Stellar: Systematic Evaluation of Human-Centric Personalized Text-to-Image Methods

Supplementary Material

Panos Achlioptas **Alexandros Benetatos** **Iordanis Fostiropoulos** **Dimitris Skourtis**

Steel Perlot Management LLC
Los Angeles, California, USA

Updated on 2023.12.07

1. Stellar Data

In this section, we provide further details on the curation process of Stellar, our proposed dataset for grounding and evaluating modern text-to-image (T2I) generation methods. Additionally, we provide and analyze both quantitatively and qualitatively the complementarity and the similarities among Stellar-$\mathcal{H}$ and Stellar-$\mathcal{T}$, highlighting the particularities of each prompt dataset.

All introduced datasets, our evaluation metrics and network, StellarNet, **will be made publicly available**, along with a formal competition, fostering progress and comparisons for newly introduced T2I personalized methods.

1.1. Stellar-$\mathcal{H}$

In our effort to create a long-standing benchmark for human-centric personalized image generation, we recognized the importance of acquiring *realistic* and *natural* prompts for this task, i.e., similar to what we can expect a *real non-expert* user to provide to a powerful personalized T2I generation system, which is is not bound by transient technological shortcoming [16]. E.g., it does not require intricate prompt-engineering to deliver accurately grounded outputs. To promote such open-ended, but also, simple (natural) prompt curation, we initiated an Amazon Mechanical Turk (AMT) study with the aim of collecting a total of *10k* prompts from multiple English-speaking annotators.

As illustrated in Fig. 1, participants of our AMT study were instructed to provide three prompts, at a time, assuming access to a system capable of generating *arbitrary* and *imaginative* images of themselves. Importantly, we guided participants to avoid prompts that sought to *explicitly* alter their physical attributes (e.g., age), request specific clothing or accessories (e.g., jewelry), or prompts that included other individuals, such as family members for which a single-image typical T2I system, is impossible to visually comprehend without additional information. Besides the above thee specific constraints, the annotators where free to describe arbitrary imaginative scenarios in natural language.

This effort resulted in the collection of 3376 submissions, each containing three prompts, from a total of 123 unique users. Importantly, in our study could only participate users with an *excellent* track-record in terms of having successfully completed (each) more than 5,000 thousand submissions in other visiolinguistic tasks with error-rates of less than 0.5%. In the end, upon applying a manually-curated content-filter to exclude potentially sensitive themes (e.g., violence or nudes), and performing minimal post-processing, we obtained the final dataset of $10k$ prompts, denoted as Stellar-$\mathcal{H}$. Detailed statistics regarding the length of its prompts and the distribution of common parts-of-speech and overall token-usage can be found in Tab. 3 - left.

1.2. Stellar-$\mathcal{T}$

For the more standardized Stellar-$\mathcal{T}$, we drew inspiration from the imaginative prompts of Stellar-$\mathcal{H}$, while imposing vocabulary constraints. This action aims to enhance the precision in evaluating personalized models, particularly their ability to faithfully represent the objects and relationships described in the prompts.

Our first step involved a detailed manual analysis of Stellar-$\mathcal{H}$. This examination revealed that the majority of prompts conform to a flexible template: `"S* [as a person in uniform] [engaging in an activity] [at a location] [at a specific time and under certain weather conditions]"`, with each bracketed component being optional.

Next, we extracted nouns and verbs from these prompts as proxies for objects and actions. Our analysis focused on the frequency of these elements, leading to the incorporation of the most recurrent ones into Stellar-$\mathcal{T}$. For related object groups, like food, we adopted their broader class categories.

To diversify Stellar-$\mathcal{T}$, we enlisted ChatGPT's aid in identifying the top 20 most recognizable items within these categories. These recommendations were largely integrated into our dataset enriching Stellar-$\mathcal{T}$ with a wide range of objects. Examples of the final object categories from Stellar-$\mathcal{T}$ can be found in Tab. 2.

Our analysis of verbs led to the development of the ac-

Instructions MUST READ at least once.

Assume you have access to a machine that can generate **arbitrary and imaginative** photos portraying yourself. To do this, this machine requires one photo of yourself and a text describing what you want the *output* image to look like.

Your task is to provide us with *three short* texts that show how you would use such a machine to make imaginative/creative photos of yourself.

RULES

1. You are free to imagine almost whatever you want for how your output image should look. E.g., you can depict yourself in arbitrary places (e.g., at a dinning room, in Paris, on Mars), doing random things (e.g., swimming, singing, relaxing), interacting with arbitrary items (e.g., swimming with a dolphin, singing in a jazz band, etc.).

   JUST a TINY SYBSET of GOOD EXAMPLES:

   myself: dancing with a bear in Alaska, myself: playing soccer at a huge stadium at night, myself: as a firefighter receiving an award, myself: watching the sunset in Santorini

   If you are curious, you can see a few examples of input text/images and output generation here (link)

2. Things you should **NOT** explicitly specify in your text:

   - Changes that ask to alter you age, skin color, weight, height, or other similar physical characteristics
   - requests that ask to wear explicitly a set of clothes, eyewear, jewelry, or footwear
   - to co-appear with specific individuals such as your family or friends

   Following these rules, here are some **BAD EXAMPLES**:

   - Myself wearing a red dress at the Eiffel Tower [don't ask for explicit changes of clothes]
   - Myself, but taller and with a bigger nose [don't ask for changes of your physical characteristics]
   - Myself celebrating with my family and children [don't ask to co-appear with specific individuals]

3. If you are not a **proficient** English speaker, do not accept this HIT.

4. Do **NOT** submit more thatn 10 HITS in this batch

Figure 1. **Instructions given to annotators for collecting prompts in natural language.** These are the mandatory instructions AMT annotators had to follow in order to crowd-source Stellar-$\mathcal{H}$. The partaking annotators were selected from a pool of annotators with *excellent* track-record in similar tasks, and where able to communicate directly with the authors for any follow-up questions. Their work resulted in a semantically rich and imaginative dataset of *10k* human-generated prompts.

Table 1. **Stellar-$\mathcal{H}$ examples and their nearest neighbors in Stellar-$\mathcal{T}$ according to the ST5 embedding space.** In each group of prompts, the top prompt (in purple ◯) is randomly chosen from Stellar-$\mathcal{H}$, while the 5 following prompts (in ◯) are its 5 nearest neighbors in Stellar-$\mathcal{T}$ in the Sentence-T5 [10] embedding space. We observe that frequently, Stellar-$\mathcal{T}$ prompts offer dense but also substantial variations of Stellar-$\mathcal{H}$ (e.g., *giving a speech next to the Statue of Liberty*) while preserving critical semantics (top group, all actions are grounded on the Statue of Liberty).

