A. Technical Appendices and Supplementary Material

A.1. Ablation study

We conduct an ablation study comparing FL2T (Ours) and CIDM [5] under varying settings. First, constraining the number of reference images (Table 3) reveals that while both models improve with more references, FL2T consistently outperforms CIDM and matches its original performance (≥ 4 images) using only three, demonstrating greater efficiency. Second, as shown in Table 2 proxy embeddings attain optimal semantic representation with two transformer decoder layers, as too few layers under-express feature relations and too many risk rank collapse [18]. Lastly, FL2T maintains superior scores across all LoRA ranks (Table 4), outperforming CIDM (rank = 4) even at a lower rank with 25% fewer parameters, highlighting the parameter efficiency of our approach.

Number of reference images	Methods	V1 - V5		V6 - V10		Avg.	
		IA (↑)	TA (↑)	IA (↑)	TA (↑)	IA (↑)	TA (↑)
1	CIDM	76.6	79.5	69.7	74.5	73.2	77.0
	Ours	**80.0**	78.7	**72.5**	74.1	**76.3**	76.4
2	CIDM	78.1	79.2	71.2	74.2	74.6	76.7
	Ours	**80.4**	78.7	**72.1**	74.0	**76.3**	76.3
3	CIDM	80.8	78.2	72.1	73.6	76.4	75.9
	Ours	**82.3**	79.0	**73.7**	73.9	**78.0**	76.4
≥4	CIDM	84.0	77.3	71.7	72.6	78.0	74.8
	Ours	**84.9**	77.0	**74.8**	88.9	**79.8**	75.4
Number of reference	**Methods**	**IMS (↑)**	**FID (↓)**	**IMS (↑)**	**FID (↓)**	**IMS (↑)**	**FID (↓)**
1	CIDM	61.7	199.0	**55.3**	291.8	58.5	245.4
	Ours	**65.0**	179.2	54.7	292.4	**59.9**	235.8
2	CIDM	66.5	178.3	57.8	284.9	62.2	231.6
	Ours	**69.0**	165.5	**58.1**	278.8	**63.6**	222.1
3	CIDM	71.3	156.5	58.5	285.3	64.9	220.9
	Ours	**72.0**	152.9	**60.3**	272.7	**66.2**	212.8
≥4	CIDM	74.0	142.0	61.5	270.0	67.7	206.0
	Ours	**78.2**	133.3	**64.4**	262.7	**71.3**	198.0

Table 3. Comparison of model performances with different number of reference images. IA and TA refer to CLIP [32] Image Alignment and Text Alignment scores respectively. The best IA and IMS scores have been denoted in **bold**, and the TA and FID values have been underlined.

A.2. Datasets, Implementation Details and Evaluation Metrics

LoRA rank	Methods	V1 - V5		V6 - V10		Avg.	
		IA (↑)	TA (↑)	IA (↑)	TA (↑)	IA (↑)	TA (↑)
3	CIDM	82.3	78.3	71.9	72.4	77.1	75.3
	Ours	**84.5**	77.6	**73.8**	88.6	**79.1**	75.6
4	CIDM	84.0	77.3	71.7	72.6	78.0	74.8
	Ours	**84.9**	77.0	**74.8**	88.9	**79.8**	75.4
5	CIDM	82.5	78.1	73.1	72.9	77.8	75.4
	Ours	**85.4**	77.3	**74.7**	73.7	**80.1**	75.5
6	CIDM	82.7	78.3	73.0	73.0	77.9	75.6
	Ours	**84.4**	77.7	**73.1**	74.2	**78.7**	76.0
LoRA rank	**Methods**	**IMS (↑)**	**FID (↓)**	**IMS (↑)**	**FID (↓)**	**IMS (↑)**	**FID (↓)**
3	CIDM	72.6	150.3	**62.2**	262.7	67.4	206.5
	Ours	**76.6**	135.1	61.8	263.9	**69.2**	199.5
4	CIDM	74.0	142.0	61.5	270.0	67.7	206.0
	Ours	**78.2**	133.3	**64.4**	262.7	**71.3**	198.0
5	CIDM	73.5	145.4	62.4	260.2	68.0	202.8
	Ours	**76.8**	132.5	**63.3**	263.7	**70.1**	198.1
6	CIDM	72.6	147.1	**61.7**	267.2	67.1	207.2
	Ours	**76.7**	134.8	60.9	273.3	**68.9**	204.0

Table 4. Comparison of model performances for different LoRA [15,45] configurations. IA and TA refer to CLIP [32] Image Alignment and Text Alignment scores, respectively. The better IA and IMS scores have been denoted in **bold**, and TA and FID values have been underlined.

