TALE: Training-free Cross-domain Image Composition via Adaptive Latent Manipulation and Energy-guided Optimization Supplementary Materials

Anonymous Authors

1 FULL ALGORITHM

We describe the complete pseudocode of the proposed training-free cross-domain image composition framework (TALE) in Algorithm 3. Note that $\mathcal{E}$ and $\mathcal{D}$ represents the autoencoder's encoder and decoder of the employed LDM [3], **PREPROCESS** and **INVERSE** denotes preprocessing pipeline and inversion technique adopted from TF-ICON [2], **DENOISE** indicates each LDM's inference step, and **NORMALIZE** and **OPTIMIZE** correspond to the Adaptive Latent Normalization (Algorithm 1) and Energy-guided Latent Optimization (Algorithm 2) respectively introduced in Section 4.2 and Section 4.3 in the main paper.

2 ADDITIONAL QUALITATIVE RESULTS

We exhibit numerous extra qualitative results of image composition across different domains in Figures 5, 6, 7, 8, 9, 10, 11, 12, 13 for a more thorough comparison. Besides, we show additional ablation study results for component effectiveness in Figure 3 and for T' selection in Figure 4.

3 USER STUDY ELABORATION

Given the primary objective of subjectively evaluating the performance of the proposed TALE framework against existing SOTA and concurrent works, employing a ranking format for user study questions, as adopted in TF-ICON [2], would be redundant. This approach would be cumbersome and time-consuming for users as there are seven baselines to compare with, and some of which may generate visually similar composite results. Therefore, we opted for an either-or format for user study questions, as detailed in Section 5.2 of the main paper. This format offers enhanced user-friendliness while effectively demonstrating user preference for our method compared to alternative approaches.

Users are requested to select better option based on comprehensive criteria:

- **Foreground Content-Style Balance:** The composited image should well-preserve the identity features of given object within the user-specified mask region while its style adapts to that of the background.
- **Background Preservation:** The complement background area outside the mask should remain unchanged.
- **Text Alignment:** The composited image should conform to the given text prompt.
- **Seamless Composition:** The composited image should be visually pleasing and free from any noticeable artifacts, such that it is challenging for users to recognize it was produced by AI or copied and pasted.

Among 310 questions, there are 56 ones for oil painting, 56 ones for cartoon animation, 63 ones for sketching, 84 ones for photorealism from the baseline benchmark, and 51 ones for mixture of domains from the extended benchmark. The average preference percentage of TALE over work W (APP_W) is calculated by:

Algorithm 3 Training-free Image Composition - TALE

```
Input: Background and foreground images ($\mathbf{x}_{bg}, \mathbf{x}_{fg}$), object segmentation mask $\mathbf{M}_{obj}$, user-specified mask $\mathbf{M}_u$, prompt $\mathbf{P}$, selective timestep $T'$, threshold $\tau$
Output: Composition image $\mathbf{x}_{res}$
$\mathbf{x}^{p}_{fg}, \mathbf{M}^{z}_{obj}$ = PREPROCESS($\mathbf{x}_{fg}, \mathbf{M}_{obj}, \mathbf{M}_u$)
$\mathbf{z}^{bg}_0, \mathbf{z}^{fg}_0 = \mathcal{E}(\mathbf{x}_{bg}, \mathbf{x}^{p}_{fg})$
$\mathbf{z}^{bg}_T, \mathbf{z}^{fg}_T$ = INVERSE($\mathbf{z}^{bg}_0, \mathbf{z}^{fg}_0, T$)
for $t = T$ to $0$ do
    $\mathbf{z}^{bg}_{t-1}, \mathbf{z}^{fg}_{t-1}$ = DENOISE($\mathbf{z}^{bg}_t, \mathbf{z}^{fg}_t$)
    if $t == T'$ then
        $\mathbf{z}^{res}_t = \mathbf{z}^{bg}_t \odot (1 - \mathbf{M}^{z}_{obj}) + \mathbf{z}^{fg}_t \odot \mathbf{M}^{z}_{obj}$
        $\mathbf{z}^{res}_{t-1}$ = DENOISE($\mathbf{z}^{res}_t$)
    else if $T' - \tau \le t < T'$ then
        $\tilde{\mathbf{z}}^{res}_t$ = NORMALIZE($\mathbf{z}^{res}_t$)
        $\hat{\mathbf{z}}^{res}_{t-1}$ = OPTIMIZE($\tilde{\mathbf{z}}^{res}_t$)
    end if
end for
$\mathbf{x}_{res} = \mathcal{D}(\hat{\mathbf{z}}^{res}_0)$
return $\mathbf{x}_{res}$
```

$$APP_W = \frac{1}{N_W} \sum_{i=1}^{N_W} \frac{c_i}{n_i}, \tag{1}$$

where N_W is the number of questions of which one option is the composited result generated by TALE and the other by work W, c_i is the number of users select TALE's option for the i^{th} question, and n_i is the total number of user responses for that question.

4 ADDITIONAL ABLATION STUDIES

4.1 Adaptive Threshold λ_t

Figure 1 demonstrates the effects of setting different fixed values for λ_t. As stated in the main paper Section 4.2, introducing λ_t helps balance the content-style trade-off within the object region. We can observe that the higher the value for λ_t, the more the color tone of object adapts to the background but the less identifying information is preserved. Empirical findings suggest that progressively increasing the value of λ_t as the composition (denoising) process progresses leads to optimal outcomes across diverse domains. This configuration is motivated by the observation that at later timesteps, a significant portion of the object content information has already

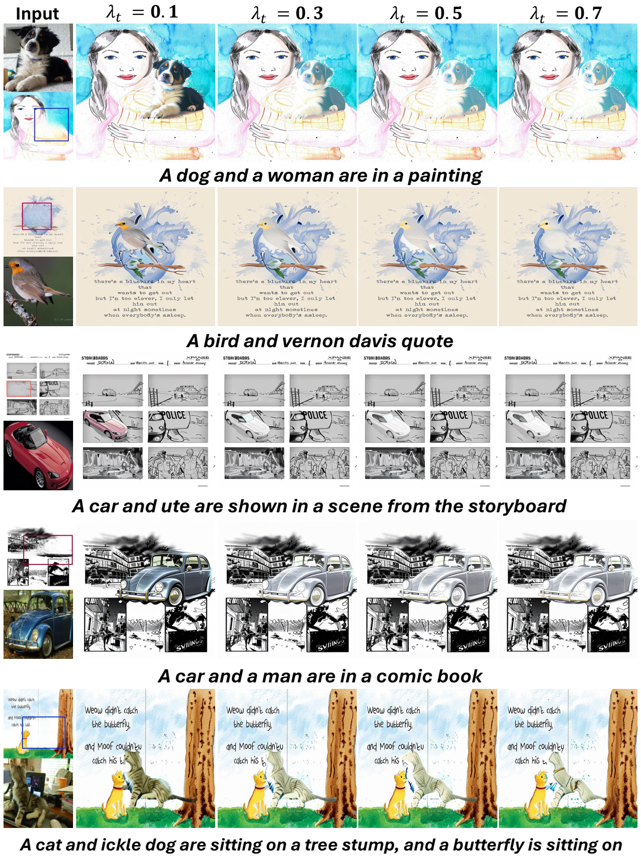

Input
$\lambda_t = 0.1$
$\lambda_t = 0.3$
$\lambda_t = 0.5$
$\lambda_t = 0.7$
A dog and a woman are in a painting
A bird and vernon davis quote
A car and ute are shown in a scene from the storyboard
A car and a man are in a comic book
A cat and ickle dog are sitting on a tree stump, and a butterfly is sitting on

Figure 1: Ablation study: Qualitative evaluation on the effect of adative threshold λ_t.