Stellar
S^* in front of the Statue of Liberty
$\vdash$ S^* at the Statue of Liberty
$\vdash$ S^* running in front of the Statue of Liberty
$\vdash$ S^* giving a speech next to the Statue of Liberty
$\vdash$ S^* wearing glasses at the Statue of Liberty
$\vdash$ S^* dining near the Statue of Liberty
S^* bungee jumping off a hot air balloon at sunrise
$\vdash$ S^* bungee jumping in the sunrise
$\vdash$ S^* bungee jumping in the morning
$\vdash$ S^* sky diving in the sunrise
$\vdash$ S^* bungee jumping in an air balloon
$\vdash$ S^* bungee jumping on a sunny day
S^* making a rice dinner
$\vdash$ S^* cooking rice
$\vdash$ S^* holding rice
$\vdash$ S^* cooking rice in the morning
$\vdash$ S^* eating rice
$\vdash$ S^* as an admiral cooking rice
S^* holding a bear on a leash
$\vdash$ S^* taking a bear for a walk
$\vdash$ S^* taking a bear for a walk in the snow
$\vdash$ S^* with a bear
$\vdash$ S^* next to a bear
$\vdash$ S^* taking a tiger for a walk
S^* climbing a rock wall as a nurse
$\vdash$ S^* as a nurse rock climbing
$\vdash$ S^* as a nurse climbing a mountain
$\vdash$ S^* as a nurse bungee jumping
$\vdash$ S^* as a nurse weightlifting
$\vdash$ S^* as a nurse running

tion templates used in Stellar-$\mathcal{T}$, such as "eating [food]", and "driving [vehicle]". We categorized similar actions into clusters (e.g., "doing sports") and, like in previous steps, utilized ChatGPT's assistance to populate these categories with relevant actions.

For the `[person in uniform]` component, we merged prevalent themes from Stellar-$\mathcal{H}$ with suggestions from ChatGPT, resulting in a comprehensive list of *people in uniform*. For location elements, we expanded upon Stellar-$\mathcal{H}$'s common locales, incorporating a list of renowned cities, countries, and monuments suggested by ChatGPT. Additionally, for the `[at a specific time and under certain weather conditions]` section, we included various times of day (e.g., morning, sunset), seasons (e.g., spring), and weather phenomena (e.g., fog).

Finally, to improve the diversity of our template's actions, locations, and objects we received feedback from in-

Food	Famous Landmark	Person in Uniform	Sports	Famous Car Brand	Nature Loc.
pizza	Golden Gate Bridge	clown	basketball	Bugatti	desert
steak	Mount Rushmore	businessman	tennis	Rolls-Royce	forest
sushi	Victoria Falls	doctor	baseball	Lamborghini	mountain
dinner	Niagara Falls	nurse	soccer	Ferrari	sea
pasta	Buckingham Palace	cop	golf	Aston Martin	lake
noodles	La Sagrada Familia	policeman	volleyball	Bentley	river
corn	Grand Palace	firefighter	rugby	Porsche	grass
popcorn	Blue Domes of Oia	fireman	football	McLaren	beach
croissant	Mount Fuji	soldier	ice hockey	Pagani	volcano
ice cream	Sydney Opera	captain	table tennis	Koenigsegg	fire
soup	White House	admiral	badminton	Maserati	**Sports**
eggs	Parthenon	scientist	cricket	Mercedes-Benz	basketball
rice	Eiffel tower	knight	**Music Instr.**	Lexus	tennis
potato chips	Pisa tower	DJ	guitar	Audi	baseball
tacos	Pyramids of Giza	Egyptian pharaoh	piano	Jaguar	soccer
hamburger	Great Pyramid of Giza	king	cello	Alfa Romeo	golf
cheeseburger	Statue of Liberty	emperor	violin	Tesla	volleyball
curry	Taj Mahal	astronaut	flute	Lotus	rugby
paella	Great Wall Of China	cowboy	bass	Zenvo	football
falafel	Petra of Jordan	wizard	horn	Rimac	ice hockey
goulash	Colosseum	pilot	drums	**Board Games**	table tennis
pad thai	Machu Picchu	President	harp	chess	badminton
kebab	Stonehenge		mandolin	cards	cricket
souvlaki	Acropolis		trumpet	poker	
hot dog	Brandenburg Gate		oboe	monopoly	
cake			saxophone	uno	

Table 2. **Example categories and associated objects explicitly annotated in Stellar-$\mathcal{T}$.** The shown categories are coupled with our sentence-producing templates to sample relevant objects (per category) and construct the candidate prompts covering Stellar-$\mathcal{T}$. Applying a template-based generation promotes a robust and fine-grained evaluation of personalization T2I systems given the implicit rich annotation we can extract in terms of prompt-grounding objects and their interaction with the subject human. For presentation purposes, we showcase categories with a relatively small number of underlying objects.

Figure 2. **Wordcloud of nouns in Stellar-$\mathcal{H}$.** The size of each word is proportional to its relative frequency in the underlying corpus. For best viewing results please use a digital version and zoom in.

ternal teams in our company.

Filtering to obtain the final *10k* prompts set. This initial procedure generated approximately *177k* prompts, a number that can be further scaled up by incorporating additional objects, locations, and actions into the templates. To refine this extensive collection down to a more manageable set of *10k* prompts, we first employed a filtering step. This involved discarding one-third of the prompts based on their semantic dissimilarity compared to those in Stellar-$\mathcal{H}$. Subsequently, to ensure the remaining set was semantically diverse we applied a furthest point sampling technique. The

Table 3. **Lexical analysis of Stellar-$\mathcal{H}$ and Stellar-$\mathcal{T}$.** We report the number of prompts, unique tokens, and machine-generated statistics based on Part Of Speech tagging [1]. For each category, we report the total elements (*TT*), the average number of elements per prompt (*PP*), and the average number of unique elements per coupled image (*PI*), in either dataset – each image is assigned 50 prompts. Note how, despite the higher diversity of Stellar-$\mathcal{H}$ shown in *TT*, the *PI* and *PP* of the two subsets are very similar, underscoring a similar complexity and diversity on each image-prompt example.

Statistic	Stellar-$\mathcal{H}$			Stellar-$\mathcal{T}$		
	TT	PP	PI	TT	PP	PI
prompts	$10k$	1	50	$10k$	1	50
tokens	$6.4k$	7.7	395.0	533	7.1	354.9
nouns	$4.5k$	3.1	96.4	368	2.7	64.4
verbs	$1.6k$	1.3	48.9	94	1.1	30.4
adjectives	827	0.3	13.8	59	0.0	6.38
adpositions	98	1.3	17.2	33	1.1	10.5

resulting collection of prompts is Stellar-$\mathcal{T}$.