Datasets. We evaluate our approach on three datasets, each selected to ensure strong semantic alignment with our task objectives (Fig. 6). For the CIFC dataset [5], we adopt the benchmark established in the original paper, which features visually rich concept cards designed to challenge models in hallucination detection. For the ImageNet [4] subset, we manually select 3–5 images per class from ten classes, ensuring that each image contains a well-centered, unobstructed primary object. For the CelebA dataset [26], we construct a subset of ten identities, with 3–5 representative images per identity. Images are chosen based on clear frontal facial orientation, uniform lighting, and minimal occlusion to preserve identity consistency. This hand-curated dataset design provides high-quality supervision, which is crucial for minimizing semantic hallucinations and promoting accurate visual-textual alignment.

For the experiments on varying numbers of reference images (Table 3), we have used a fixed set of images picked randomly (without human intervention).

Implementation Details. We use Stable Diffusion (SD-1.5) [35] as the pretrained model for all experiments. The training is conducted with a fixed initial learning rate of 1.0×10^{-3} for updating textual embeddings and 1.0×10^{-4} for optimizing the U-Net. Empirically, we set $\gamma_1 = 0.1$ and $\gamma_2 = 0.1$ in Eq. 6.

Evaluation Metrics. Following the experiments under our problem setting as in Sec 3, we evaluate our generated images across 2 metrics - Image Alignment (IA) and Text Alignment (TA). Image Alignment (IA) scores are computed using CLIP [32] image encoder, comparing the similarity of features between generated images and reference images. Similarly, we utilize the text encoder of CLIP [32] to evaluate the text-image similarity between the input prompt and synthesized image for the Text Alignment (TA) scores. Additionally, we utilize *Identity Matching Score (IMS)* that measures the semantic closeness of the generated image and reference image [25]. It is computed as the cosine similarity score between embeddings of generated images and mean of reference image embeddings. For the CIFC dataset [5] and the ImageNet dataset [4], we utilize ResNet-152 [13] as the image encoder. On the other hand,

	Methods	V1		V2		V3		V4		V5		V6		V7		V8		V9		V10		Avg.	
		IMS (↑)	FID (↓)	IMS (↑)	FID (↓)	IMS (↑)	FID (↓)	IMS (↑)	FID (↓)	IMS (↑)	FID (↓)	IMS (↑)	FID (↓)	IMS (↑)	FID (↓)	IMS (↑)	FID (↓)	IMS (↑)	FID (↓)	IMS (↑)	FID (↓)	IMS (↑)	FID (↓)
CIFC [5]	**CIDM** [5]	84.7	82.2	74.5	166.1	66.7	156.7	65.1	211.9	78.9	93.2	53.6	348.4	80.8	119.0	49.7	345.6	70.0	195.6	53.3	341.5	67.7	206.0
	Ours with proxy-guidance	87.7	71.8	75.1	169.3	75.6	145.4	71.3	202.5	81.1	77.7	55.8	355.2	89.5	98.4	48.3	353.4	69.8	189.9	58.6	316.4	71.3	198.0
	Δ	+3.0	-10.4	+0.6	+3.2	+8.9	-11.3	+6.2	-8.6	+2.2	-15.5	+2.2	+6.8	+8.7	-20.6	-1.4	+7.8	-0.2	-5.7	+5.3	-25.1	+3.6	-8.0
CelebA [26]	**CIDM** [5]	63.9	233.2	65.6	219.7	60.1	250.0	76.0	166.8	55.3	246.7	61.4	234.5	58.0	193.4	56.1	273.5	51.2	185.7	57.0	197.0	60.5	220.0
	Ours with proxy-guidance	64.4	225.8	69.3	212.2	63.5	239.8	70.2	168.9	57.4	188.7	63.9	221.1	58.1	218.0	59.5	236.9	55.4	170.4	63.0	195.4	62.5	209.0
	Δ	+0.5	-7.4	+3.7	-7.5	+3.4	-10.2	-5.8	+2.1	+2.1	-58.0	+2.5	-13.4	+0.1	+24.6	+3.4	-36.6	+4.2	-15.3	+6.0	-1.6	+2.0	-11.0
INet [4]	**CIDM** [5]	81.6	129.6	68.3	113.6	85.4	75.7	85.0	85.1	73.5	91.5	78.4	86.2	64.9	135.1	79.4	58.8	80.1	88.1	87.4	57.3	78.4	92.1
	Ours with proxy-guidance	85.2	103.1	75.2	107.1	88.5	57.1	84.1	79.3	85.7	83.5	84.8	69.1	75.6	110.3	77.2	64.3	80.2	79.8	95.4	31.6	83.2	78.5
	Δ	+3.6	-26.5	+7.1	-6.5	+3.1	-18.6	-0.9	-6.2	+12.2	-8.0	+6.4	-17.1	+10.7	-24.8	-2.2	+5.5	+0.1	-8.3	+8.0	-25.7	+4.8	-13.6

