning Text-to-Image Models with Human Preference. *ArXiv* abs/2303.14420 (2023).

[62] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. 2023. Imagereward: Learning and evaluating human preferences for text-to-image generation. *arXiv preprint arXiv:2304.05977* (2023).

[63] Jingfeng Yao, Xinggang Wang, Lang Ye, and Wenyu Liu. 2023. Matte Anything: Interactive Natural Image Matting with Segment Anything Models. *arXiv preprint arXiv:2306.04121* (2023).

[64] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. 2022. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789* 2, 3 (2022), 5.

[65] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. 2023. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790* (2023).

[66] Chunhui Zhang, Li Liu, Yawen Cui, Guanjie Huang, Weilin Lin, Yiqian Yang, and Yuehong Hu. 2023. A Comprehensive Survey on Segment Anything Model for Vision and Beyond. *arXiv preprint arXiv:2305.08196* (2023).

[67] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. In *IEEE International Conference on Computer Vision*.

[68] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. 2023. Inversion-based style transfer with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10146–10156.

[69] Chenliang Zhou, Fangcheng Zhong, and Cengiz Öztireli. 2023. CLIP-PAE: Projection-Augmentation Embedding to Extract Relevant Features for a Disentangled, Interpretable and Controllable Text-Guided Face Manipulation. In *ACM SIGGRAPH 2023 Conference Proceedings*. 1–9.

A HYPER-PARAMETERS SETTING OF LORA

Table 1 illustrated the default settings of model training hyperparameters. We choose 8-bit AdamW [15] as the optimizer and use cosine annealing with warm restarts [36] as the learning rate (LR) scheduler. Other hyper-parameters are preset based on the training domain translated from the user intents (Sect. 4.1). Stable diffusion consists of a U-net and a text encoder, with different learning rates during training. The LoRA module has two essential parameters: Dimension and Alpha [21]. Dimension is the size of the low-rank update matrices, which determines the number of trainable parameters. Alpha is a scaling factor that affects weight updates. Weight updates become more aggressive when $\frac{alpha}{rank}$ is set to a higher value.

Table 1. Detailed setting of training hyperparameters

Domain	U-net LR	Text encoder LR	Dimension	Alpha
Painting	1e-4	1e-5	64	32
Human portrait	1e-4	5e-5	128	64
2D character	1e-4	1e-5	32	32
Product	1e-4	5e-5	64	32