Similarities and Differences of Stellar-$\mathcal{H}$ and Stellar-$\mathcal{T}$. Our standardization efforts are reflected in the object and verb statistics of the two Stellar datasets, as detailed in Tab. 3. A key observation is the similarity in the average number of part-of-speech (POS) elements per prompt and image. These statistics demonstrate that we have successfully maintained comparable levels of complexity and diversity in the image-prompt pairings across both datasets. Additionally, an aspect of our standardization process is the significant reduction of unnecessary modifiers, highlighted by the near-zero count of *adjectives* in Stellar-$\mathcal{T}$.

Moreover, when merging the objects from popular object detection datasets like Open Images [7], COCO [2] and Objects365 [15], we observe that these objects cover *2.6x* higher percentage of the vocabulary in Stellar-$\mathcal{T}$ compared to Stellar-$\mathcal{H}$ (37% vs. 14%). A similar trend is observed in the relationships within these datasets when contrasted with VG200 [18] and VRD [9] datasets (35% vs. 2%).

Despite these differences, both qualitative (Fig. 3) and quantitative (Tab. 4) analyses suggest a high degree of semantic similarity between Stellar-$\mathcal{T}$ and Stellar-$\mathcal{H}$ with Stellar-$\mathcal{T}$'s advantage being on the capability for more robust evaluation of a personalization T2I system as we discuss in Sec. 1.2. For comparison purposes, we include COCO [2], a commonly used captioning dataset. For these comparisons, we employed a tailored subset of COCO, denoted as COCO*, which includes only human-centric content. We analyzed *10k* samples from each dataset using Sentence T5 (ST5) [10] embeddings. Qualitatively, these embeddings are visualized in the t-SNE space (Fig. 3).

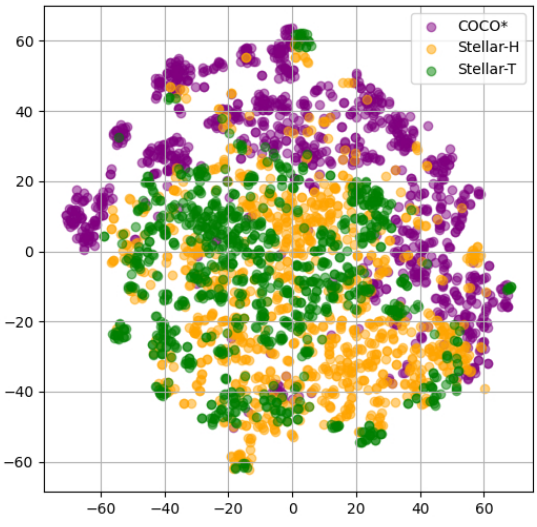

Figure 3. **Visually contrasting the semantic proximity between Stellar-$\mathcal{H}$, Stellar-$\mathcal{T}$ and COCO*.** Here, we project and visualize ST5 embeddings projected in two dimensions via t-SNE for equally-sized subsampled prompts of the aforementioned datasets. We observe a big overlap (clustering) between Stellar-$\mathcal{H}$ and Stellar-$\mathcal{T}$ indicative of their semantic proximity. On the other hand, there is a distinct separation of both datasets from COCO*, highlighting their uniqueness from such objective captions. (COCO* is a subset of captions of COCO that explicitly refer to people, with more than $100k$ caption examples).

Table 4. **Quantitative semantic similarity of the datasets.** This analysis showcases the semantic similarities between Stellar-$\mathcal{T}$ and Stellar-$\mathcal{H}$, and their distinction from existing captioning datasets such as COCO. A Gaussian Mixture Model (GMM) trained on Sentence-T5 embeddings from each prompt sample in Stellar-$\mathcal{H}$ was used to calculate the average log-likelihood of the embeddings from Stellar-$\mathcal{T}$ and COCO* being part of the Stellar-$\mathcal{H}$ distribution. Despite Stellar-$\mathcal{T}$ being more constrained in scope than Stellar-$\mathcal{H}$, the data reveals a notable semantic alignment between them. In contrast, COCO* though richer in prompt quantity, exhibits a different distribution from that of Stellar. This distinction underscores the unique semantic composition of the Stellar datasets.

Dataset	**Stellar-$\mathcal{H}$**	**Stellar-$\mathcal{T}$**	**COCO***
Stellar-$\mathcal{H}$	1869	1858	1765

To quantitatively compare the datasets, we employ an optimal Gaussian Mixture Model (GMM), as determined by the BIC score, trained on the ST5 embeddings from each prompt of Stellar-$\mathcal{H}$. By calculating the average log-likelihood of the embeddings from Stellar-$\mathcal{T}$ and COCO*

being part of this distribution we get a quantitative measure of the similarity of these datasets (Tab. 4). Underscoring the semantic similarity between Stellar-$\mathcal{H}$ and Stellar-$\mathcal{T}$ is also the evaluation of the ablated methods (Sec. 6) in both datasets. Notably, comparing Tab. 5 and Tab. 3, we observe that the metrics are nearly identical when evaluating either of the subsets of Stellar.

Finally, we estimated the percentage of Stellar-$\mathcal{H}$ prompts that can be covered by the templates in Stellar-$\mathcal{T}$. Specifically, we sampled 200 random prompts from Stellar-$\mathcal{H}$ and manually evaluated if there was a specific template covering each of them. For example "S* watching a basketball match in London" can be effectively captured by the template "S* watching [event] in [city]". However, "S* kicking a football with their elbow" can not be effectively captured by any templates in Stellar-$\mathcal{T}$, even the relatively similar "S* kicking [kickable objects]" as it can not capture the semantic nuance the modifier "with their elbow". With this in mind, we find that we can cover a least 75% of Stellar-$\mathcal{H}$ with templates from Stellar-$\mathcal{T}$.

The outcomes of both qualitative and quantitative analyses consistently indicate that the two subsets of Stellar—Stellar-$\mathcal{H}$ and Stellar-$\mathcal{T}$—exhibit a high degree of semantic similarity. In contrast, they are markedly distinct from the COCO* dataset, underscoring the unique nature of the Stellar dataset in comparison to traditional captioning datasets.

1.3. Stellar-Images

The image portion of the STELLAR dataset contains a subsample of the CelebAMask-HQ's test set. CelebAMask-HQ contains close-up images of celebrities with several pictures per subject and annotations on characteristics specific to the image, i.e.,whether the subject is a male or female. We construct our image dataset by first employing a state-of-the-art (SoTA) face detector [3] to extract the bounding box of each face. Our criteria for inclusion required that the detected face must cover at least 20% of the image area and that the model's confidence level in the detection must be no less than 99%. Furthermore, we excluded any images where their identity in the original CelebAMask had inconsistent gender annotation for the same person (e.g., a Celeb identity marked as `female` in one instance and `male` in another). Additionally, we eliminated identities represented by fewer than two images within the dataset. We then randomly selected 200 subjects and ensured diversity by selecting a balanced gender (`woman`/`man`) and age (`old`/`young`) using the annotations in CelebA. From each chosen identity, only two images were randomly selected for inclusion. Consequently, this process resulted in a collection of 400 high-quality human face images representing 200 unique identities.

