Composing Concepts from Images and Videos via Concept-prompt Binding

Xianghao Kong[1] , Zeyu Zhang[1] , Yuwei Guo[2] , Zhuoran Zhao[1,3] , Songchun Zhang[1] , Anyi Rao[1]
[1] HKUST [2] CUHK [3] HKUST(GZ)

https://refkxh.github.io/BiCo_Webpage

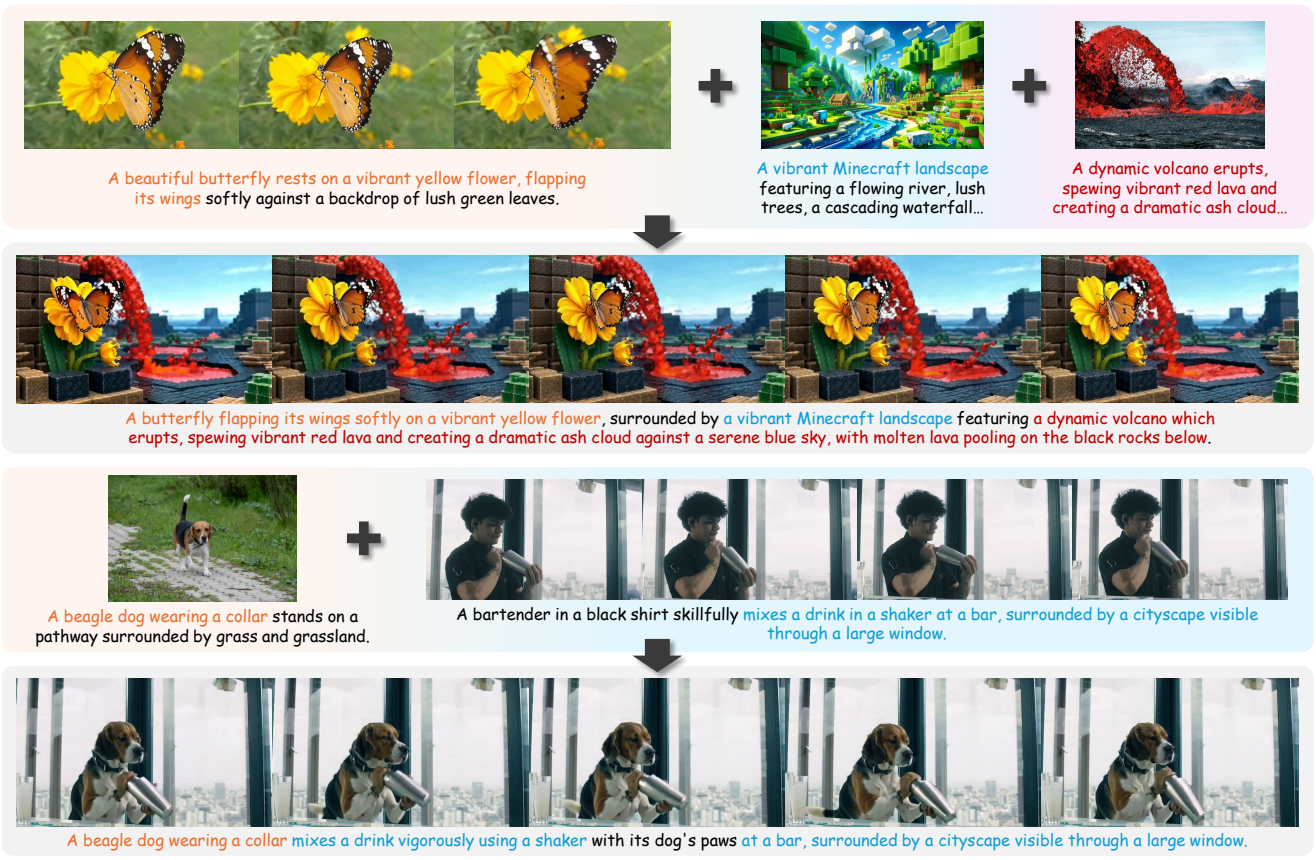

Figure 1. Illustration of BiCo, a one-shot method that enables flexible visual concept composition by binding visual concepts with the corresponding prompt tokens and composing the target prompt with bound tokens from various sources (§1).