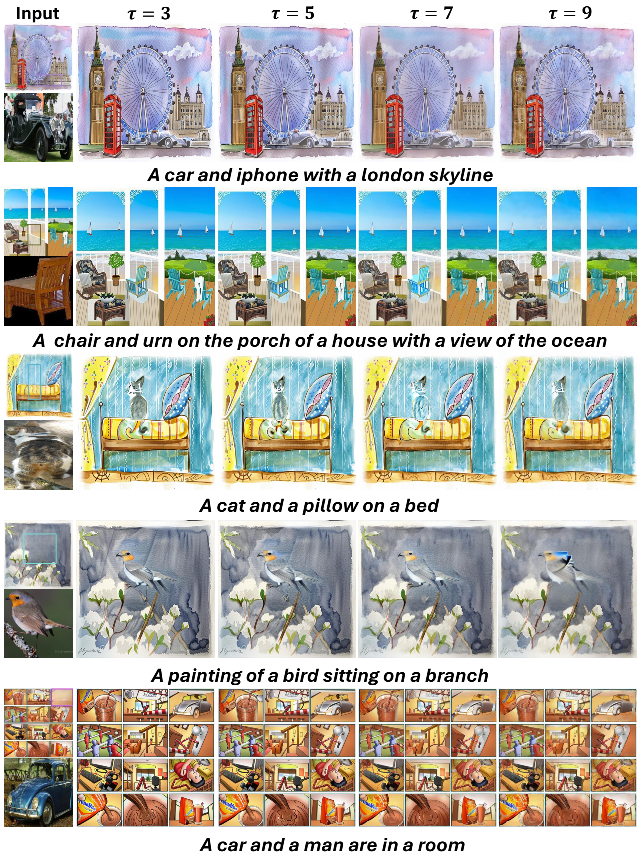

Input
$\tau = 3$
$\tau = 5$
$\tau = 7$
$\tau = 9$
A car and iphone with a london skyline
A chair and urn on the porch of a house with a view of the ocean
A cat and a pillow on a bed
A painting of a bird sitting on a branch
A car and a man are in a room

Figure 2: Ablation study: Qualitative evaluation on the effect of timestep constraint τ.

been incorporated. Consequently, a higher trade-off with style information becomes feasible, enabling the generation of more refined and stylistically consistent results.

4.2 Timestep Constraint τ

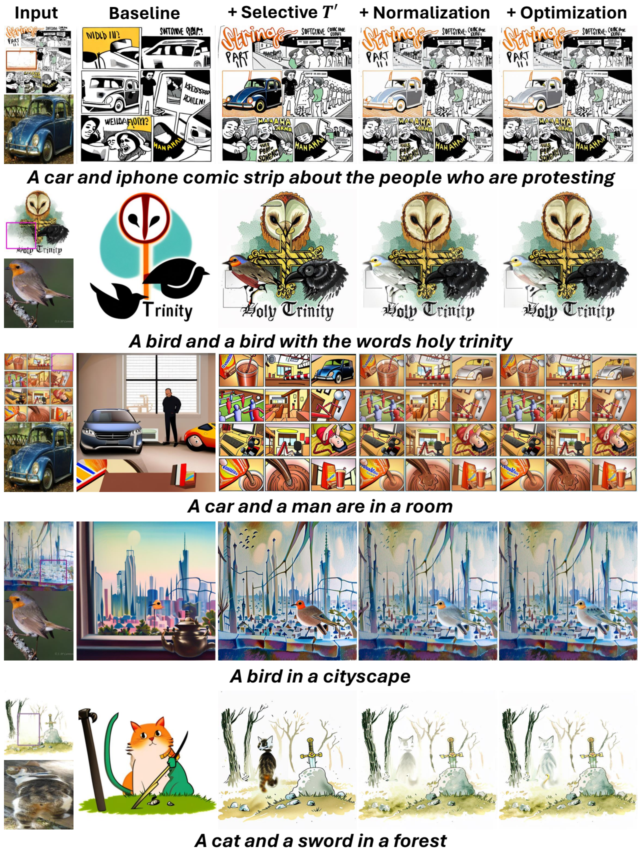
Figure 2 assesses the effects of setting different values for τ. As mentioned in the main paper Section 4.3, applying Adaptive Latent Normalization and Energy-guided Latent Op- timization for every timestep $t \in [0, T')$ may induce unwanted artifacts in transition area. Moreover, it is also observed that setting large value for τ can lead to loss of content information and thus content-style imbalance in object area.

5 LIMITATIONS AND FUTURE WORK

The primary limitation of our propose framework TALE is the inability to generate object views that deviate significantly from the provided reference image. Consequently, the selection of input objects images may be constrained in certain instances. This stems from the reliance of TALE on the preprocessed object segmentation mask $\mathbf{M}^z_{obj}$ for both Adaptive Latent Manipulation and Energy-guided Latent Optimization components. Though using the rescaled user-specified mask $\mathbf{M}^z_u$ as alternative can allow for some flexibility to incorporate objects of different views into composited results, this often compromises the preservation of object identity and induces unwanted artifacts in transition areas. To address this challenge, further research could exploit personalized concept learning methods, such as Textual Inversion [1] and InST [6], to encode identifying features of objects fused with background style information into special text embeddings which can require additional training or fine-tuning. Besides, it is worth mentioning that TALE leverages pre-trained LDM [3] hence also inherits its drawbacks and biases that may lead to undesired outcomes in certain scenarios.

6 SOCIETAL IMPACTS

TALE empowers individuals without professional artistic expertise to engage in image-guided composition. However, there are some potential risks associated with employing our framework. For instance, it could be misused for malicious purposes, such as harassment or dissemination of false information. Additionally, image composition is intrinsically linked to image generation, necessitating awareness of potential biases introduced by diffusion models trained on web-scraped data like LAION [5]. Notably, LAION unintendedly contains inappropriate NSFW contents. Consequently,