Table 5. **Additional Experiments.** FL2T outperforms CIDM across the three datasets on Identity Matching Scores (IMS) [25] and Fretchet Inception Distance (FID) [17], effectively showcasing its ability to preserve identities better and improved generation capabilities.

we have utilized VGG-Face [38] for CelebA [26]. Further, *Fretchet Inception Distance (FID)* evaluates the quality and diversity of images by comparing the Inception-v3 feature distributions of the reference images and generated images [17].

A.3. Additional Experiments

Beyond CLIP scores, we utilize two other metrics, *Identity Matching Score (IMS)* and *Fretchet Inception Distance (FID)*. IMS that measures the semantic closeness of the generated image and reference image [25]. Specifically, this metric is designed to measure the identity match between of the concepts in generated images. Whereas FID compares the distributions of generated and reference images using Inception-v3 features [17]. Based on these metrics, as shown in Table 5, we observe that FL2T outperforms CIDM across the three datasets, showing trends akin to CLIP [32] scores.

On comparison with Multi-concept Customization [20], FL2T delivers consistent identity gains , boosting IMS by 12–31 points across concepts, with strong IA improvements on most concepts. Importantly, it achieves lower (better) FID indicating high quality image generation results as shown in Table 6.

A.4. Comparing Attention and Concatenation Operations

The main contribution of this work revolves around "positively" exploiting the higher-order interactions between concepts. Towards this end, we analyze two operations that are commonly utilized to extract richer representations from a set of embeddings. Consider a set of embeddings, $S = \{X, Y, Z\} \in \mathbb{R}^d$. We assume that trivial components such as linear layers for scaling in attention and downsampling in concatenation are present.

Attention. Attention computes the cosine similarity between X and Y and adds the projected component of X on Y to X -

$$\text{Attn}(X; (X, Y)) = X + (X \cdot Y^T)Y$$

To develop a deeper understanding, we consider three vectors and let a_{XY} be the cosine similarity between vectors X and Y. Then, after the first attention operation where $S_1 = \{X_1, Y_1, Z_1\}$ is the output :

$$\begin{aligned} X_1 &= X + a_{XY}Y + a_{XZ}Z \\ Y_1 &= a_{XY}X + Y + a_{YZ}Z \\ Z_1 &= a_{XZ}X + a_{YZ}Y + Z \end{aligned}$$

Subsequently, extending this to a second attention layer provides us:

$$\begin{aligned} X_2 = {} & [1 + (a^2_{XY} + a^2_{XZ})(a^2_{XY} + a^2_{YZ} + a^2_{XZ} + 3) + 6a_{XY}a_{YZ}a_{XZ}]X \\ & + [a_{XY}(3a^2_{YZ} - 2) + (a_{XY} + a_{XZ}a_{YZ})(a^2_{XY} + a^2_{YZ} + a^2_{XZ} + 6)]Y \\ & + [a_{XZ}(3a^2_{YZ} - 2) + (a_{XZ} + a_{XY}a_{YZ})(a^2_{XY} + a^2_{YZ} + a^2_{XZ} + 6)]Z \end{aligned}$$

The number of pairwise interactions between the vectors of S has significantly increased in the second attention layer, thereby allowing the model to capture higher-level dependencies between the vectors. It is important to note that attention is a generalized form of weighted aggregation.

Concatenation. On the other hand, concatenation captures non-linear interactions between each element of the vectors. This can introduce unintended noise and alter the original embedding negatively. The operation is defined below where a_i, b_i and c_i are polynomials for the i-th element.

$$\text{Concat(X, Y, Z)} = \sum_{i=0}^{d-1} a_i X_i + b_i Y_i + c_i Z_i$$

A.5. Bounding Model Drift under Unnormalized Attention Coefficients

We quantify one-step model drift by the norm of the aggregated gradient. Lemma A.1 shows a universal upper bound (the same crude bound as uniform summation), and Theorem A.1 gives a simple, explicit construction proving that unnormalized attention can *strictly reduce* drift relative to uniform summation.