Abstract

Visual concept composition, which aims to integrate different elements from images and videos into a single, coherent visual output, still falls short in accurately extracting complex concepts from visual inputs and flexibly combining concepts from both images and videos. We introduce Bind & Compose, a one-shot method that enables flexible visual concept composition by binding visual concepts with corresponding prompt tokens and composing the target prompt with bound tokens from various sources. It adopts a hierarchical binder structure for cross-attention conditioning in Diffusion Transformers to encode visual concepts into corresponding prompt tokens for accurate decomposition of complex visual concepts. To improve concept-token binding accuracy, we design a Diversify-and-Absorb Mechanism that uses an extra absorbent token to eliminate the impact of concept-irrelevant details when training with diversified prompts. To enhance the compatibility between image and video concepts, we present a Temporal Disentanglement Strategy that decouples the training process of video concepts into two stages with a dual-branch binder structure for temporal modeling. Evaluations demonstrate that our method achieves superior concept consistency, prompt fidelity, and motion quality over existing approaches, opening up new possibilities for visual creativity.

1. Introduction

Visual concept composition aims to integrate different elements from images and videos into a single, coherent visual output. This process is a reflection of human artists' creation: combining ingredients from various inspirations to form a brand new masterpiece [15]. Consequently, it plays a fundamental role in visual creativity and filmmaking [62]. With the rapid advancement of diffusion-based visual content generation models [16, 20, 29–31, 35, 40, 42, 54, 58, 61, 63], an increasing number of works [1, 3, 11, 14, 18, 26, 32–34, 55, 56] have been exploring the field of visual concept composition by exploiting the generative models' strong capability of concept grounding and customization.

Despite considerable efforts devoted to this field, challenges still remain in accurately extracting complex concepts from visual inputs and flexibly combining concepts from both images and videos. First, the capability to precisely extract specific concepts from various sources is of great significance for visual content creators. Nevertheless, existing mainstream methods [1, 3, 14, 26, 32, 34, 55, 56] use either adapters like LoRA [25] or learnable embeddings with explicit or implicit masks to realize concept selection, which fall short in decoupling complex concepts with occlusions and temporal alterations, and extracting non-object concepts such as styles. Second, it is a common practice to integrate different visual elements from both images and videos in the visual content creation process [62]. However, previous works are confined to animating designated subjects from images with motion from videos [26, 55, 56], without further exploration of flexibly combining various attributes (*e.g.*, visual styles and lighting variations) from both images and videos. Although there has been recent effort on flexible concept composition [18] in the image domain, achieving universal visual concept composition for both images and videos remains an underexplored problem.

To this end, we introduce Bind & Compose (BiCo), a one-shot method that enables flexible visual concept composition by binding visual concepts with the corresponding textual tokens, with satisfactory compatibility between image and video concepts (Fig. 1). Our method first leverages the powerful concept grounding capability [59] of text-to-video (T2V) diffusion models [54] to bind textual tokens with their corresponding visual concepts through one-shot training, achieving implicit decomposition without mask input. Then, concept composition is done through selecting any desired bound tokens from various sources and composing them into the final prompt tokens, which serves as the model condition. This paper mainly encompasses the following three technical contributions: **First**, to achieve reliable decomposition of complex visual concepts for flexible manipulation, we propose a hierarchical binder structure for the cross-attention mechanism [52] in Diffusion Transformer (DiT) [41] blocks to effectively encode visual concepts into corresponding textual tokens. When composing concepts from multiple sources, concept tokens in the target prompt are passed through different binders correspondingly to integrate visual features, enabling text-conditioned concept composition without explicit mask input. **Second**, to improve the accuracy of concept-token binding for more precise concept decomposition, we design a Diversify-and-Absorb Mechanism (DAM) that diversifies the one-shot prompts while retaining key concepts, and introduces an extra absorbent token during training to eliminate the impact of concept-irrelevant details. **Third**, to enhance the compatibility between image and video concepts during composition, we present a Temporal Disentanglement Strategy (TDS) that decouples the training process of video concepts into two stages. In the first stage, the binders are trained with individual frames without temporal concepts, which aligns with the training setting of image concepts. In the second stage that trains the binders on videos, we adopt a dual-branch binder structure to better cater to temporal concepts while inheriting knowledge from the first stage.