A car and iphone comic strip about the people who are protesting

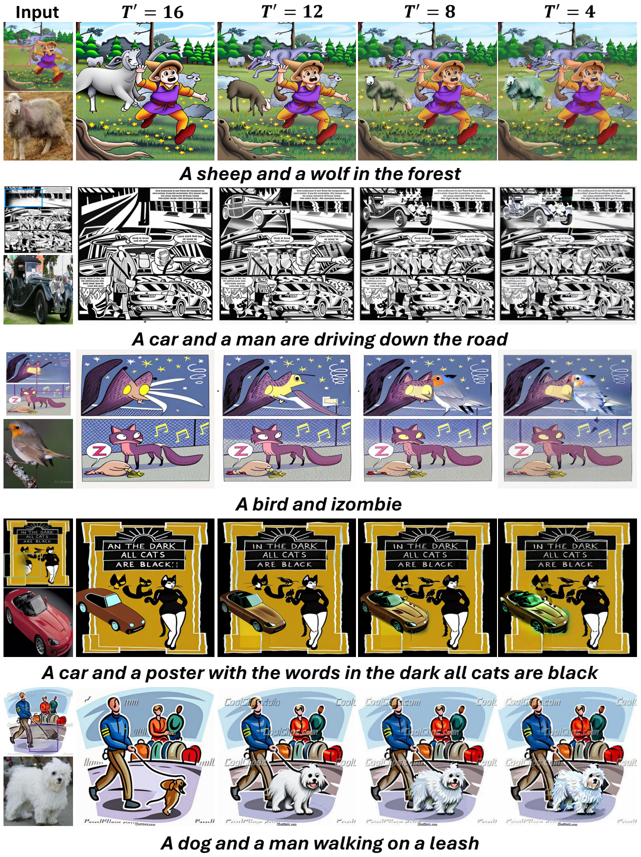
A bird and a bird with the words holy trinity

A car and a man are in a room

A bird in a cityscape

A cat and a sword in a forest

Figure 3: Ablation study: Qualitative evaluation on effectiveness of each component.



A sheep and a wolf in the forest

A car and a man are driving down the road

A bird and izombie

A car and a poster with the words in the dark all cats are black

A dog and a man walking on a leash

Figure 4: Ablation study: Qualitative evaluation on selection of T'.

diffusion models trained on LAION, such as LDM [3] and Imagen [4], may exhibit social and cultural biases. Therefore, using such models can raise ethical concerns hence warrants careful consideration. Finally, the ability to compose across artistic domains could be misapplied for copyright infringement, as users could create images of similar style without the artist's consent. While the generated artwork may currently be readily distinguishable from the original, future technologies could make such infringement challenging to differentiate. Consequently, we urge users to use this method with caution and only for legitimate purposes.

REFERENCES

[1] 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. https://doi.org/10.48550/ARXIV.2208.01618

[2] Shilin Lu, Yanzhu Liu, and Adams Wai-Kin Kong. 2023. TF-ICON: Diffusion-Based Training-Free Cross-Domain Image Composition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2294–2305.

[3] 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.

[4] 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.

[5] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. arXiv:2210.08402 [cs.CV]

[6] 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 (CVPR)*. 10146–10156.

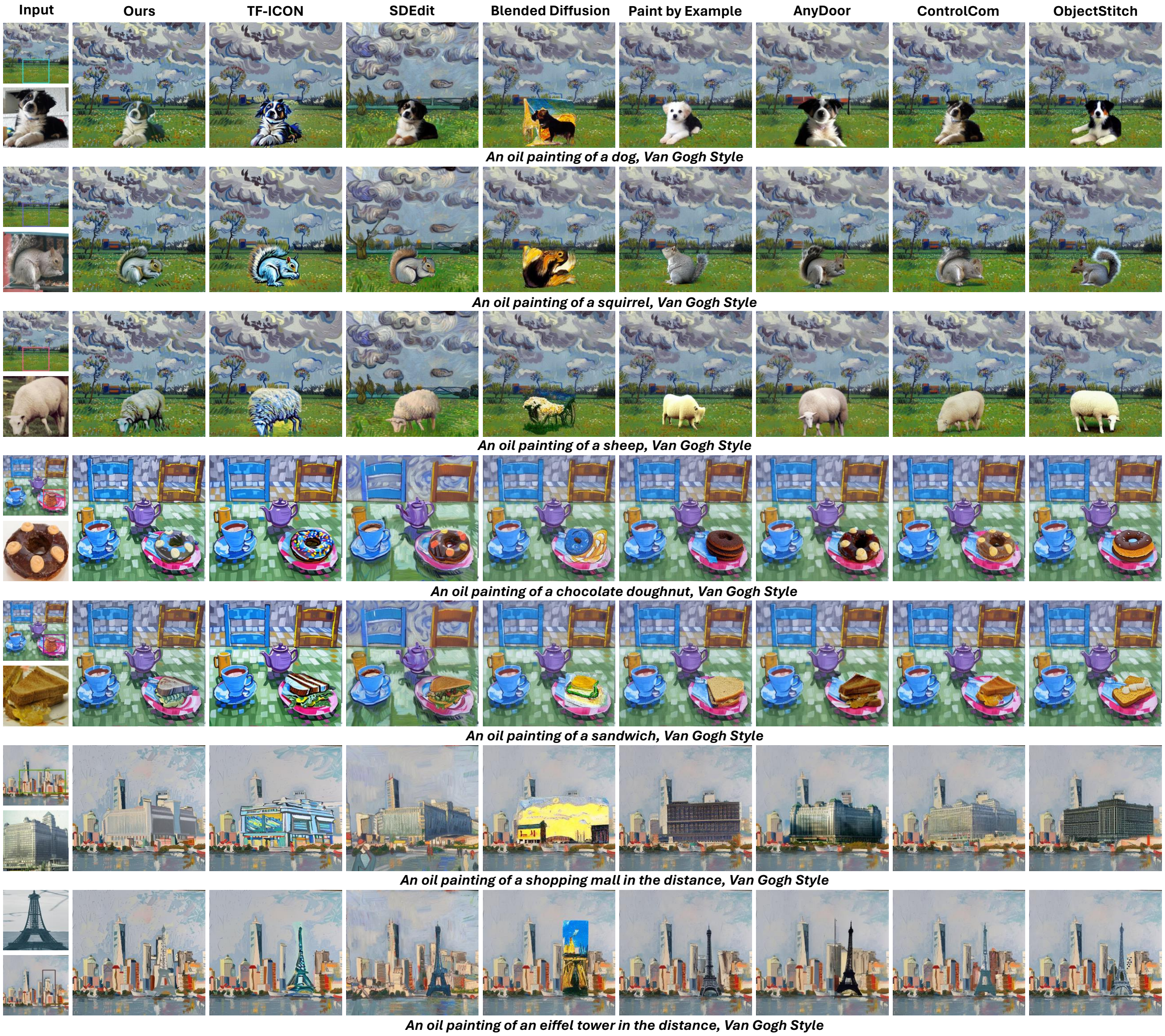

Input
Ours
TF-ICON
SDEdit
Blended Diffusion
Paint by Example
AnyDoor
ControlCom
ObjectStitch
An oil painting of a dog, Van Gogh Style
An oil painting of a squirrel, Van Gogh Style
An oil painting of a sheep, Van Gogh Style
An oil painting of a chocolate doughnut, Van Gogh Style
An oil painting of a sandwich, Van Gogh Style
An oil painting of a shopping mall in the distance, Van Gogh Style
An oil painting of an eiffel tower in the distance, Van Gogh Style

Figure 5: Qualitative comparison with prior SOTA and concurrent works in image composition for the oil painting domain on baseline benchmark. Zoom-in for details.