Similarly, we applied this procedure to the validation set of CelebAMask-HQ, selecting another set of 200 unique identities with two images each, thereby forming the validation split of our dataset. We use this split for model selection and validation purposes.

2. Details on Evaluation Metrics

2.1. Stellar-$\mathcal{T}$ usefulness to our Metrics

A simplified visual representation of our proposed metrics is provided in Fig. 4. This illustration also underscores the significance of Stellar-$\mathcal{T}$ as its annotations are necessary for the calculation of the object-centric metrics `GOA` and `RFS`. Additionally, the `SIS` score necessitates the use of additional images of the input identity. This requirement is conveniently met by the Stellar dataset, thereby reinforcing its utility.

2.2. Metrics Correlation

The extensive correlation matrix depicted in Fig. 5 reveals insightful patterns among various existing and newly introduced metrics in personalized image generation evaluation. Two distinct clusters emerge from this analysis: one representing text-to-image evaluation metrics ( ) and the other encompassing our identity-based metrics ( ).

Focusing on the text-to-image metrics, highlighted in Fig. 5 with a circle, a notable high correlation is observed among these metrics. For instance, the Pearson correlation coefficient ρ reaches 0.6 between CLIP$_T$ and HPSv2. Interestingly, these metrics also exhibit a strong correlation with `GOA`, our metric designed to assess object faithfulness between the prompt and the generated image. This trend suggests that traditional text-to-image metrics predominantly emphasize the faithfulness of representing objects rather than the relationships between them. This observation further underscores the need for specialized relation metrics, such as `RFS`. As indicated in the figure, `RFS` introduces a distinct evaluative dimension, showing minimal correlation with other metrics and thereby providing a unique perspective in the assessment of image generation quality.

In the segment of identity-based metrics (Fig. 5 -), our three proposed metrics—`IPS`, `APS`, and `SIS`—form a tightly correlated cluster. Despite their high inter-correlations, these metrics exhibit minimal correlation with other metrics, reinforcing their significance in introducing new evaluative dimensions for personalized text-to-image generation. It is crucial to note, however, that the strong correlations observed among `IPS`, `APS`, and `SIS` should not be misconstrued as redundancy. Each metric addresses distinct aspects of identity representation: `SIS` focuses on the consistency of identity generation, while `APS` provides a fine-grained assessment of identity preservation, complementing the more holistic approach of `IPS`. This comple-

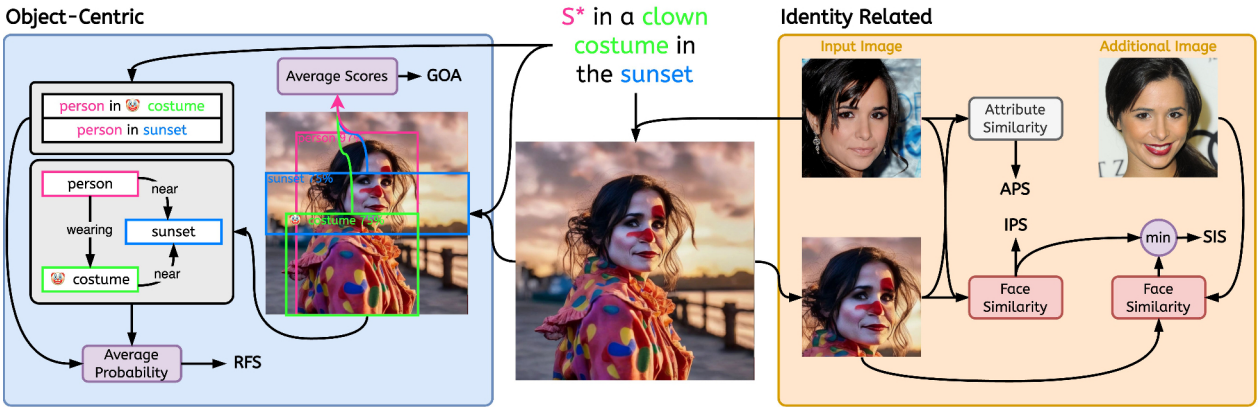

Figure 4. **Graphical depiction of our proposed evaluating pipeline and metrics for personalized T2I generations.** The abstract illustration underscores the significance of Stellar-$\mathcal{T}$ as its annotations are necessary for the calculation of the object-centric metrics GOA and RFS (left-most rectangle). Moreover, the SIS score necessitates the use of *additional* images for evaluating the output's fidelity to the input identity, and the APS requires specialized physical/facial attributes (right-most rectangle). These requirements are conveniently met by Stellar's image dataset, thereby reinforcing its utility.

mentarity is particularly evident in cases where IPS might not capture subtle nuances, as seen in the 2nd row of Fig. 3.

2.3. Human Evaluation

To ground the assessment of personalized image generation models in human perception and to quantitatively determine the correlation of our proposed metrics with human judgment, we conducted two human-centered studies with a total of $\sim$*2.5k*. The methodologies and findings of these studies are presented in Sec. 6.1. Here we provide additional information about each of the studies.

***Overall* Study.** In our *Overall* study, participants were presented with four different outputs generated using four distinct ablation methods: StellarNet, ELITE*, Dreambooth, and Textual Inversion. They were asked to select the one they preferred most overall, as shown in Fig. 6. This preference was based on how well each output simultaneously adhered to the given prompt and accurately represented the input image. In total, we sent evaluators 500 input-output pairs, gathering *1.5k* responses. Upon analyzing these responses, we retained those input-output pairs where a majority consensus was reached, amounting to 95% of the total queries. This high rate of consensus highlights a significant agreement among human evaluators in their preferences.

The results showed a notable 78% preference for StellarNet. More crucially, there was a significant Kendall-τ correlation score of 0.45 between human judgment and the IPS score, the highest correlation observed among all the ablated metrics. This outcome underscores the paramount importance of input identity representation in aligning with human evaluators' preferences, provided there is at least a minimal alignment with the prompt.

Second Study. The focus of our second study was to assess the effectiveness of GOA and RFS in independently evaluating the faithfulness of representing objects and their relations in generated images, as described in the input prompts. To this end, human evaluators were engaged with two distinct tasks (Fig. 7). Again, we presented evaluators with 500 input-output pairs, garnering *1k* responses. To ensure the reliability of our findings, we considered only those queries where the majority of evaluators reached a consensus in their responses, ending up with 70% of the 500 queries.

Firstly, they were presented with two images and asked to determine which image (if any) more accurately depicted the objects mentioned in the prompt. Secondly, evaluators were tasked with identifying which of the two images (if any) more effectively captured the interactions between humans and objects, as described in the prompt.

