Composing Concepts from Images and Videos via Concept-prompt Binding

Supplementary Material

This document includes more details, extra experimental results, corresponding analyses, and further discussions of BiCo. The document is organized as follows:

- §A provides detailed VLM prompts for the prompt diversification process in DAM.
- §B gives more details on the user studies.
- §C provides more details on the two-stage inverted training strategy and conducts further ablations.
- §D explains the justifications for the absorbent token and provides empirical evidence.
- §E illustrates more qualitative comparisons.
- §F performs another case study to facilitate the understanding of different components of BiCo.
- §G further discusses the limitations with failure cases and the societal impacts of BiCo.

Please refer to the webpage for video results.

A. Detailed Prompts for DAM

In the prompt diversification process, we utilize a powerful VLM Qwen2.5-VL [1] to generate diversified concept prompts while retaining the key conceptual words unchanged. During the key concept extraction stage, the VLM is asked to extract essential spatial and temporal concepts from the visual inputs. For image inputs, we use the following textual prompt to extract spatial concepts:

> You are an image captioning specialist whose goal is to extract the concepts in words or phrases that compose the input image. You need to adhere to the formatting of the examples provided strictly.
> **Task Requirements**:
> 1. Concepts stand for names of objects, colors, styles, etc;
> 2. The overall output should be in English;
> 3. The concepts should be brief but concrete, each concept is either a single word or a small phrase. Avoid vague concepts such as "background";
> 4. You should be precise and concise;
> 5. You should output all the extracted concepts within a "spatial" category as the example.
>
> **Example of the concept output**:
> {"spatial": ["brown cat", "sunglasses", "sketch", "sunny", "grassland"]}
> Please output in JSON format (pure text, without markdown formatting).

For video inputs, the following textual prompt is adopted to extract both spatial and temporal concepts:

> You are a video captioning specialist whose goal is to extract the spatial and temporal concepts in words or phrases that compose the input video. You need to adhere to the formatting of the examples provided strictly.
> **Task Requirements**:
> 1. Spatial concepts stand for names of objects, colors, styles, etc;
> 2. Temporal concepts refer to the motion, transitions, and probably viewpoint changes in the video;
> 3. The overall output should be in English;
> 4. The concepts should be brief but concrete, each concept is either a single word or a small phrase. Avoid vague concepts such as "background";
> 5. You should be precise and concise;
> 6. You should output all the extracted concepts within a "spatial" category as the example.
>
> **Example of the concept output**:
> {"spatial": ["brown cat", "sunglasses", "sketch", "sunny", "grassland"], "temporal": ["jumping", "running", "falling", "gently flowing", "bright to dark", "near to far"]}
> Please output in JSON format (pure text, without markdown formatting).

During the spatiotemporal concept composition stage, the VLM is asked to combine the extracted concepts into a number of full prompts according to the visual input. For images and the first-stage training of videos with a focus on spatial concepts, we use the following prompts:

> You are an image captioning specialist whose goal is to write high-quality English prompts by referring to the extracted concepts and the input image, making them complete and expressive.
> **Task Requirements**:
> 1. Use the given concepts to describe the image in a concise sentence;
> 2. You should make sure that the generated caption matches the image content;
> 3. You can rearrange or paraphrase these concepts to form diverse captions;
> 4. No matter what language the user inputs, you must always output in English.
>
> **Example of the English captions**:
> 1. A boat in a river, with trees and houses on the

riverbank, and a foggy sky.
2. A large brown bear in front of a rocky enclosure. The backdrop features a rustic stone wall and scattered boulders.
3. A human pose standing with arms crossed in front of a black background.

Directly output the English caption text.

For the second-stage training of videos, the following prompt is adopted:

You are a video captioning specialist whose goal is to write high-quality English prompts by referring to the extracted spatial and temporal concepts and the input video, making them complete and expressive.