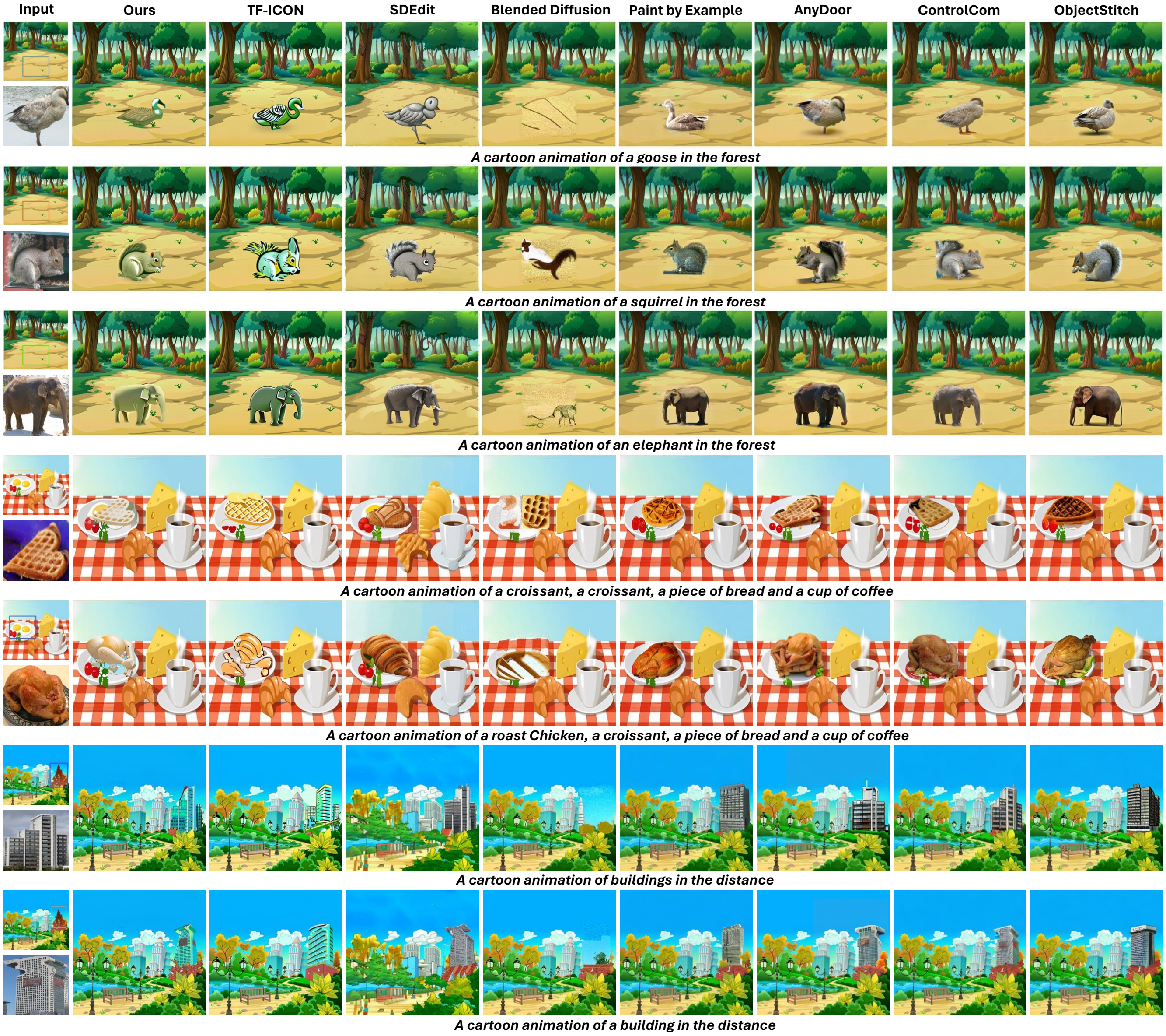

Input
Ours
TF-ICON
SDEdit
Blended Diffusion
Paint by Example
AnyDoor
ControlCom
ObjectStitch
A cartoon animation of a goose in the forest
A cartoon animation of a squirrel in the forest
A cartoon animation of an elephant in the forest
A cartoon animation of a croissant, a croissant, a piece of bread and a cup of coffee
A cartoon animation of a roast Chicken, a croissant, a piece of bread and a cup of coffee
A cartoon animation of buildings in the distance
A cartoon animation of a building in the distance

Figure 6: Qualitative comparison with prior SOTA and concurrent works in image composition for the cartoon animation domain on baseline benchmark. Zoom-in for details.