Extensive experiments demonstrate that BiCo simultaneously achieves superior concept consistency, prompt fidelity, and motion quality when performing visual concept composition. It also outperforms previous baseline approaches in both concept manipulation flexibility and visual quality of the composed video. With support for a variety of innovative video creation tasks, BiCo demonstrates a strong potential to serve as a promising solution for creators to experiment with their whimsies.

2. Related Work

T2V Diffusion Models. The emergence of diffusion models [24, 46, 50] has significantly advanced the realm of visual content generation. Recently, DiT [41] has become the de facto standard of the denoising model's architecture, surpassing U-Net [47] with its strong scaling capability [28] and flexibility for integrating multi-modal conditions [16]. Flow Matching [37, 38] introduces a new linear paradigm to transit between the Gaussian distribution and the target distribution, improving the theoretical properties and simplifying the conceptual framework. Based on these works, a number of T2V diffusion models [20, 29, 42, 54, 58, 61] have emerged with text-to-video cross-attention or joint attention for conditioning. Despite these methods achieving satisfactory quality and consistency of generation, they are designed for general T2V generation, with limited support for personalization and concept composition.

Video Personalization. Video personalization aims to integrate the appearance or motion of a designated object into a pre-trained video generation model, enabling the model to reproduce these properties when generating with other prompts. Building upon the progress in the image domain [9, 27, 48], several approaches handle the temporal

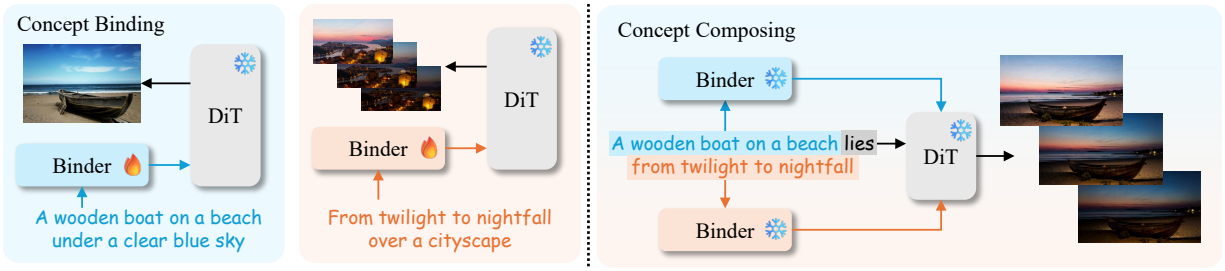

Figure 2. **Overview of BiCo** (§3.1). BiCo first adopts a binder structure to learn visual concepts into corresponding prompt tokens, and then composes different concepts by passing corresponding prompt tokens through different adapters for the updated prompt as condition.

consistency problem by adding LoRAs [25] to the temporal layers of T2V models [8, 22] or learning the motion embeddings from reference videos [39, 56, 64]. Set-and-Sequence [4] enables the simultaneous learning of both appearance and motion from a single video by designing the spatio-temporal weight space within the LoRA architecture. Grid-LoRA [2] further enables reusable video personalization by introducing a grid-based LoRA system that spatially organizes input and output. However, we cannot accurately designate the concept to extract and the way the concepts are combined. The number and type of inputs are also confined, limiting the flexibility of concept composition.

Visual Concept Composition. Composing multiple visual elements from images and videos into a coherent output remains a challenging task. There have been early explorations in decomposing image concepts [5, 21, 53]. Break-A-Scene [5] relies on explicit mask inputs to achieve concept decomposition, which limits its availability to common users and its ability to extract non-object concepts. Other methods [21, 53] extract multiple concepts from a single image by jointly learning several tokens, each corresponding to a visual concept. However, the content learned by each token is unpredictable. To achieve concept composition in the image domain, existing works either use explicit spatial conditioning [32, 34, 60], which falls short in overlapping or non-object concepts, or fuse multiple LoRAs [19, 44, 49], which restricts the type and number of concepts to compose or requires joint optimization among all source images. TokenVerse [18] learns a modulation term for each text token to achieve prompt-controlled concept composition. Despite enhanced flexibility, it relies on text-conditioned modulation architectures in DiT [41] models, limiting its universality to modern T2V models [51, 54]. To extend concept composition to handle videos, previous methods [26, 55, 56] incorporate dedicated designs to decouple appearance and motion, supporting only the composition of subjects from images and motions from videos. BiCo simultaneously enables complex concept decomposition (non-object concepts and multiple concepts from a single input) and flexible concept composition (selective composition via prompts and composing image and video concepts), offering endless possibilities for visual creators.