The study's findings were significant: there was a *1.4x* increase in the Kendall-τ correlation between GOA and human judgment for object faithfulness, compared to the best existing text-to-image metric. Likewise, the correlation between RFS and human judgment on relationship accuracy showcased a *1.5x* improvement over the best existing text-to-image metric. These results highlight the superior capability of GOA and RFS in aligning with human evaluators' perceptions of object and relationship accuracy in generated images.

3. StellarNet Architecture & Training Details

StellarNet architecture draws inspiration from various pre-existing works in the personalized T2I generation space. As

Table 5. **Quantitative evaluation of StellarNet against popular SoTA using existing and introduced metrics on Stellar-$\mathcal{H}$.** In the Main paper we evaluate all methods on our purposed metrics on Stellar-$\mathcal{T}$ Tab. 3. In this table we present results for the same evaluation on Stellar-$\mathcal{H}$. We can observe that the metrics are similar to the ones in Tab. 3. Our results highlight the similarity between the two datasets and the benefit of using Stellar-$\mathcal{T}$, where ground-truth object annotations allow one to evaluate object-centric metrics i.e., `RFS` and `GOA`.

	Models				
Metrics	**DreamBooth** [14]	**ELITE*** [17]	**Text. Inv.** [4]	**StellarNet** (Ours)	*Type*
Aesth. (↑)	5.250	5.066	5.214	**5.641**	
CLIP$_I$ (↑)	0.304	0.374	0.468	**0.521**	Img-to-Img
DreamSim (↓)	0.786	0.704	0.615	**0.566**	
CLIP$_T$ (↑)	**0.404**	0.369	0.313	0.378	
HPSv1 (↑)	0.198	0.193	0.189	**0.204**	
HPSv2 (↑)	0.271	0.267	0.262	**0.274**	Text-to-Img
ImageReward (↑)	-0.038	-0.409	-0.913	**0.423**	
PickScore (↑)	0.211	0.205	0.198	**0.213**	
APS (↑)	0.299	0.449	0.419	**0.685**	
IPS (↑)	0.246	0.368	0.299	**0.622**	Personalized (Ours)
SIS (↑)	0.228	0.342	0.273	**0.564**	

explained in Sec. 5 of the Main paper, our method consists of the Dynamic Textual Inversion (`DTI`) module and learnable LoRA offset weights that steer the UNet backbone of the pre-trained model we build upon. An overview of the architecture can be seen in Fig. 8.

Dynamic Textual Inversion. `DTI` deals with projecting the input image from the pixel space to the textual word embedding space. We call the result of this projection $\mathbf{S}^*$. For our architecture, we provide the `DTI` module with masked input images, where all the background pixels have been zeroed out. Visually, this can be seen on the left part of Fig. 8.

The module is versatile enough to be integrated with multi-encoder text-to-image models such as SDXL [11]. In this setup, there are two possible approaches: one option is to provide the same word embedding $\mathbf{S}^*$ to all encoders. Alternatively, we can generate distinct sets of embeddings $\mathbf{S}^*_\mathbf{i}$ tailored for each text encoder. Furthermore, these embeddings can either be derived from the same Image Encoder but processed through different MLP mappers, or we can opt for separate image encoders, each corresponding to an individual text encoder. This modularity of `DTI` allows for flexible and efficient integration with various text-to-image model architectures.

Drawing inspiration from ELITE [17], our model constructs $\mathbf{S}^*$ as a series of embeddings, each derived using a distinct image2text mapper for various layers of the image encoder. These embeddings encapsulate varying levels of detail. In practice, although multi-word embeddings are learned, we primarily utilize the one associated with the final layer of the image encoder during inference. This selective use helps to isolate key identity features from higher frequency details (e.g.,the background), thereby enhancing the model's editing capability.

Low Rank Adaptation Weight Offsets. Beyond employing `DTI`, we refine our model by fine-tuning the pre-trained conditional UNet backbone to more effectively interpret the newly formed $\mathbf{S}^*$ embeddings. However, in order to successfully fine-tune the UNet, we would need a very large amount of data and careful tuning of the training hyperparameters. Thus, for efficient and stable training, particularly with large models like SDXL, we utilize Low Rand Adaptation (LoRA [5]) weight offsets to prevent model collapse. Moreover, these weight offsets open avenues for future exploration, such as introducing an auxiliary parameter, λ. This parameter can modulate the influence of the fine-tuned weights, thereby offering a mechanism to balance between editability and accurate subject representation. Visually, this can be seen in the middle part of Fig. 8.

Loss. To complement the expectation for masked images in the `DTI` module, we formulate and use a masked Mean Square Error (MSE) loss function. This loss function is designed to avoid penalizing the model for variations in background details compared to the input image. Instead, it focuses solely on the non-masked pixels, as determined by the mask provided in the `DTI` module. The formula for this masked MSE loss is as follows:

$$MSE_{masked} = \frac{1}{N}\sum MSE_{2D}(I, O) \cdot \text{MASK}_{bin} \quad (1)$$

Here, MSE_{2D} represents the Mean Square Error, before the final mean redaction, calculated in the latent space and

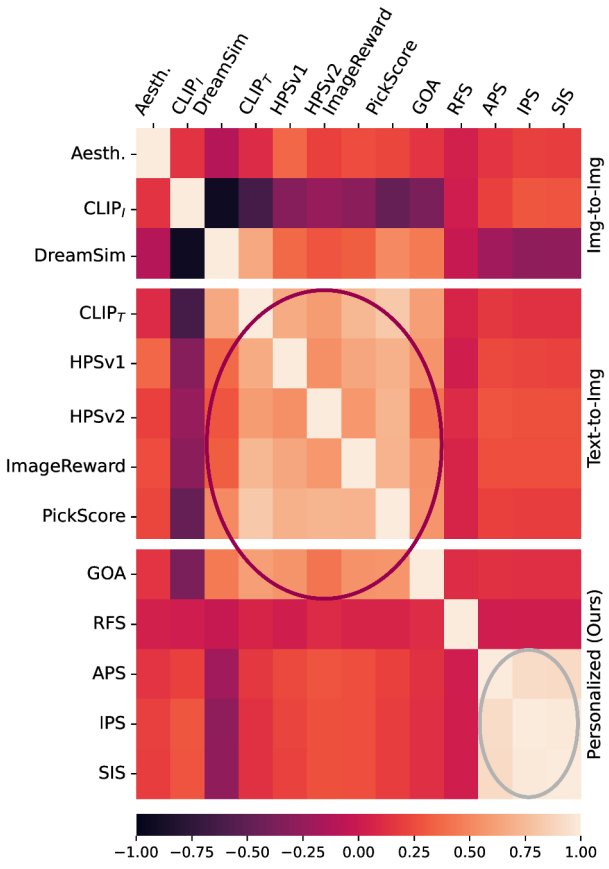

Figure 5. **Correlation heatmap among key introduced and ablated metrics.** Personalized metrics (bottom) ( ) have a small correlation with Img-to-Img (top) and Text-to-Img (middle) metrics with a Pearson correlation of $\rho = 0.06$ and $\rho = 0.25$, which suggests that they provide an additional dimension by which T2I systems can be evaluated. This is in contrast to, Text-to-Img metrics ( ) that have an increased correlation (e.g., CLIP$_T$ and HPSv2, with $\rho = 0.6$). Note also how the Text-to-Img metrics have a high correlation with `GOA` but very little correlation with `RFS`, indicating that they mostly assess the representation of the objects in the image and not the interactions. Overall, our metrics appear to capture several orthogonal dimensions (and crucial for evaluating the output's quality, per Main paper Fig. 1–right); that can not be fully assessed with existing metrics.