Task Requirements:
1. Use the given concepts to describe the video in a concise sentence;
2. You should make sure that the generated caption matches the video content;
3. You can rearrange or paraphrase these concepts to form diverse captions;
4. No matter what language the user inputs, you must always output in English.

Example of the English captions:
1. A boat sailing in a river, creating white ripples in the water, with trees and houses on the riverbank, and a foggy sky.
2. A large brown bear ambles slowly across a rocky enclosure. The backdrop features a rustic stone wall and scattered boulders.
3. A human pose standing with arms crossed in front of a black background, turning slowly from left to right.

Directly output the English caption text.

B. User Study Details

We recruited volunteers from various backgrounds to conduct the user study. Each user is given a subset of 10 groups of test cases and is asked to rate the concept consistency, prompt fidelity, and motion quality on a 5-point Likert scale. The detailed questions are as follow:

- **Concept Preservation**: How well do you think that the composed video preserves the concepts from the corresponding visual sources?
- **Prompt Fidelity**: How well do you think that the composed video follows the input prompt?
- **Motion Quality**: Please rate the motion quality of the generated video. You can consider the motion smoothness, consistency, naturalness, etc. Please note that **still frames without motion are considered low quality**.

Table 1. **Extra Ablations on Two-stage Inverted Training Strategy** (§C). Results in **bold** are the best.

Two-stage	Inverted	Concept↑	Prompt↑	Motion↑	Overall↑
		2.60	2.70	2.43	2.58
✓		3.53	3.77	3.53	3.61
✓	✓	**4.43**	**4.47**	**4.32**	**4.40**

C. Extra Details and Ablations on Two-stage Inverted Training Strategy

The probability distribution for the discretized timestep t_i in inverted training is:

$$p(t_i) = \begin{cases} (1-\beta) \cdot \frac{1}{N_{<\alpha}} & , d(t_i) < \alpha \\ \beta \cdot \frac{1}{N_{\geq \alpha}} & , d(t_i) \geq \alpha \end{cases}, \tag{1}$$

where $d(t_i) \in [0, 1]$ indicates the position of t_i in the scheduler, and N_* is the total number of discretized timesteps in the interval $*$. $\alpha = 0.875$ is selected according to the training recipe of Wan2.2 to distinguish higher and lower noise levels. While β can be selected in a reasonable range to emphasize the higher noise levels, we empirically found that setting $\beta = \alpha$ exchanges the total probability mass between the higher and lower noise levels and yields satisfactory performance given that higher noise levels originally account for a smaller probability than lower noise levels.

We provide additional quantitative ablation results under the same settings in §4.3 to facilitate understanding of the two-stage inverted training strategy. Results are shown in Tab. 1, where *Two-stage* means that training the global binder before training the whole hierarchical binder structure, and *Inverted* stands for focusing more on high noise levels in the first stage. We can observe that both techniques are crucial for achieving satisfactory optimization of the binders.

D. Analysis on Absorbent Token

In T2V models, text tokens are already associated with corresponding visual concepts as a good initial value for further personalization. This association is the foundation for our binders to learn sample-specific features. With a new absorbent token, it is expected that the model encodes irrelevant information into this token instead of other tokens with good initialization for corresponding visual concepts. The absorbent token is expected not to capture specific concepts, but to prevent other conceptual tokens from being distracted from established initial associations.

We demonstrate the effectiveness of the absorbent token by reconstructing a target image with the trained binder and visualizing the cross-attention maps of the target subject tokens (Akita dog) and the absorbent token in Fig. 1. As ob-

served, the absorbent token does capture irrelevant details like plants. Removing the trained absorbent token during inference also enhances attention on the target.

Figure 1. Visualizations of cross-attention maps of target subject tokens (Akita dog) (§D).