3. Methodology

3.1. Overview

Given M concept images or videos $\{\boldsymbol{V}_c^j\}_{j=1}^M$ with their corresponding textual prompt tokens $\{\boldsymbol{p}_c^j\}_{j=1}^M$, BiCo aims at composing the visual concepts from the inputs according to the designated prompt $\boldsymbol{p}_d$ to generate a coherent visual output $\boldsymbol{V}$. As illustrated in Fig. 2, it first learns each text token's corresponding visual appearance or motion via a light-weight binder module for each visual input, and then combines tokens from different source images or videos to generate a target video that composes the individual concepts. Specifically, during concept binding, a binder structure attached to a DiT-based T2V model [54] is utilized to encode the correspondence between visual concepts and textual tokens through one-shot training on different inputs $\{\boldsymbol{V}_c^j\}_{j=1}^M$ and $\{\boldsymbol{p}_c^j\}_{j=1}^M$ respectively. When integrating concepts from various sources, different parts of the designated prompt $\boldsymbol{p}_d$ representing visual concepts are passed through their corresponding binders to compose the updated prompt $\boldsymbol{p}_u$, which contains visual concept information and is then fed into DiT blocks to serve as the condition for the composed visual output.

Preliminary: Text Conditioning in T2V Models. Current mainstream T2V models [20, 29, 42, 54, 58, 61] adopt the DiT [41] architecture with tens of blocks to predict the denoising vector. Each DiT block contains attention layers, an MLP, and a modulation mechanism for the timestep condition. To achieve text conditioning, a prevalent method is to insert a cross-attention layer in each DiT block, which takes the latent tokens $\boldsymbol{x}_{in}$ as queries and the textual prompt tokens $\boldsymbol{p}$ as keys and values:

$$\boldsymbol{x}_{out} = \texttt{cross\_attention}(\boldsymbol{x}_{in}, \boldsymbol{p}, \boldsymbol{p}), \quad (1)$$

where $\boldsymbol{x}_{out}$ stands for the updated latent tokens. Through the cross-attention process, the textual information is injected into the DiT model and serves as the condition when predicting the denoising vector.

3.2. Hierarchical Binder Structure

To fully exploit the powerful capability of visual-text association within T2V models for accurate decomposition of

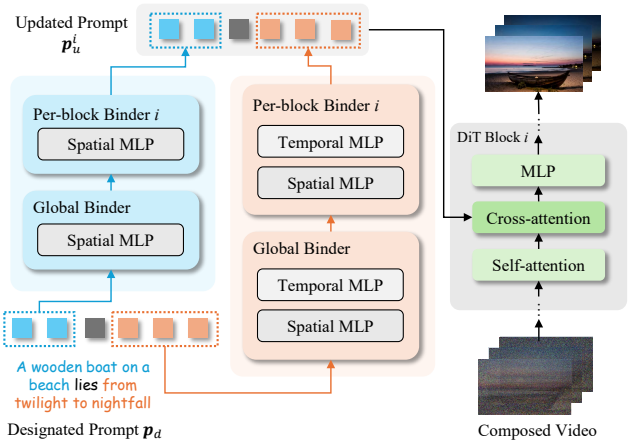

Figure 3. **Hierarchical Binder Structure** (§3.2). It consists of global and per-block binders, where each binder contains an MLP with residual connections. For video inputs, a dual-branch binder structure with spatial and temporal MLPs is incorporated to better address temporal concepts.

complex visual concepts, we attach binders to DiT cross-attention conditioning layers to encode visual concepts into corresponding prompt tokens. Since DiT blocks have distinct behaviors during the denoising process [57], a hierarchical binder structure is designed with a global binder for the overall association and per-block binders for tailored association (Fig. 3). Specifically, each binder $f(\cdot)$ consists of an MLP with a zero-initialized learnable scaling factor γ in a residual style, and takes the prompt tokens $\boldsymbol{p}$ as input:

$$f(\boldsymbol{p}) = \boldsymbol{p} + \gamma \cdot \texttt{MLP}(\boldsymbol{p}). \tag{2}$$