formatted as a 2D array. MASK$_{bin}$ is the downsampled binary mask of the input image, where the background is assigned a zero value and the subject is indicated with a value of 1. N denotes the total number of elements in the 2D array. This formulation ensures that the loss calculation is effectively concentrated on the areas of interest—namely, the subject of the image—thereby facilitating a more targeted and accurate learning process.

Additionally, to constrain the information in the $\mathbf{S}^*$ embeddings, so that they manage to focus on the facial charac-

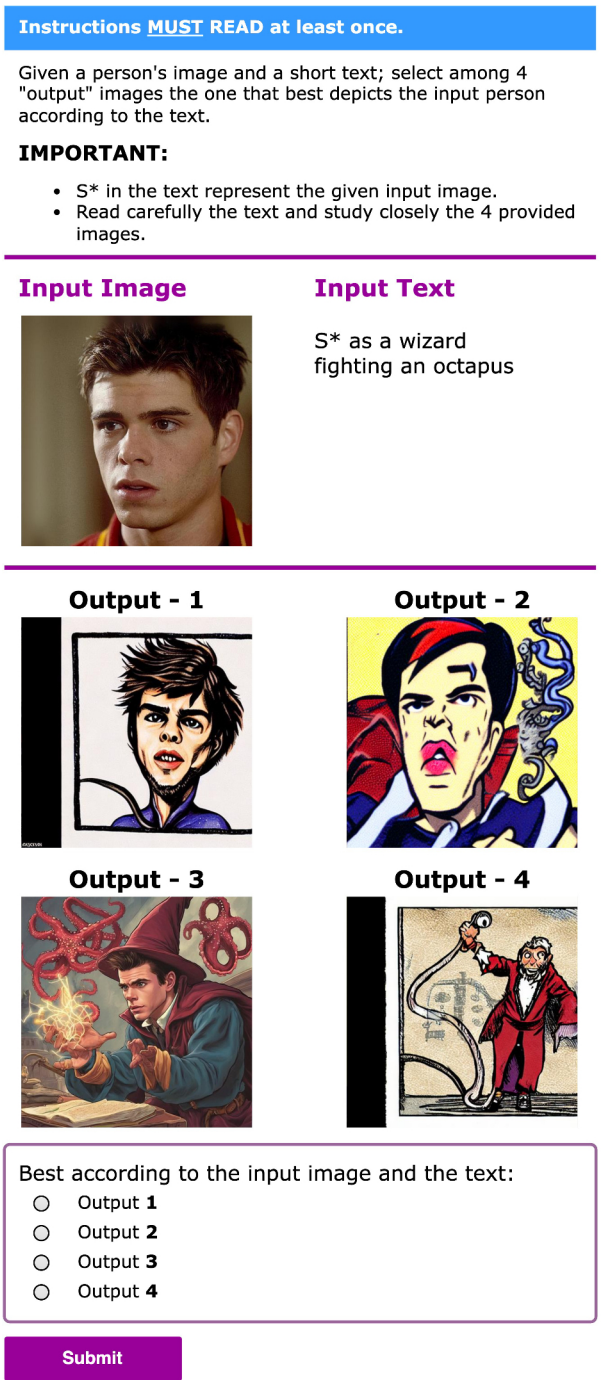

Figure 6. **Instructions given to annotators for evaluating *holistically* the quality of personalized T2I systems.** These are the instructions given to AMT users in order to assess the alignment and predictability of our proposed metrics with human judgment, in terms of finding the *overall* best generation. I.e., in the posed question no fine-grained sub-questions are included (unlike those asked in the experiment depicted in Fig. 7).

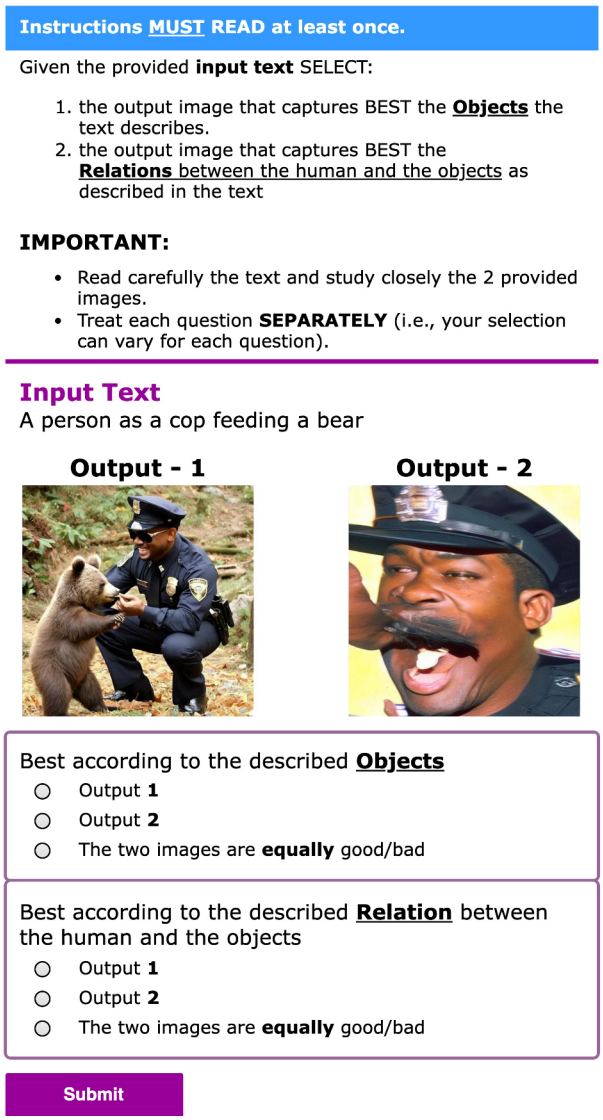

Figure 7. **Instructions given to annotators for evaluating images based on object-centric context.** These are the initial instructions given to users in Amazon Mechanical Turk (AMT) in order to accurately assess the alignment of our proposed metrics with human judgment on *object-centric context-based* evaluation.

teristics of the input identities, instead of the high-frequency details of the image, we add a regularization term for these embeddings on the total training loss. As a result, the final loss is calculated as follows:

$$Loss = MSE_{masked} + \alpha \cdot REG_{S^*} \quad (2)$$

Where α is a hyperparameter that controls the weight of the regularization term REG_{S^*} in the overall loss function. Visually, this can be seen on the right part of Fig. 8.