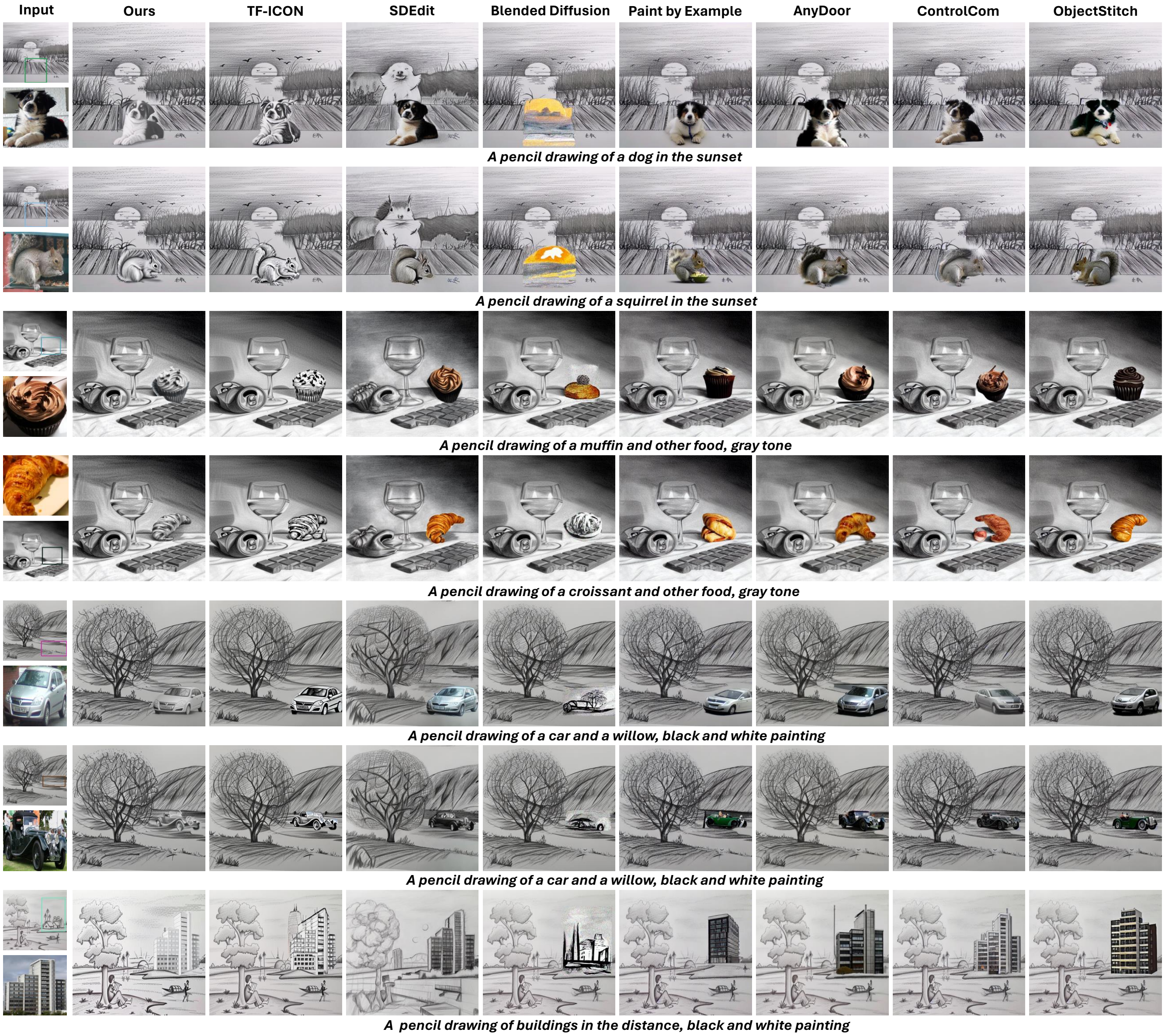

Input
Ours
TF-ICON
SDEdit
Blended Diffusion
Paint by Example
AnyDoor
ControlCom
ObjectStitch
A pencil drawing of a dog in the sunset
A pencil drawing of a squirrel in the sunset
A pencil drawing of a muffin and other food, gray tone
A pencil drawing of a croissant and other food, gray tone
A pencil drawing of a car and a willow, black and white painting
A pencil drawing of a car and a willow, black and white painting
A pencil drawing of buildings in the distance, black and white painting

Figure 7: Qualitative comparison with prior SOTA and concurrent works in image composition for the sketching domain on baseline benchmark. Zoom-in for details.

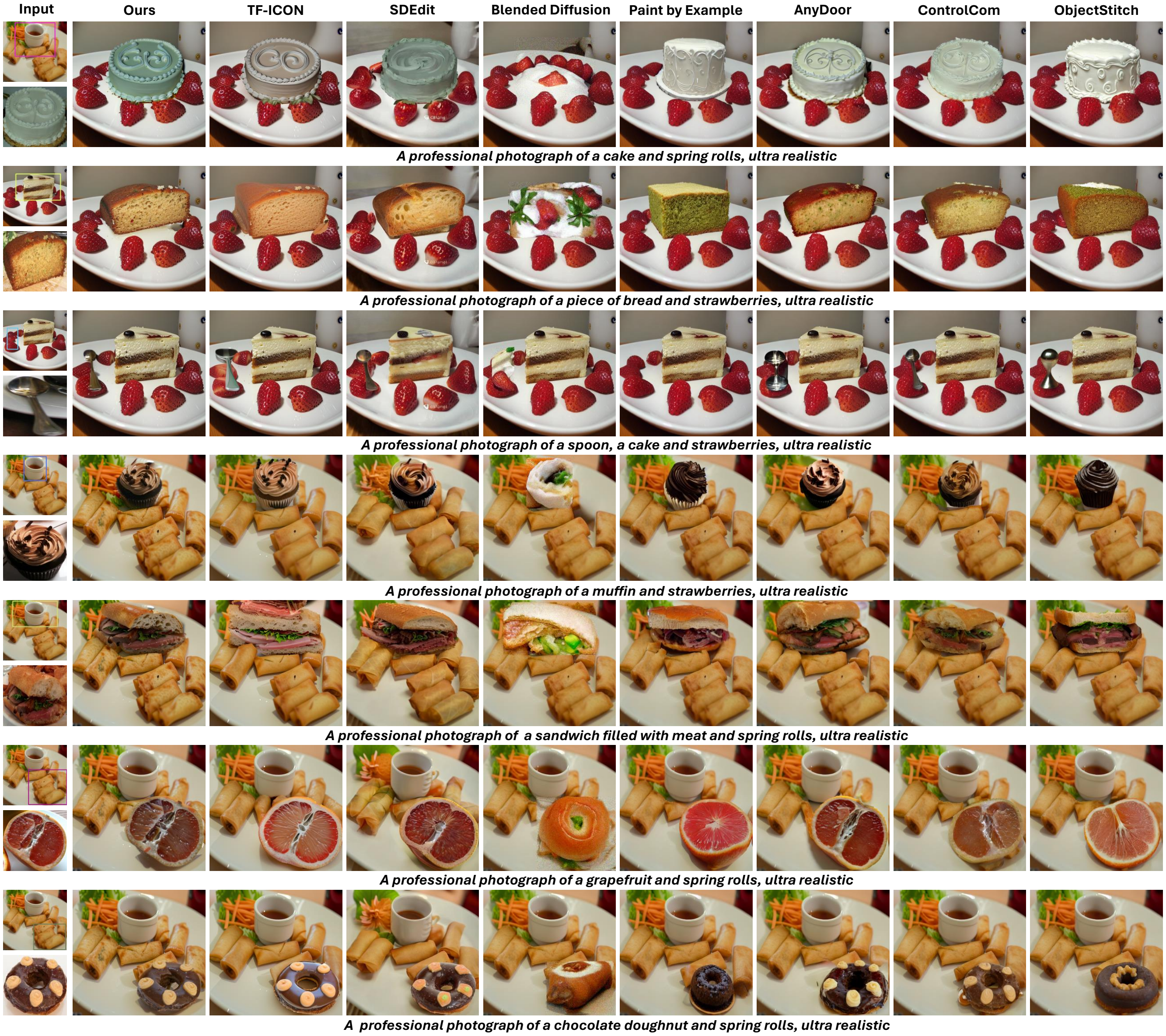

Input
Ours
TF-ICON
SDEdit
Blended Diffusion
Paint by Example
AnyDoor
ControlCom
ObjectStitch
A professional photograph of a cake and spring rolls, ultra realistic
A professional photograph of a piece of bread and strawberries, ultra realistic
A professional photograph of a spoon, a cake and strawberries, ultra realistic
A professional photograph of a muffin and strawberries, ultra realistic
A professional photograph of a sandwich filled with meat and spring rolls, ultra realistic
A professional photograph of a grapefruit and spring rolls, ultra realistic
A professional photograph of a chocolate doughnut and spring rolls, ultra realistic

Figure 8: Qualitative comparison with prior SOTA and concurrent works in image composition for the photorealism domain on baseline benchmark. Zoom-in for details.