For video inputs, a dual-branch binder structure with spatial and temporal MLPs is incorporated to better address temporal concepts (detailed in §3.4). For the training process, the concept prompt tokens $\boldsymbol{p}_c$ are first passed through a global binder $f_g(\cdot)$ for a global update to get $\boldsymbol{p}_g$. Then, for the i-th DiT block, $\boldsymbol{p}_g$ are fed into a per-block binder $f_l^i(\cdot)$ to obtain the updated prompt $\boldsymbol{p}_u^i$, which are used as the key and value inputs for the cross-attention layer. For the inference process, we first decompose the designated prompt tokens $\boldsymbol{p}_d$ according to the correspondence with visual concepts, and then feed each concept-related part into the corresponding binder. Finally, we compose the updated prompt $\boldsymbol{p}_u^i$ with the result of each concept binder. This design enables flexible manipulation of visual concepts by composing the designated prompt $\boldsymbol{p}_d$.

Two-stage Inverted Training Strategy. Recent studies point out that the denoising process of diffusion models is divided into several stages with different functions [13, 36]. It has been discovered that prioritizing the training on higher noise levels yields better performance [13]. To this end, we utilize a two-stage inverted training strategy to enhance the optimization process for hierarchical binders. Specifically, we define a noise level threshold α to separate high and low noise levels. In the first stage, we only train the global binder with the probability of α for high noise levels ($\geq \alpha$) and the probability of $1-\alpha$ for low noise levels ($< \alpha$). This setting emphasizes the high noise levels and reduces the optimization steps on low noise levels. In the second training stage, both global and per-block binders are trained without inverting the probability of noise levels.

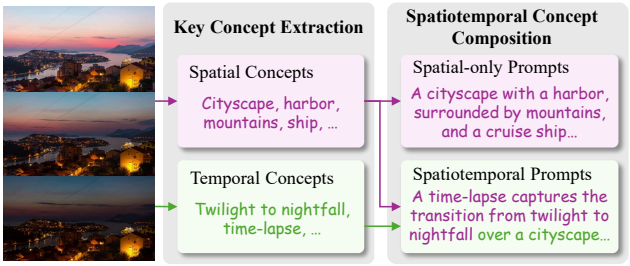

Figure 4. **Prompt Diversification** (§3.3). The VLM extracts key spatial and temporal concepts from the visual input, and then composes them into diverse spatial-only or spatiotemporal prompts.

3.3. Diversify-and-Absorb Mechanism (DAM)

Establishing accurate concept-token bindings is a notable challenge, especially in one-shot cases. To enable precise association between concepts and prompt tokens in binders, we introduce DAM, which takes advantage of the powerful visual comprehension and reasoning capability of Vision-language Models (VLMs) [7] to diversify concept prompts while retaining the key conceptual words unchanged during the concept binding process. As shown in Fig. 4, the prompt diversification process is divided into two stages: key concept extraction and spatiotemporal concept composition. In the key concept extraction stage, the VLM is asked to extract critical concepts from the input image or video and divide them into spatial and temporal concepts. During spatiotemporal concept composition, the VLM composes the extracted concepts into a designated number of diverse prompts with the visual input reference. For images and the first-stage training of videos with a focus on spatial concepts (detailed in §3.4), only spatial concepts are used to form the full prompt. For the second-stage training of videos, the VLM uses both spatial and temporal concepts to generate the complete prompt.

The diversified prompts may not cover all the details in the visual inputs, and those uncovered visual elements can entangle with other prompt tokens, resulting in degraded concept-prompt binding quality. To address this issue, a learnable absorbent token is introduced to minimize the impact of concept-irrelevant details during concept binding by absorbing those distracting visual details. Concretely, when binding the j-th visual concept source $\boldsymbol{V}_c^j$ with the corresponding textual prompt tokens $\boldsymbol{p}_c^j$, we initialize an absorbent token p_a^j, and concatenate it with $\boldsymbol{p}_c^j$ along the sequence dimension as the input of the hierarchical binder

structure. The embeddings of the token p_a^j are updated with other learnable parameters during optimization. When it comes to concept composing, the absorbent token p_a^j is discarded to suppress undesired details.