Methods	**V1**		**V2**		**V3**		**V4**		**V5**		**V6**		**V7**		**V8**		**V9**		**V10**		**Avg.**	
	IA (↑)	**TA** (↑)	**IA** (↑)	**TA** (↑)	**IA** (↑)	**TA** (↑)	**IA** (↑)	**TA** (↑)	**IA** (↑)	**TA** (↑)	**IA** (↑)	**TA** (↑)	**IA** (↑)	**TA** (↑)	**IA** (↑)	**TA** (↑)	**IA** (↑)	**TA** (↑)	**IA** (↑)	**TA** (↑)	**IA** (↑)	**TA** (↑)
Multi-concept Customization [20]	74.4	77.6	76.6	73.4	73.4	74.7	69.7	73.9	78.2	80.7	65.9	76.0	73.6	66.3	68.7	76.2	62.7	75.8	71.2	83.5	71.4	75.8
Ours with proxy guidance	**84.4**	74.0	**86.1**	81.6	**84.4**	70.1	**82.2**	81.1	**87.2**	78.0	**69.3**	72.9	**85.3**	70.4	**60.7**	77.6	**82.1**	76.6	**76.5**	71.6	**79.8**	75.4
Δ	+10.0	-3.6	+9.5	+8.2	+11.0	-4.6	+12.5	+7.2	+9.0	-2.7	+3.4	-3.1	+11.7	+4.1	-8.0	+1.4	+19.4	+0.8	+5.3	-11.9	+8.4	-0.4
Methods	**IMS** (↑)	**FID** (↓)	**IMS** (↑)	**FID** (↓)	**IMS** (↑)	**FID** (↓)	**IMS** (↑)	**FID** (↓)	**IMS** (↑)	**FID** (↓)	**IMS** (↑)	**FID** (↓)	**IMS** (↑)	**FID** (↓)	**IMS** (↑)	**FID** (↓)	**IMS** (↑)	**FID** (↓)	**IMS** (↑)	**FID** (↓)	**IMS** (↑)	**FID** (↓)
Multi-concept Customization [20]	65.4	160.9	62.8	98.7	64.1	189.5	63.7	222.6	66.2	86.5	67.8	386.6	61.9	91.2	60.5	386.9	62.3	220.0	64.8	379.3	63.5	216.2
Ours with proxy guidance	**85.2**	71.8	**75.2**	169.3	**88.5**	145.4	**84.1**	202.5	**85.7**	77.7	**84.8**	355.2	**75.6**	98.4	**77.2**	353.4	**80.2**	189.9	**95.4**	316.4	**83.2**	198.0
Δ	+19.8	-7.3	+12.4	-8.4	+24.4	+44.1	+20.4	+20.1	+19.5	+8.8	+17.0	+31.4	+13.7	-7.2	+16.7	+33.5	+17.9	+20.1	+30.6	+62.9	+19.7	+18.2

Table 6. **Comparison with multi-concept method.** We compare FL2T with Multi-concept customization [20] on the CIFC dataset [5] across all metrics, CLIP Image Alignment (IA), CLIP Text Alignment (TA), Identity Matching Scores (IMS) [25] and Fretchet Inception Distance (FID) [17], and notice that FL2T outperforms Multi-concept Customization across all metrics.

Operation	**Pairwise Interactions (per layer)**	**After M Layers**	**Permutation Invariance**
Summation $(x + y)$	nd	Mnd	✓
Concatenation $([x; y])$	0	0	✓
Attention	n^2d	Mn^2d	✓

Table 7. **Pairwise interactions.**We compute the number of pairwise interactions for combining two vectors $x, y \in \mathbb{R}^d$ across n embeddings. We assume the same k, d for M layers. The attention function showcases its superiority by showing the largest number pairwise interactions and is permutation invariant.

Setup and Assumptions Let $(\mathbb{R}^d, \langle\cdot,\cdot\rangle)$ be a real inner-product space with norm $\|v\| = \sqrt{\langle v, v\rangle}$. We have N concepts with losses $\ell_i(\theta)$ and gradients

$$m_i \ := \ \nabla_\theta \ell_i(\theta) \in \mathbb{R}^d, \qquad i = 1, \ldots, N.$$

We compare two aggregate gradients:

$$M_{\mathrm{CIDM}} \ := \ \sum_{i=1}^{N} m_i$$

$$M_{\mathrm{FL2T}} \ := \ \sum_{i=1}^{N} \lambda_i \, m_i, \ \ \lambda_i \in [-1, 1].$$

We impose *no* normalization on (λ_i) (in particular, $\sum_i \lambda_i$ need not equal 1). One-step drift with learning rate $\eta > 0$ is proportional to $\|M\|$, so we compare $\|M_{\mathrm{FL2T}}\|$ to $\|M_{\mathrm{CIDM}}\|$.