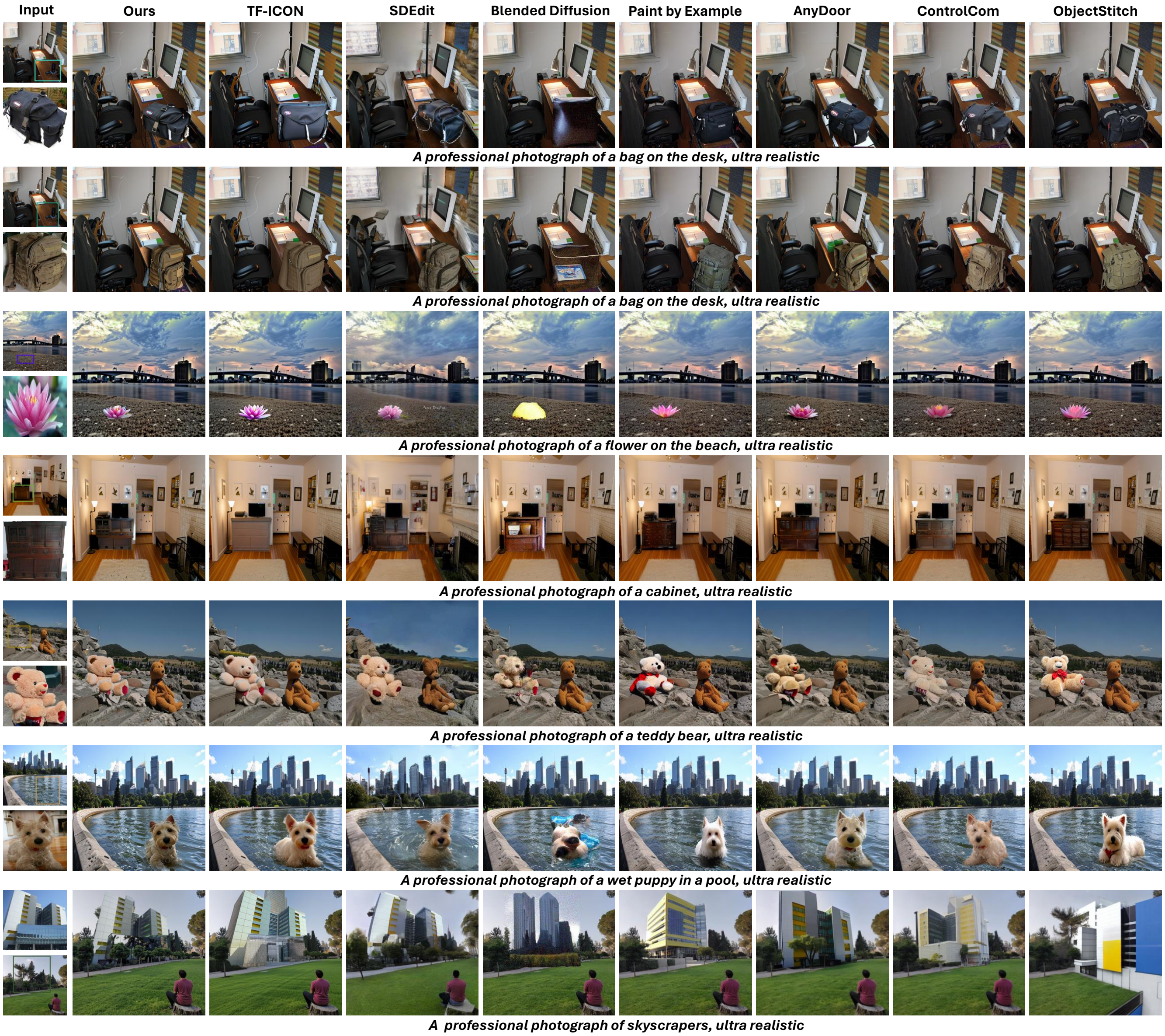

Input
Ours
TF-ICON
SDEdit
Blended Diffusion
Paint by Example
AnyDoor
ControlCom
ObjectStitch
A professional photograph of a bag on the desk, ultra realistic
A professional photograph of a bag on the desk, ultra realistic
A professional photograph of a flower on the beach, ultra realistic
A professional photograph of a cabinet, ultra realistic
A professional photograph of a teddy bear, ultra realistic
A professional photograph of a wet puppy in a pool, ultra realistic
A professional photograph of skyscrapers, ultra realistic

Figure 9: Qualitative comparison with prior SOTA and concurrent works in image composition for the photorealism domain on baseline benchmark. Zoom-in for details.