3.4. Temporal Disentanglement Strategy (TDS)

The ability to compose concepts from both images and videos is of great significance to visual content creators. However, significant temporal heterogeneity exists between images and videos [12], especially the temporal domain shift caused by the absence of motion in images. This hinders compatibility when directly composing concepts from both sources. To enable flexible composition of image and video concepts with satisfactory quality, we devise TDS, which aligns the learning paradigm of images and videos by decoupling the training process for video concepts into two stages. In the first stage, we train the binders on individual video frames without temporal concepts in the input prompt. This setting remains the same as the training setting of image concepts, with a focus on binding spatial concepts. In the second stage that trains the binders on full videos and complete prompts with temporal concepts, we adopt a dual-branch binder structure to decouple the learning of spatial and temporal concepts and inherit the knowledge from the first stage. Specifically, we extend the MLP in the original binder with an extra temporal MLP branch $\texttt{MLP}_t$, and then fuse them with a learnable gating module $g(\cdot)$:

$$\texttt{MLP}(\boldsymbol{p}) \leftarrow (1 - g(\boldsymbol{p})) \cdot \texttt{MLP}_s(\boldsymbol{p}) + g(\boldsymbol{p}) \cdot \texttt{MLP}_t(\boldsymbol{p}), \quad (3)$$

where the weights of $\texttt{MLP}_s$ are taken from the first stage and $g(\cdot)$ is zero-initialized to provide the optimization process with a good initial state. Such a disentanglement strategy alleviates the temporal heterogeneity between images and videos and achieves better results when composing concepts from both images and videos.

4. Experiments

4.1. Implementation Details

We select Wan2.1-T2V-1.3B [54] as the base model to apply BiCo. The MLP structure in binders consists of two linear layers with layer normalization [6] and GELU [23] activation. The binders are trained with a learning rate of 1.0×10^{-4} for 2400 iterations per stage. The noise level threshold α in §3.2 is set to 0.875. We set the length for composed videos to 81 frames during inference. All other hyperparameters remain the same as Wan2.1 [54]. Experiments are conducted on NVIDIA RTX 4090 GPUs.

4.2. Comparisons to Prior Works

To demonstrate the superiority of BiCo over existing visual concept composition works, we conduct quantitative and qualitative comparisons with 4 representative methods:

Table 1. **Quantitative Comparisons with Prior Arts** (§4.2.1). Results in **bold** are the best. † Implemented on Wan2.1 [54].

Method	CLIP-T↑	DINO-I↑	Concept↑	Prompt↑	Motion↑	Overall↑
Text-Inv†	25.96	20.47	2.14	2.17	2.94	2.42
DB-LoRA†	30.25	27.74	2.76	2.76	2.51	2.68
DreamVideo	27.43	24.15	1.90	1.82	1.66	1.79
DualReal	31.60	32.78	3.10	3.11	2.78	3.00
BiCo (Ours)	**32.66**	**38.04**	**4.71**	**4.76**	**4.46**	**4.64**

Textual Inversion (Text-Inv) [17], DreamBooth-LoRA (DB-LoRA) [48], DreamVideo [56], and DualReal [55]. We adapt Text-Inv and DB-LoRA on the same T2V model [54] as BiCo to support video concepts. Since existing methods that support both images and videos only take one image (subject) and one video (motion) as input, we limit our comparisons to composing concepts from one image and one video for fair comparisons in this section.

4.2.1. Quantitative Comparisons

We construct 40 test cases with images and videos from the DAVIS [43] dataset and the Internet for evaluation. Both automatic metrics and human evaluations are adopted for assessing the concept composition performance. For automatic metrics, we use *CLIP-T* to measure the alignment between the generated video and the textual prompt with CLIP [45] feature similarities, and choose *DINO-I* to quantify the preservation of visual concepts with the harmonic mean of DINO [10] feature similarities between the composed video and all visual inputs. For human evaluations, we asked 28 volunteers to rate the composed video in the following aspects with a 5-point Likert scale: 1) *Concept Preservation*: how well the composed video preserves the concepts from corresponding visual sources. 2) *Prompt Fidelity*: how well the composed video follows the input prompt. 3) *Motion Quality*: the motion quality of the composed video considering motion smoothness, consistency, naturalness, etc. We compute the average score of the 3 aspects as the *Overall Quality*. Please refer to the supplementary for more details on user study settings.

As displayed in Tab. 1, BiCo consistently outperforms all other methods in both automatic metrics and human evaluations. Compared to the prior art DualReal [55], our method achieves a **+54.67%** improvement on the subjective *Overall Quality*. In addition, BiCo also supports the extraction of non-object concepts, learning multiple concepts from a single input, arbitrary image/video input types, and flexible concept composition via prompt manipulation, where previous methods fall short.