Lemma A.1 (Universal Upper Bound)**.** *For any coefficients* $\lambda_i \in [-1, 1]$*,*

$$\|M_{\mathrm{FL2T}}\| \ = \ \Big\| \sum_{i=1}^{N} \lambda_i m_i \Big\| \ \le \ \sum_{i=1}^{N} |\lambda_i| \, \|m_i\| \ \le \ \sum_{i=1}^{N} \|m_i\|.$$

Proof. By the triangle inequality and homogeneity of norms,

$$\Big\| \sum_{i=1}^{N} \lambda_i m_i \Big\| \ \le \ \sum_{i=1}^{N} \|\lambda_i m_i\| \ = \ \sum_{i=1}^{N} |\lambda_i| \, \|m_i\| \ \le \ \sum_{i=1}^{N} \|m_i\|$$

because $|\lambda_i| \le 1$ for all i. □

Theorem A.1 (Existence of Reduced Drift)**.** *If* $M_{\mathrm{CIDM}} \neq 0$*, then there exists* $\boldsymbol{\lambda} \in [-1, 1]^N$*, not all equal to* 1*, such that*$\|M_{\mathrm{FL2T}}\| \ < \ \|M_{\mathrm{CIDM}}\|$.

Proof. Set $M_{\mathrm{CIDM}} = \sum_{k=1}^{N} m_k$. Then

$$\begin{aligned} \sum_{k=1}^{N} \langle M_{\mathrm{CIDM}}, \, m_k \rangle &= \Big\langle M_{\mathrm{CIDM}}, \, \sum_{k=1}^{N} m_k \Big\rangle \\ &= \langle M_{\mathrm{CIDM}}, \, M_{\mathrm{CIDM}} \rangle \\ &= \|M_{\mathrm{CIDM}}\|^2 \ > \ 0. \end{aligned}$$

Consequently, there exists an index $k^\star$ for which

$$\langle M_{\mathrm{CIDM}}, m_{k^\star} \rangle > 0.$$

Fix such an index $k^\star$. For any $\varepsilon \in (0, 1]$, define coefficients

$$\lambda_{k^\star} = 1 - \varepsilon, \qquad \lambda_j = 1 \quad \text{for all } j \neq k^\star.$$

These coefficients lie in $[-1, 1]$ and are not all equal to 1, so the resulting vector

$$M_{\mathrm{FL2T}} = \sum_{i=1}^{N} \lambda_i m_i = M_{\mathrm{CIDM}} - \varepsilon \, m_{k^\star}$$

is feasible. Expanding the squared norm yields

$$\begin{aligned} \|M_{\mathrm{FL2T}}\|^2 &= \|M_{\mathrm{CIDM}} - \varepsilon m_{k^\star}\|^2 \\ &= \|M_{\mathrm{CIDM}}\|^2 - 2\varepsilon \, \langle M_{\mathrm{CIDM}}, m_{k^\star} \rangle + \varepsilon^2 \|m_{k^\star}\|^2 \end{aligned}$$

Since $\langle M_{\mathrm{CIDM}}, m_{k^\star} \rangle > 0$, the quadratic expression on the right-hand side is strictly smaller than $\|M_{\mathrm{CIDM}}\|^2$ whenever

$$0 < \varepsilon < \min \left\{ 1, \ \frac{2 \, \langle M_{\mathrm{CIDM}}, \, m_{k^\star} \rangle}{\|m_{k^\star}\|^2} \right\}.$$

Thus $\|M_{\mathrm{FL2T}}\|^2 < \|M_{\mathrm{CIDM}}\|^2$, which implies $\|M_{\mathrm{FL2T}}\| < \|M_{\mathrm{CIDM}}\|$. Hence, there exists a feasible coefficient vector $\boldsymbol{\lambda}$ strictly reducing the norm, completing the proof. □