E. Additional Qualitative Comparisons

We provide more composed videos in Fig. 2 for additional qualitative comparisons with other methods. Fig. 2a demonstrates a motion transfer task, where Textual Inversion [2] and DreamVideo [5] fails to combine the visual concepts. DualReal [4] suffers from inadequate visual concept preservation and unintended concept leakage (*e.g.*, the green leaves). Although DB-LoRA [3] mostly follows the designated prompt to integrate visual concepts, there are significant drifts of visual concepts from the original inputs (*e.g.*, the direction of the squirrel). BiCo achieves the best result in composing the visual concepts according to the given prompt while maintaining the consistency of visual concepts with the input image and video.

Fig. 2b illustrates a creative style transfer task to integrate the line art sketch style with the subject in a video. All previous methods [2–5] fail in this task to learn and compose the style concept. This sample further verifies the flexible versatile controllability of BiCo.

F. Extra Case Study

We further illustrate the functions of BiCo's components with another concrete visual concept composition sample in Fig. 3. Comparing #2 to #1, we can observe that the hierarchical binder structure enables our method to encode more visual information into binders, resulting in better concept preservation results. The prompt diversification operation (#3) and the absorbent token (#4) in DAM enhance the accuracy of concept-prompt binding, better preserving background details in the composed videos. The effectiveness of the absorbent token can also be verified by the enhanced background preservation in #7 compared to #5. TDS further improves the composition quality by enhancing the compatibility between image and video concepts, as illustrated by comparing #7 to #4 and #5 to #3. The two-stage inverted training strategy significantly stabilizes the optimization process, bringing considerably better results in the same optimization steps (#7 to #6). The video results can be found in the webpage.

G. More Discussions

G.1. Limitations

The significance of each prompt token for T2V generation is unevenly distributed. Some tokens that represent subjects and motions play a more important role than the function words. In addition, when a concept is visually complex or deviates significantly from the *average looking* of the text token, the binder's representation capability for each token may be insufficient to accommodate all the visual information. Nevertheless, BiCo treats each token equally in the concept composition process, which can result in unintended concept drifts. For instance, in the upper part of Fig. 4, BiCo fails to accurately reproduce the colorful whimsical hat in the composed video, where the hat's appearance differs considerably from an average hat. We plan to integrate adaptive designs to highlight critical tokens in our future work.

Furthermore, BiCo also falls short when the composition requires some common sense reasoning. For example, the composed video in the lower part of Fig. 4 simply adds an additional leg to the Doberman Pinscher to hold the gun instead of raising an existing leg, resulting in a total of 5 legs in a single dog. This issue may be alleviated by integrating the strong reasoning capabilities of VLMs to design a more comprehensive captioning and composing paradigm.

G.2. Societal Impacts

BiCo enables flexible visual concept composition for both images and videos through a one-shot paradigm, enabling practitioners to experiment with visual concepts from multiple sources to implement their creativity. For individual creators, the one-shot nature of our method allows them to integrate AI-assisted visual content composition into their workflows without extensive training. For commercial teams, our method provides them with a new opportunity to flexibly combine their intermediate results and other assets, boosting the novelty of the produced visual content.

On the other hand, with BiCo's powerful capability to manipulate visual concepts, it can be used to produce fabricated images and videos that appear highly realistic, posing significant challenges for verifying the authenticity of visual media. Such content can distort public perception and raise privacy concerns when fake contents featuring an individual are generated in an unauthorized way.