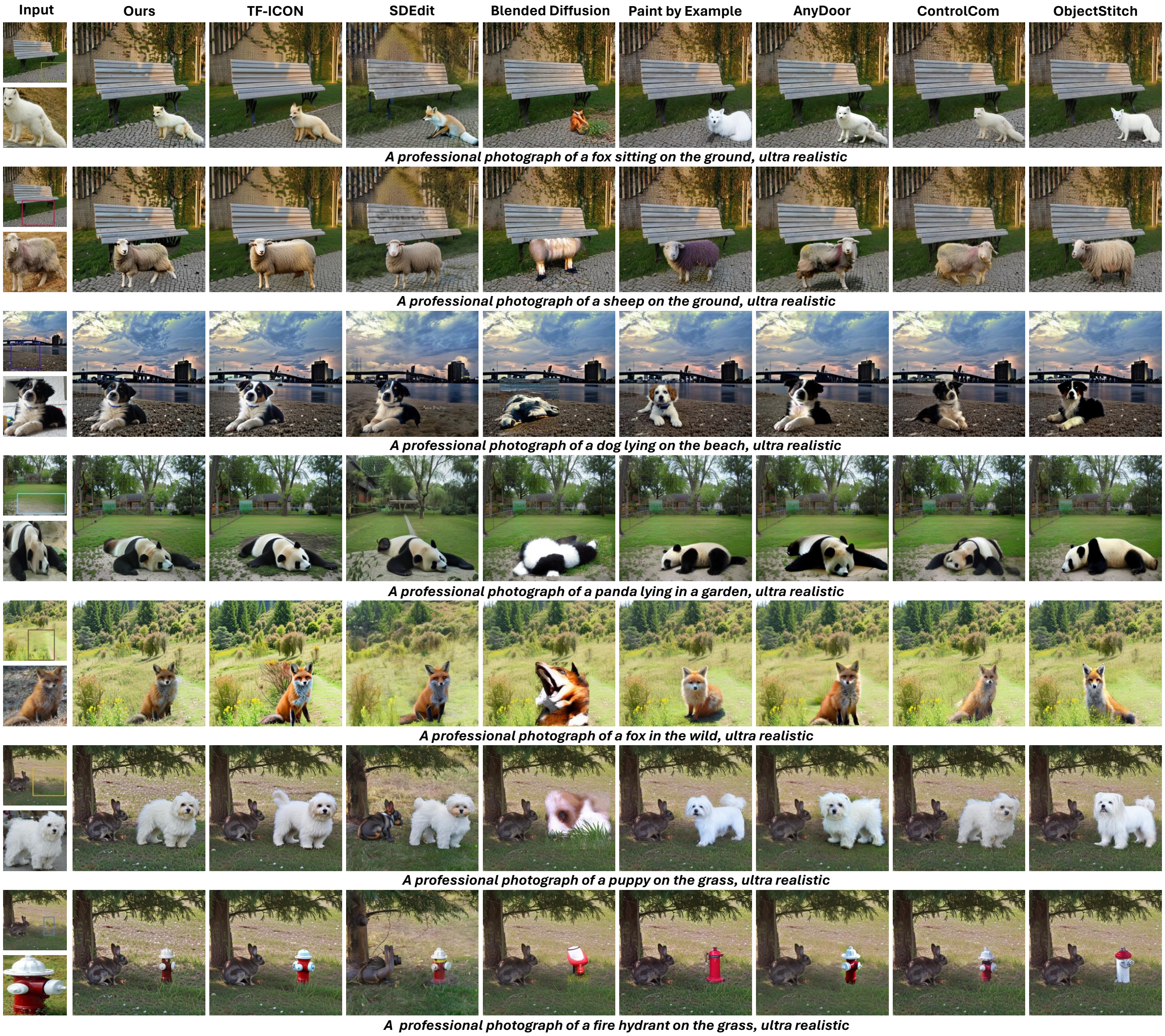

Input
Ours
TF-ICON
SDEdit
Blended Diffusion
Paint by Example
AnyDoor
ControlCom
ObjectStitch
A professional photograph of a fox sitting on the ground, ultra realistic
A professional photograph of a sheep on the ground, ultra realistic
A professional photograph of a dog lying on the beach, ultra realistic
A professional photograph of a panda lying in a garden, ultra realistic
A professional photograph of a fox in the wild, ultra realistic
A professional photograph of a puppy on the grass, ultra realistic
A professional photograph of a fire hydrant on the grass, ultra realistic

Figure 10: Qualitative comparison with prior SOTA and concurrent works in image composition for the photorealism domain on baseline benchmark. Zoom-in for details.

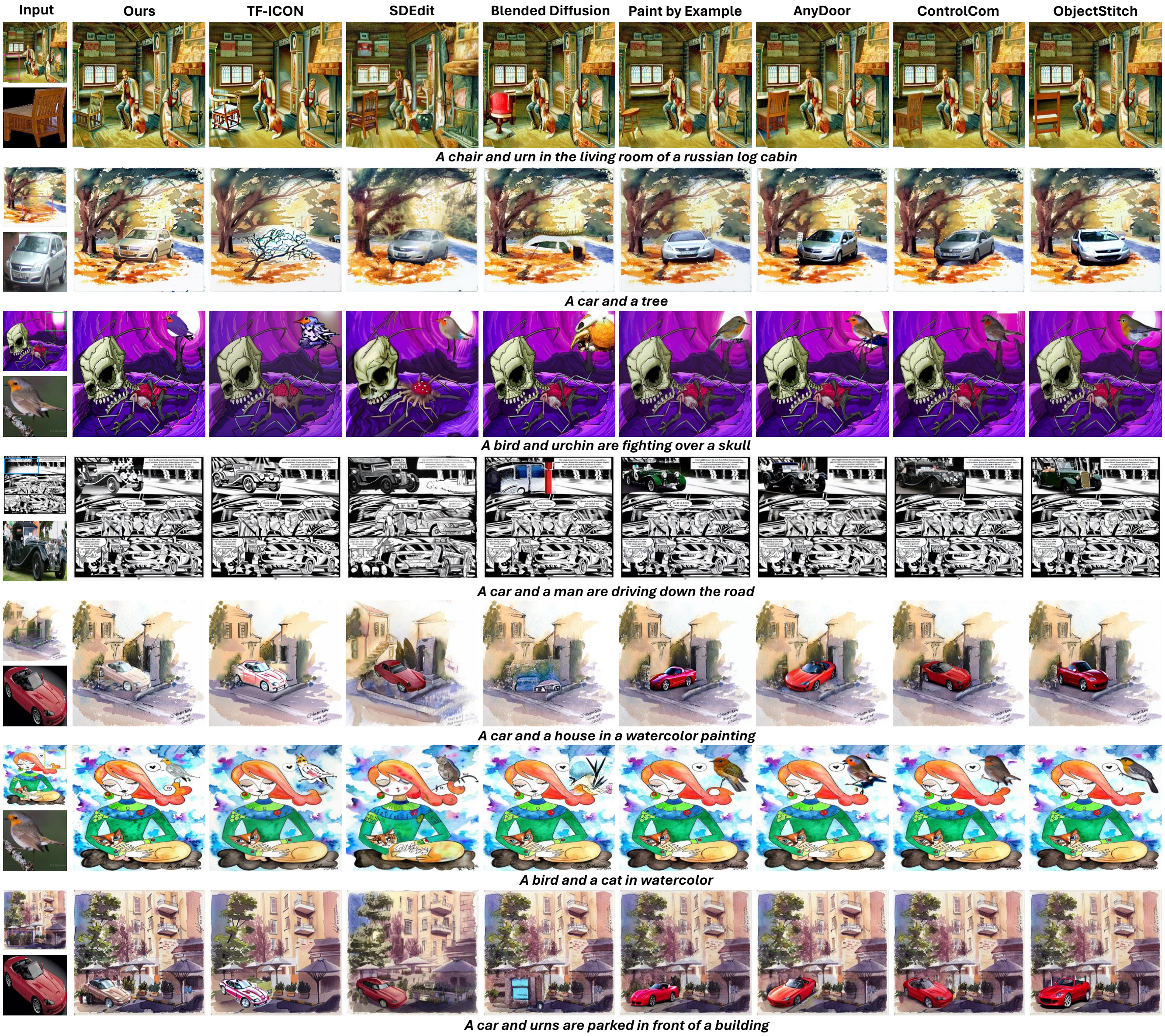
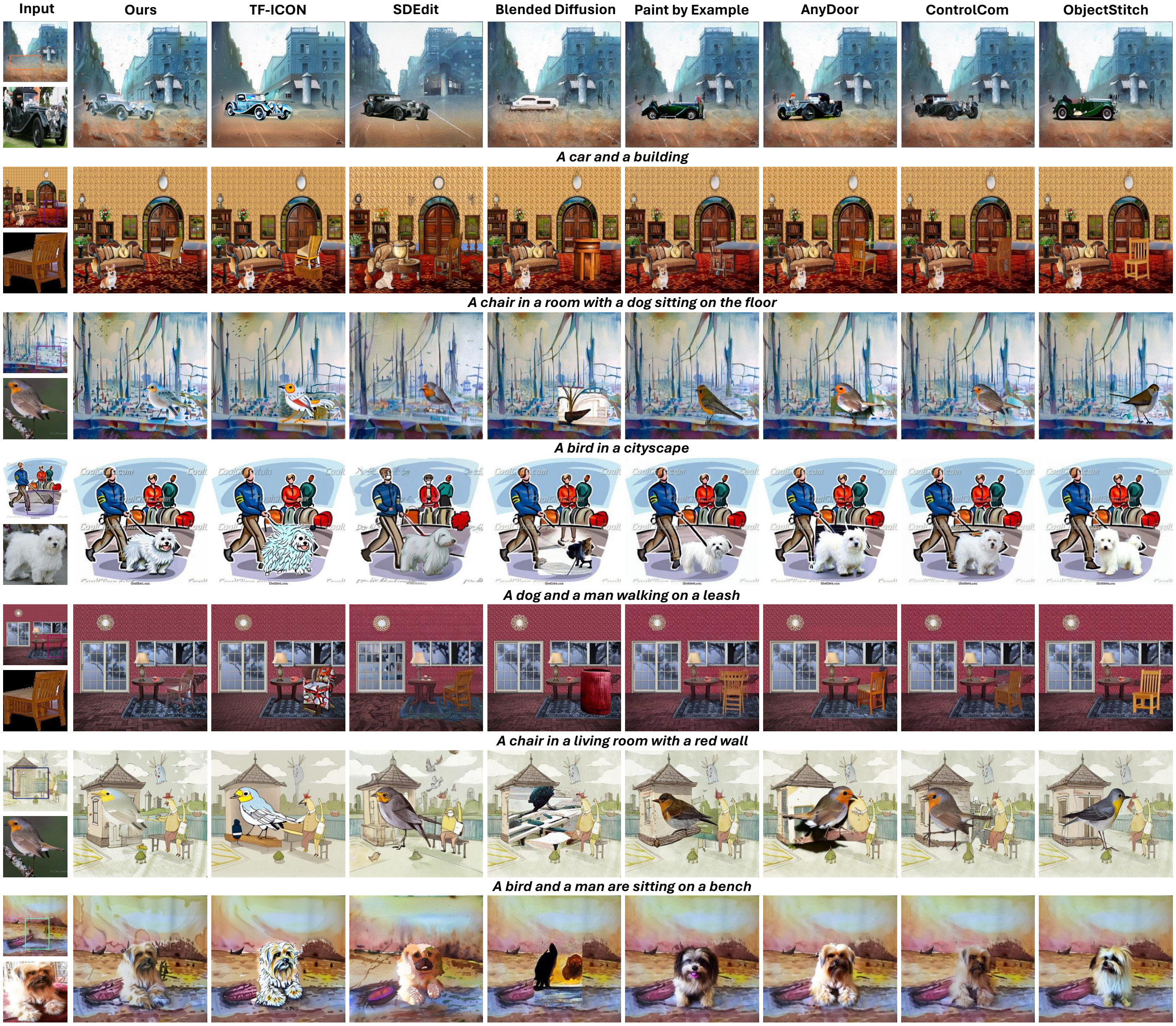