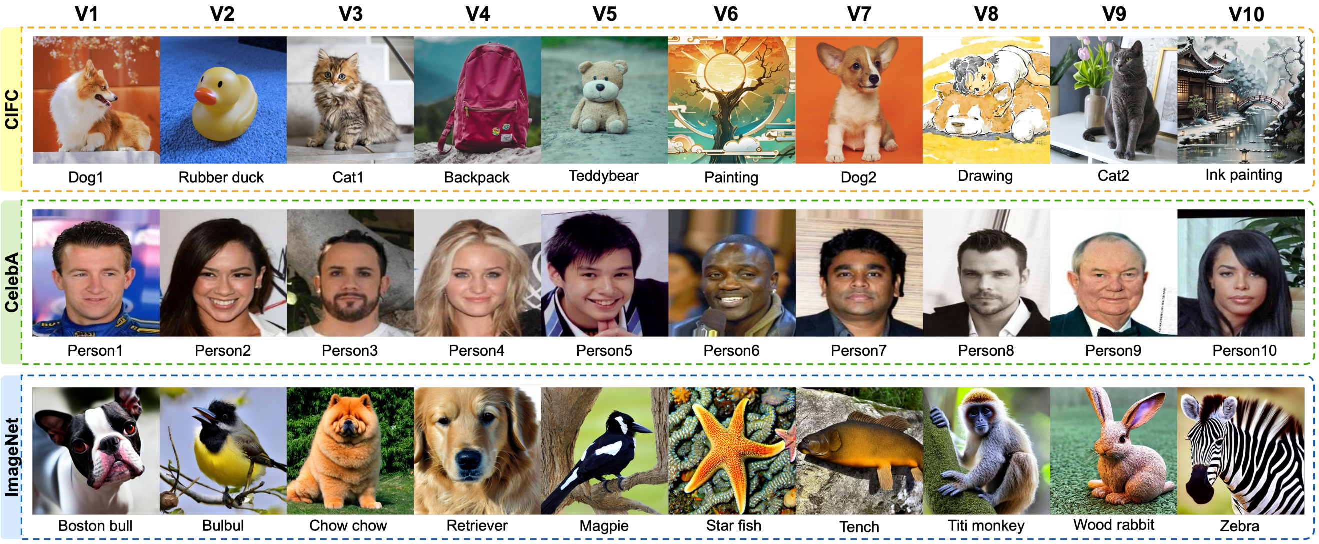

Figure 6. **Datasets used for the experiments.** We present an overview of the images used in our experiments for the CIDM [5], CelebA [26] and ImageNet [4] datasets.

Remark (degenerate case). If $M_{\text{CIDM}} = 0$, then $\|M_{\text{CIDM}}\| = 0$ is already minimal, so no strict decrease is possible; nonetheless, Lemma A.1 still holds.

A.6. Background on Custom Diffusion Models.

Latent diffusion models (LDMs) [8, 37] rely on conditional inputs, such as text prompts [31, 34] or images [16, 30], to guide the generation of images. These models utilize an encoder $E(\cdot)$ and a decoder $D(\cdot)$ to facilitate image synthesis in the latent space. Custom diffusion models (CDMs) [7, 11, 27] extend LDMs by incorporating low-rank adaptation (LoRA) [15, 45] to fine-tune pretrained diffusion models [1, 35] for personalized concept learning.

Given a personalized image-text pair (x, p), the encoder $E(\cdot)$ maps x to a latent representation z, with z_t denoting the noisy latent feature at timestep t ($t = 1, \ldots, T$). The text encoder $\Gamma(\cdot)$, such as a pretrained CLIP model [32], maps the text prompt p to a textual embedding $c = \Gamma(p)$. The objective for learning a personalized concept (x, p) at timestep t is defined as:

$$L_{CDM} = \mathcal{E}_{z\sim E(x),c,\varepsilon\sim\mathcal{N}(0,I),t}\left[\|\varepsilon - \varepsilon_{\theta'}(z_t|c,t)\|_2^2\right] \quad (7)$$

where $\varepsilon_{\theta'}(\cdot)$ represents the denoising UNet [31, 35] that gradually denoises z_t by estimating the Gaussian noise $\varepsilon \sim \mathcal{N}(0, I)$. The parameter set θ' corresponds to $\theta' = \theta_0 + \Delta\theta$, where $\theta_0 = \{W_l^0\}_{l=1}^L$ denotes the pretrained weights in LDMs, and $\Delta\theta = \{\Delta W_l\}_{l=1}^L$ corresponds to the LoRA-updated parameters [9, 20]. Here, $W_l^0, \Delta W_l \in \mathbb{R}^{a\times b}$ are the pretrained and low-rank weight matrices in the l-th transformer layer of θ', respectively, where a and b are matrix dimensions. Following [36, 53], the low-rank update ΔW_l can be factorized as $\Delta W_l = A_l B_l$, where $A_l \in \mathbb{R}^{a\times r}$ and $B_l \in \mathbb{R}^{r\times b}$ with rank $r \ll \min(a, b)$.