Table 6. **Textual prompts used for training StellarNet.** During training for each image we randomly sample one of the prompts below to train the system. These prompts are constructed to align with the expected images from CelebAHQ [8].

a portrait photo of a S^* person
a photo of a S^* person
a photo of a S^* person face
a cropped photo of a S^* person
a cropped photo of a S^* person face
a high quality photo of a S^* person face
a good photo of a S^* person face
a photo of a S^* person face

StellarNet. For StellarNet, we build on top of SDXL (base only without refiner) and we use two CLIP image encoders matching the CLIP text encoders of SDXL. Each of the MLPs mapping image to textual features has 2 layers (an input and an output layer) with 1024 hidden dimensions and input/output dimensions matching those of the CLIP encoders. For fine-tuning the UNet, following the original paper's authors [5], we only train offsets for the attention layers of the UNet using LoRA with rank$=4$.

Training Details. In all our experiments and results, we set the auxiliary parameter λ to 1. StellarNet is trained on the train set of CelebAHQ consisting of $\sim$*24k* images with masks, for $\sim$*23k* steps with an effective batch size of 64 on 8 A100 GPUs. The regularization weight α is set to 0.01, the LoRA learning rate is set to 1×10^{-7} and the `DTI` learning rate is set to 2×10^{-6}. During training, we randomly paired one image with one of the prompts in Tab. 6, which are constructed to align with the expected images from CelebAHQ [8].

4. Failure Cases

StellarNet, like most current personalized T2I systems, exhibits three primary failure cases that are noteworthy. Firstly, the model tends to homogenize facial features when generating multiple people in a single scene leading to images like those in Fig. 10 - middle left.

Secondly, the model demonstrates a challenge in decoupling variable characteristics from the subject in the input image. For example, if the input image features a person wearing glasses or a hat, the model often replicates these accessories in the output, regardless of their relevance per the prompt (Fig. 10 - middle).

Thirdly, while less pronounced, there is an indication of social semantic bias originating from the underlying SDXL model. In cases involving certain semantic contexts, such as professions stereotypically associated with a specific gender (e.g., nurse, cop), the model tends to generate identities

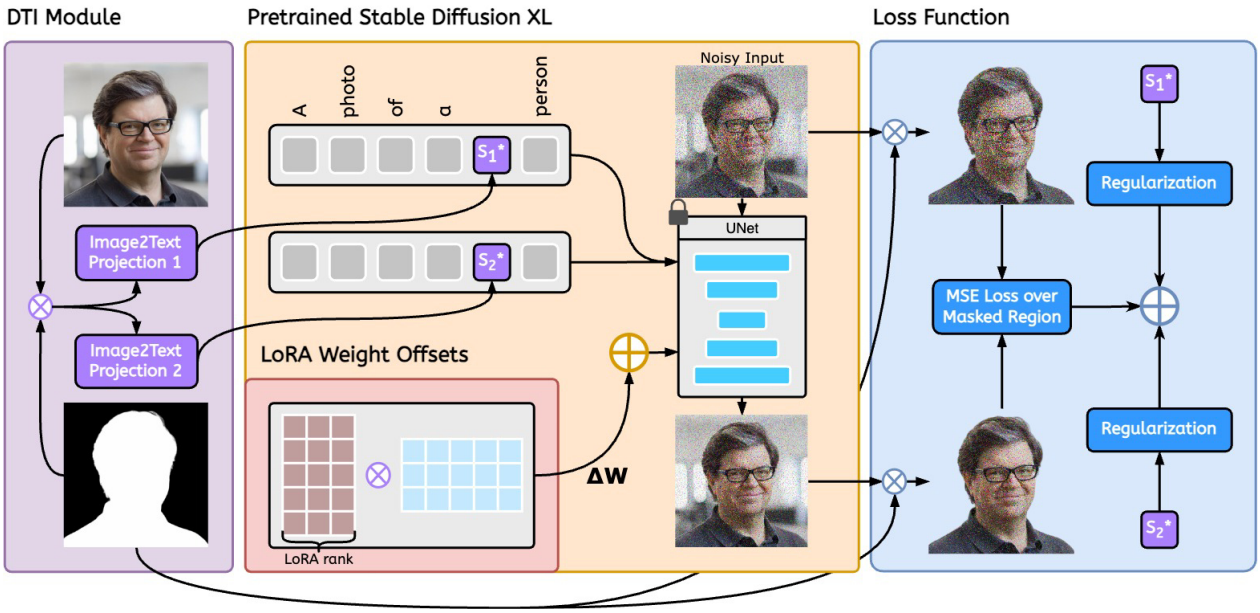

Figure 8. **StellarNet architecture overview**. The Dynamic Textual Inversion (DTI) module uses a binary subject mask to invert the input identity image into textual embeddings, S^*, that are used to condition the Stable Diffusion backbone (**left**). This S^* is used as part of the textual prompt that is passed to the pre-trained text-to-image model (SDXL [11] in our case) to generate images of the given identity (**middle**). Additionally, we use LoRA [5] weight offsets for the UNet [13] backbone of SDXL to learn to better understand the new $S*$ embedding and generate the desired personality (**middle left**). To avoid unnecessary penalization of the model, we use the MSE loss only over the masked outputs, and we additionally regularize the S^* embeddings to avoid learning high-frequency information (**right**). During training, we only update the weights of the Image2Text projection and the LoRA weights.

conforming to these social biases. Observing Fig. 10 (right column), there is an instance where the system completely changed the gender of the input image.

It is important to note, however, that these tendencies do not manifest uniformly across generations, as depicted in the bottom row of Fig. 10, where the model does not exhibit these problems.

Lastly, we also note various edge cases for our system. Notably, there are instances where the system fails to generate a human figure in the output, particularly in prompts involving certain objects like cars, as seen on the left of Fig. 9. Another issue arises when the model confuses or mixes up semantic objects from the prompt, leading to weird depictions in the generated image (Fig. 9 - middle). Furthermore, the system occasionally exhibits confusion in interpreting poorly structured prompts, resulting in outputs that do not align well with the intended request (Fig. 9 - right). These edge cases, though not frequent, highlight areas where StellarNet's understanding and representation capabilities can be further refined.

5. Miscellaneous

An additional figure, with more examples comparing StellarNet with competing methods, can be found in Fig. 11. Also, a visualization of the nouns in Stellar-$\mathcal{H}$ can be seen in Fig. 2.