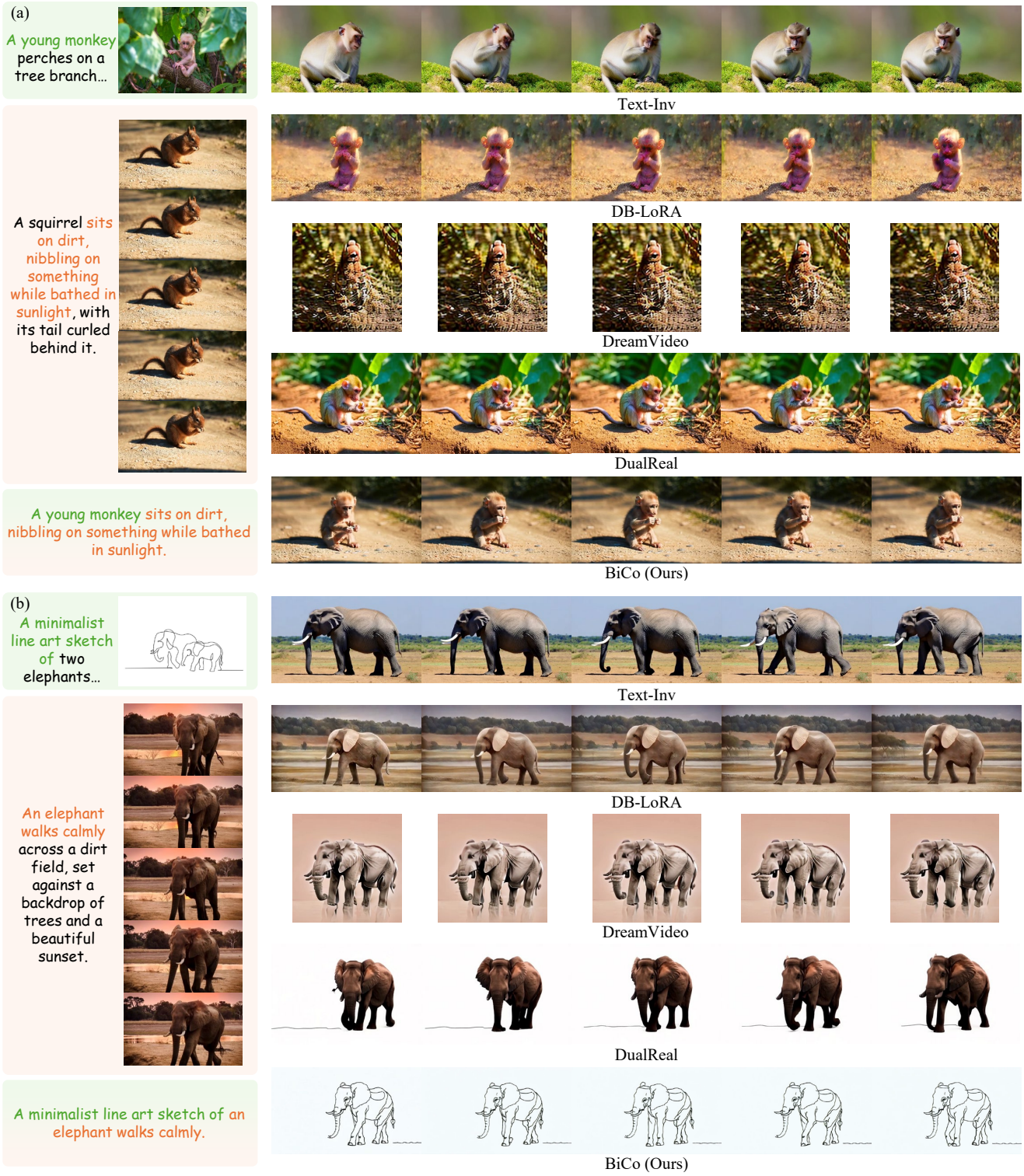

Figure 2. **Additional Qualitative Comparisons** (§E). The input visual concepts and composed prompts are on the left.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jian-