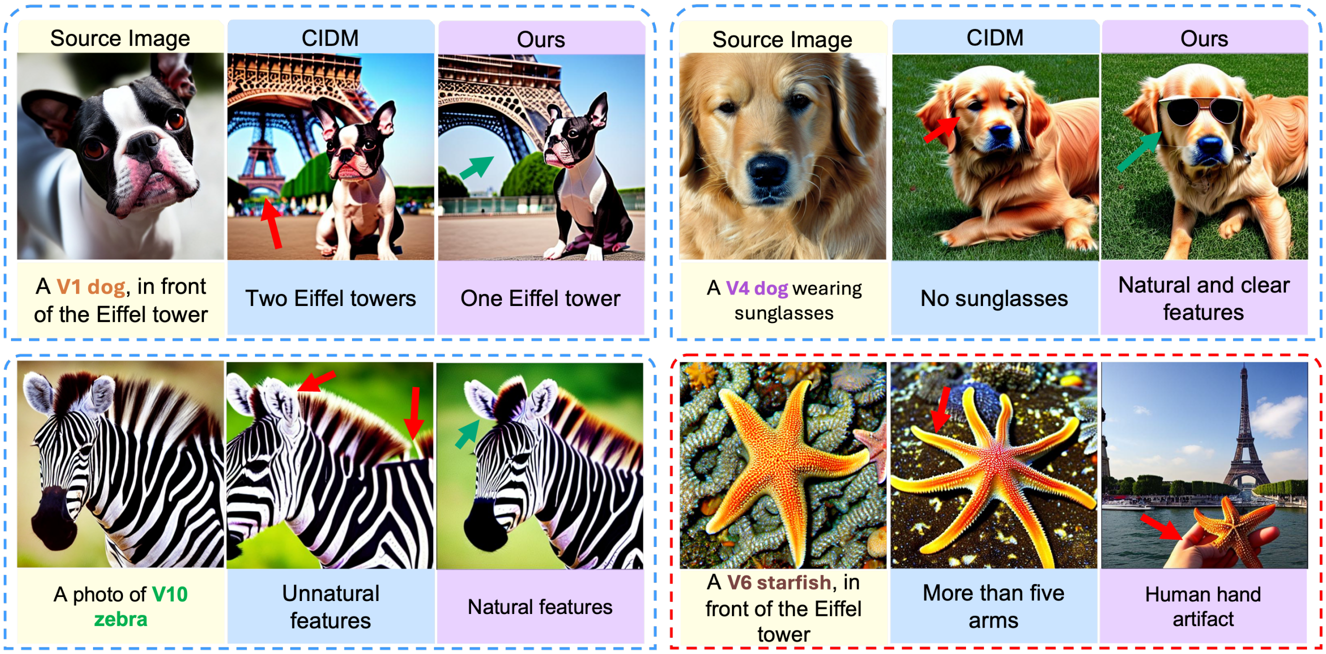

Figure 7. **Qualitative Analysis on the ImageNet dataset [4].** We compare the synthesized images by CIDM [5] and FL2T (Ours). The images are generated with a source image and an associated text prompt as input. Images with red and green arrows indicate regions of undesirable and desirable qualities, and their reasons are stated below each image. FL2T preseverves features and conforms to the text prompt better than CIDM. We have also shown a failure case of FL2T (bottom right in red box), where we observe a human hand artifact, but the identity of the starfish is preserved (only five arms).