Input
Ours
TF-ICON
SDEdit
Blended Diffusion
Paint by Example
AnyDoor
ControlCom
ObjectStitch
A chair and urn in the living room of a russian log cabin
A car and a tree
A bird and urchin are fighting over a skull
A car and a man are driving down the road
A car and a house in a watercolor painting
A bird and a cat in watercolor
A car and urns are parked in front of a building

Figure 11: Qualitative comparison with prior SOTA and concurrent works in image composition for mixture of domains on extended dataset. Zoom-in for details.



A car and a building

A chair in a room with a dog sitting on the floor

A bird in a cityscape

A dog and a man walking on a leash

A chair in a living room with a red wall

A bird and a man are sitting on a bench

A dog at sunset, watercolor painting

Figure 12: Qualitative comparison with prior SOTA and concurrent works in image composition for mixture of domains on extended dataset. Zoom-in for details.

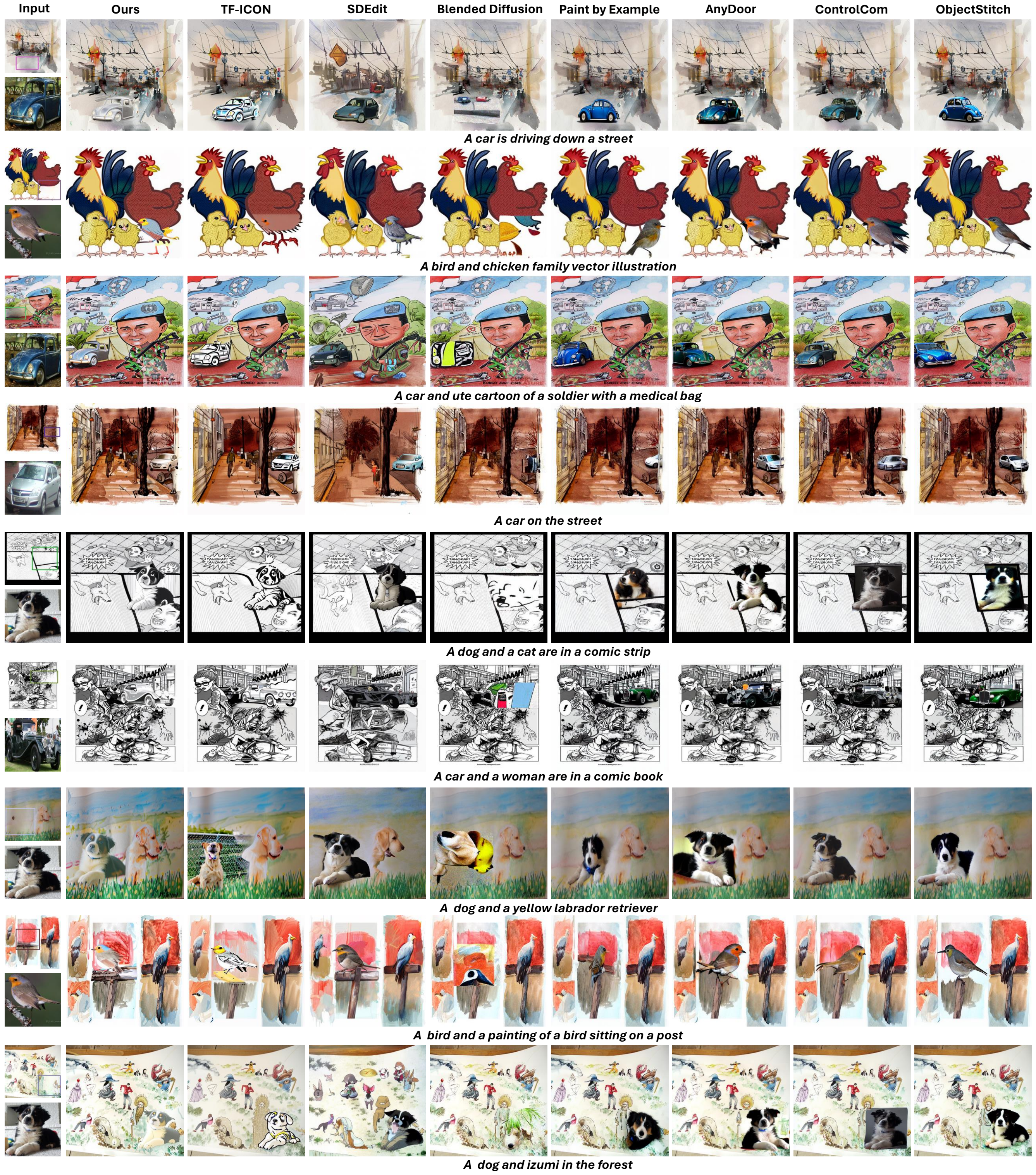

Input
Ours
TF-ICON
SDEdit
Blended Diffusion
Paint by Example
AnyDoor
ControlCom
ObjectStitch
A car is driving down a street
A bird and chicken family vector illustration
A car and ute cartoon of a soldier with a medical bag
A car on the street
A dog and a cat are in a comic strip
A car and a woman are in a comic book
A dog and a yellow labrador retriever
A bird and a painting of a bird sitting on a post
A dog and izumi in the forest

Figure 13: Qualitative comparison with prior SOTA and concurrent works in image composition for mixture of domains on extended dataset. Zoom-in for details.