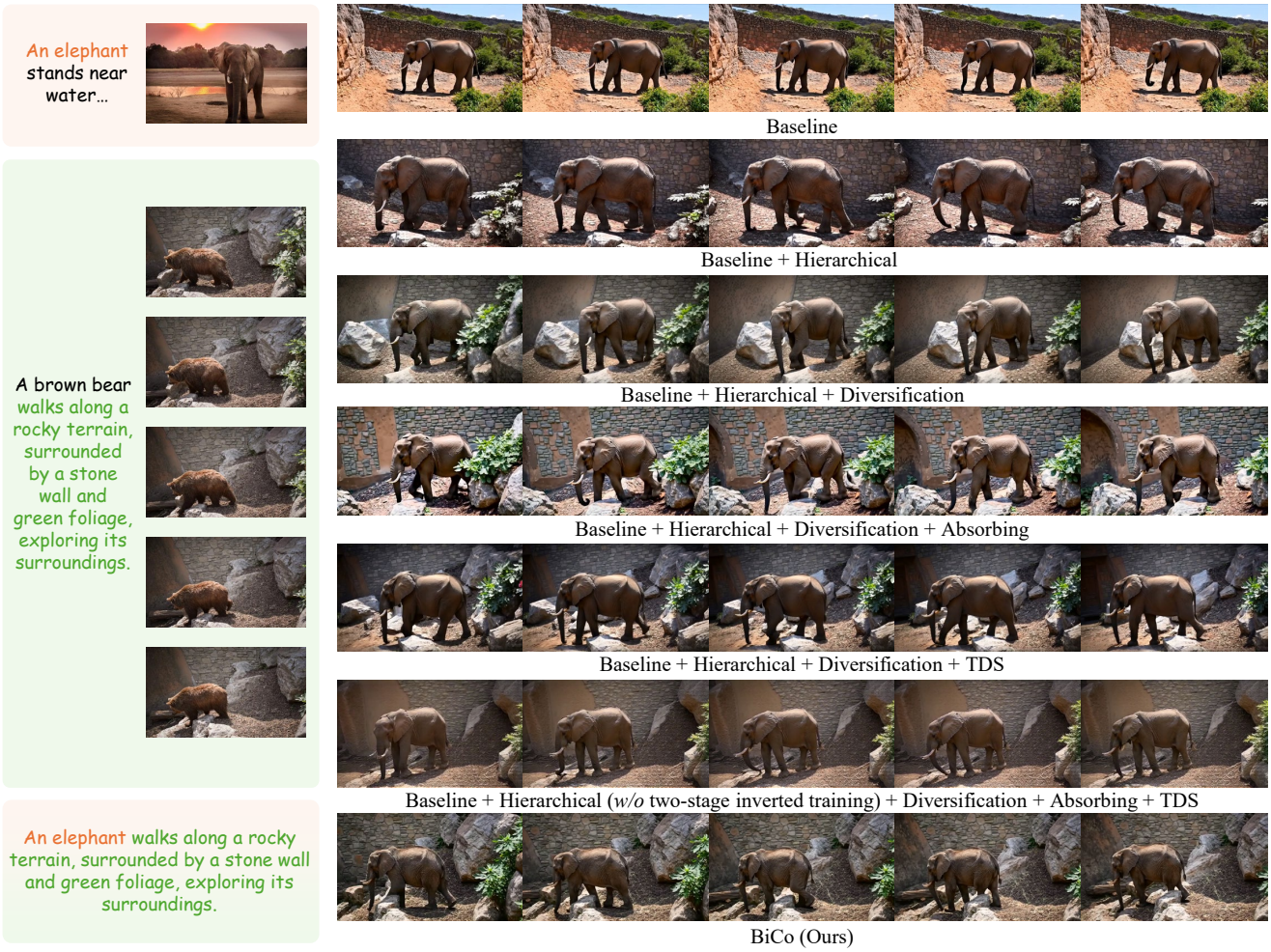

Figure 3. **Extra Case Study for Components** (§F). The input visual concepts and composed prompts are on the left.

qiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. 1

[2] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023. 3

[3] 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 *CVPR*, pages 22500–22510, 2023. 3

[4] Wenchuan Wang, Mengqi Huang, Yijing Tu, and Zhendong Mao. Dualreal: Adaptive joint training for lossless identity-motion fusion in video customization, 2025. 3

[5] Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. Dreamvideo: Composing your dream videos with customized subject and motion. In *CVPR*, pages 6537–6549, 2024. 3

Figure 4. **Failure Cases** (§G.1). In each case, the upper row shows the visual inputs, and the lower row presents the composed video.