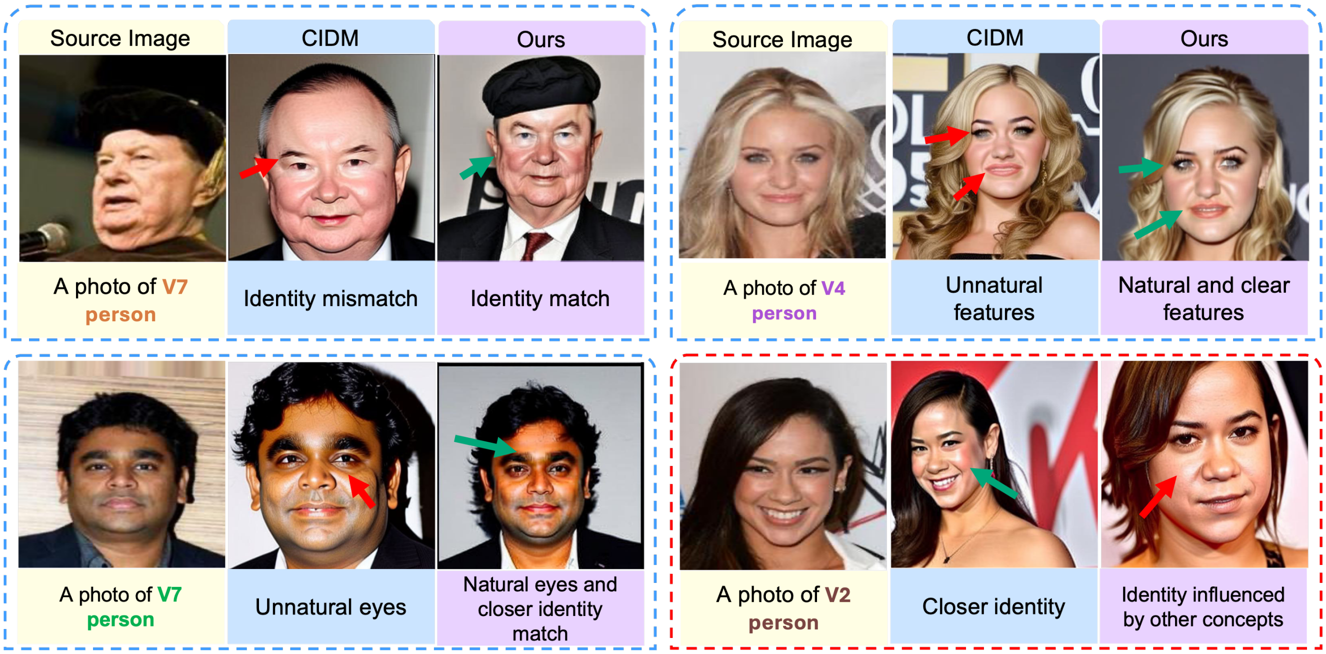

Figure 8. **Qualitative Analysis on the CelebA dataset [26].** We compare the synthesized images by CIDM [5] and FL2T (Ours). The images are generated with a source image and an associated text prompt as input. Images with red and green arrows indicate regions of undesirable and desirable qualities, and their reasons are stated below each image. FL2T preseverves identity and features better than CIDM. We have also shown a failure case of FL2T (bottom right in red box), where the identity of V2 person was influenced by other concepts in the dataset.

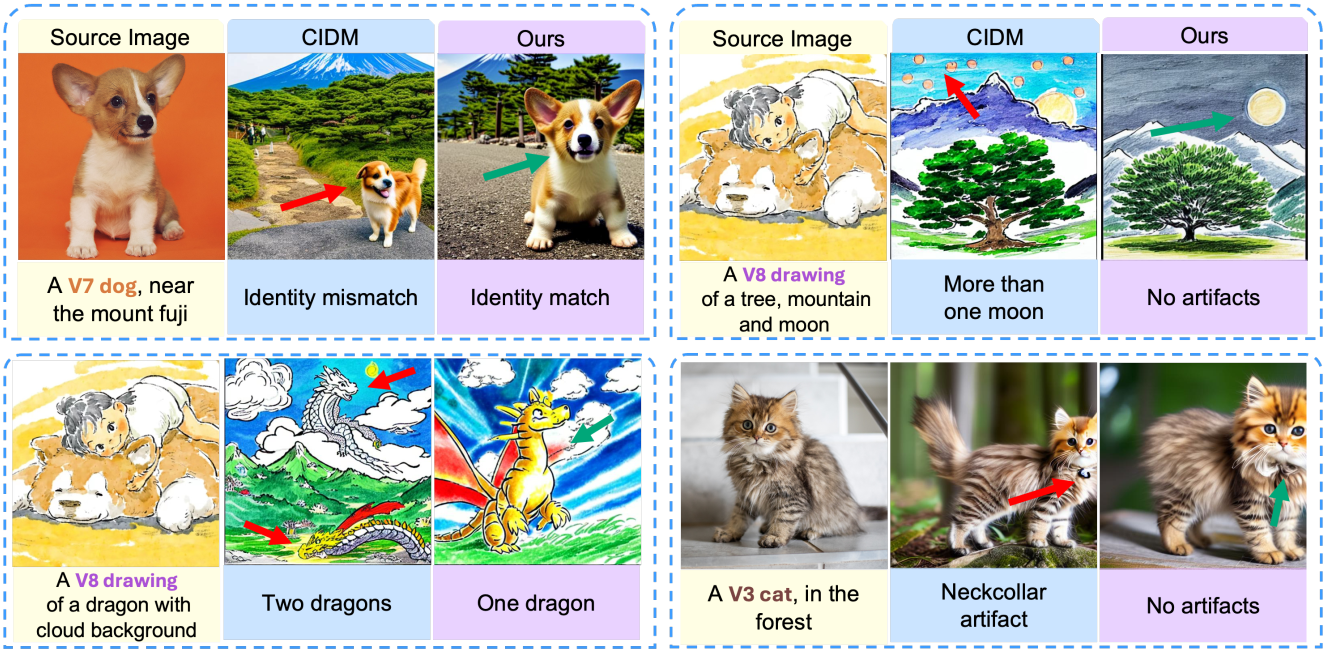

Figure 9. **More Qualitative Samples on the CIFC dataset [5].** We compare the synthesized images by CIDM [5] and FL2T (Ours). The images are generated with a source image and an associated text prompt as input. Images with red and green arrows indicate regions of undesirable and desirable qualities, and their reasons are stated below each image.