Note: All input images in the paper and this supplementary material (denoted as *Original Image* in the figures) are from CelebAMask-HQ [8].

(a) S* onto a Rimac at a foggy day (b) S* as a sumo fighter fighting a tiger (c) S* sitting on sunbed made of bronze

Figure 9. **Extreme failure cases of StellarNet.** This figure presents a few rare edge cases of our system. Fig. 9a hiding the subject's identity. Fig. 9b merging of semantics from the prompt. Fig. 9c inability to accurately follow confusing prompts.

Figure 10. **Common failure cases of StellarNet documented also in similar T2I systems.** This figure highlights three well-known failure cases in current personalization and T2I systems. Firstly, there is a tendency for such a system to blend characteristics when multiple people of the same gender are in an image (**left**) [6]. Secondly, T2I systems struggle to decouple variable characteristics such as glasses and hats from the input, often carrying these over to the output (**middle**) [17]. Finally, there appears to be a social bias in the backbone model (Stable Diffusion), with prompts involving certain professional roles often leading to changes in the gender representation of the input subject (**right**) [12]. The **top** row shows the original input images, while the lower rows demonstrate variations with the same prompt but different seeds illustrating the aforementioned failure modes and their dependency on not-controllable (random) seeding.

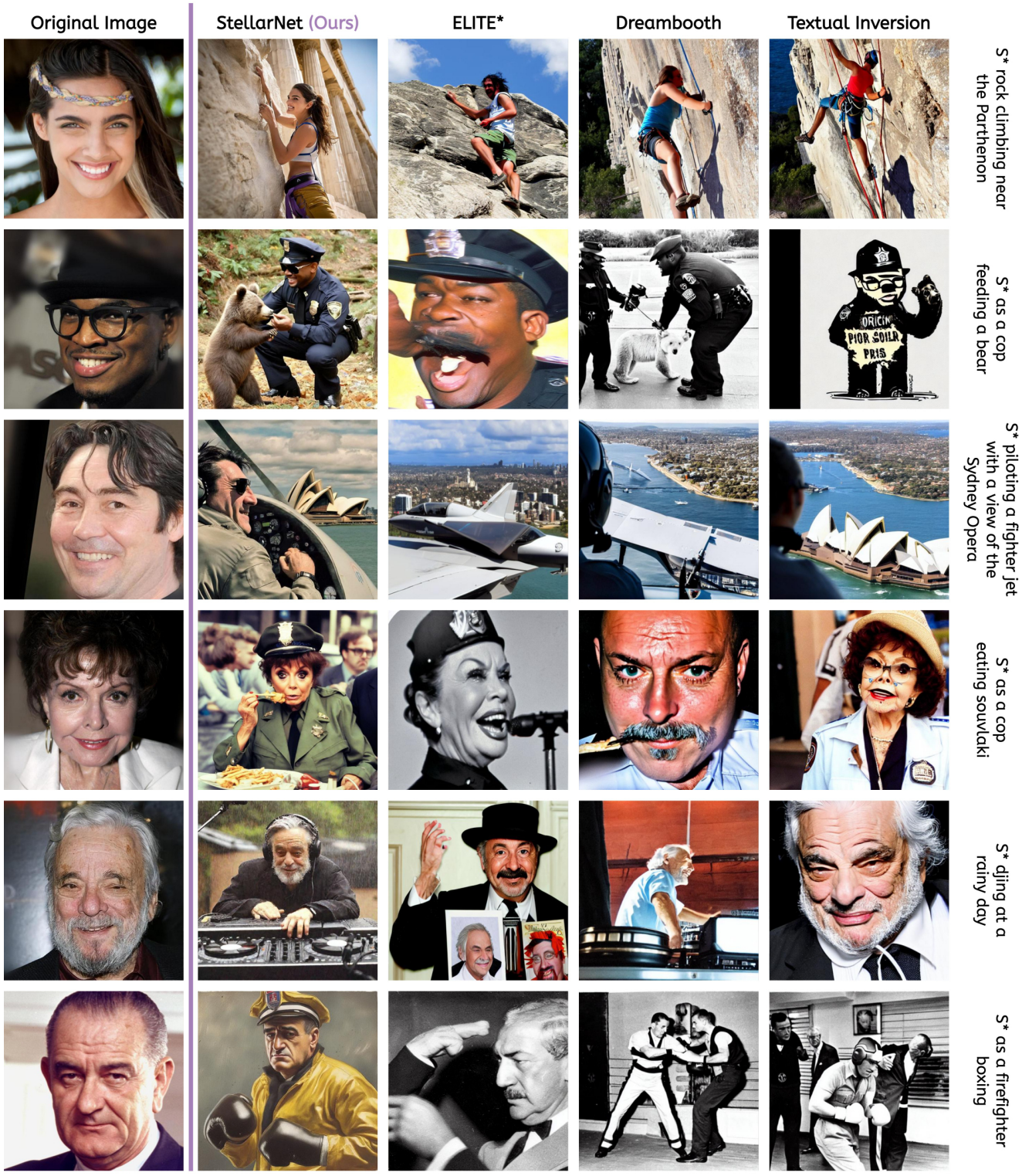

Figure 11. **More qualitative comparisons.** Complementing the rest of the paper, here we present more qualitative examples comparing StellarNet with ELITE*, DreamBooth, and Textual Inversion. With S^* in the prompt, we denote the input image identity.

References

[1] Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL*, 2019. 4

[2] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and Lawrence C. Zitnick. Microsoft COCO Captions: Data collection and evaluation server. *Computing Research Repository (CoRR)*, abs/1504.00325, 2015. 4

[3] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 5

[4] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *Computing Research Repository (CoRR)*, abs/2103.00020, 2022. 7

[5] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. 7, 9, 10

[6] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 12

[7] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *International Journal of Computer Vision (IJCV)*, 2020. 4

[8] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 9, 10

[9] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision (ECCV)*, 2016. 4

[10] Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*. Association for Computational Linguistics, 2022. 2, 4

[11] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *Computing Research Repository (CoRR)*, abs/2307.01952, 2023. 7, 10

[12] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *Computing Research Repository (CoRR)*, abs/2112.10752, 2022. 12

[13] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *Computing Research Repository (CoRR)*, abs/1505.04597, 2015. 10

[14] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 7

[15] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *International Conference on Computer Vision (ICCV)*, 2019. 4

[16] Zijie J. Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and Duen Horng Chau. DiffusionDB: A large-scale prompt gallery dataset for text-to-image generative models. *Computing Research Repository (CoRR)*, abs/2210.14896, 2022. 1

[17] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. ELITE: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *International Conference on Computer Vision (ICCV)*, 2023. 7, 12

[18] Danfei Xu, Yuke Zhu, Christopher B. Choy, and Li Fei-Fei. Scene graph generation by iterative message passing. In *CVPR*, 2017. 4