4.2.2. Qualitative Comparisons

We visualize the composed videos in Fig. 5 to provide an intuitive comparison with other methods. It shows a creative motion transfer task, where Textual Inversion [17] and DreamVideo [56] fails to combine the visual concepts. DualReal [55] does not accurately follow the prompt to

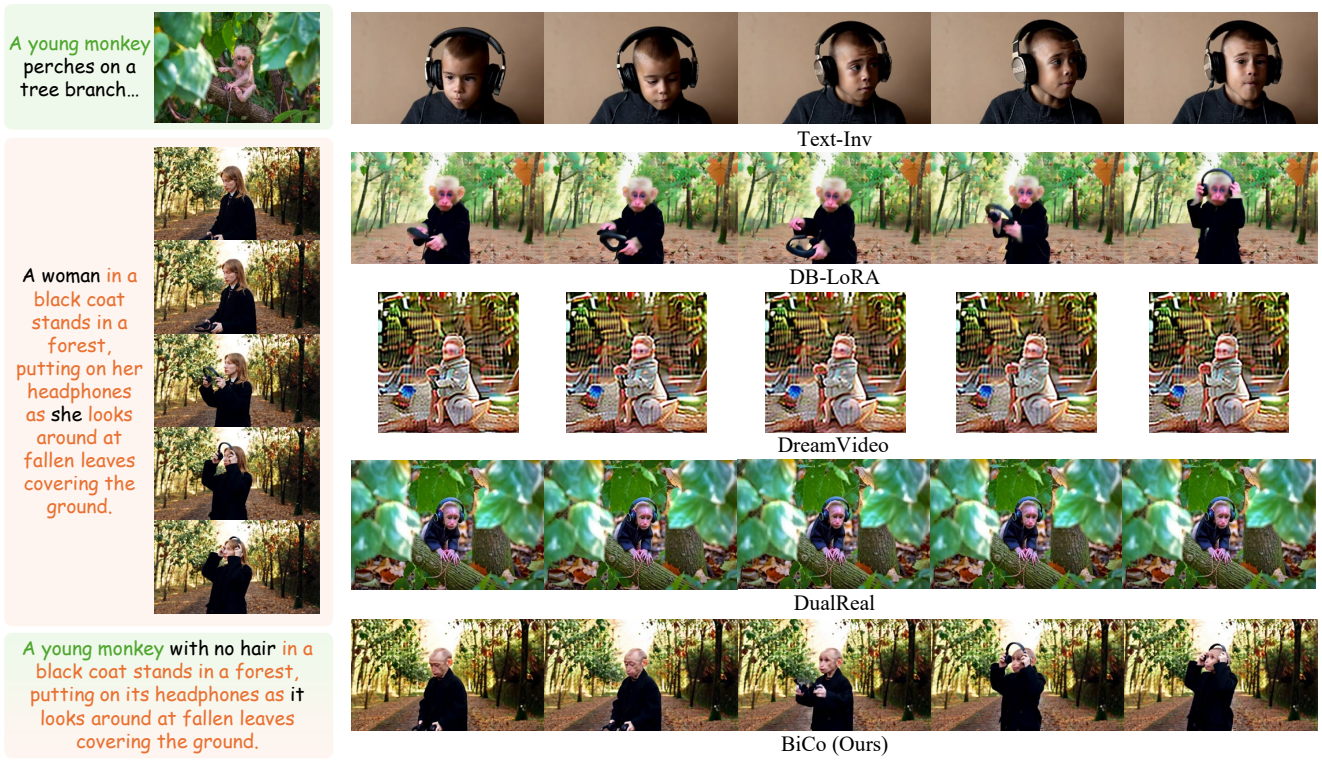

Figure 5. **Qualitative Comparisons with Previous Methods** (§4.2.2). The input visual concepts and composed prompts are on the left.

Table 2. **Ablations of BiCo** (§4.3). Results in **bold** are the best. ▲ stands for without two-stage inverted training strategy.

Hrc.	Div.	Abs.	TDS	Concept↑	Prompt↑	Motion↑	Overall↑
				2.16	2.60	2.26	2.34
✓				2.63	2.88	2.93	2.81
✓	✓			3.40	3.34	3.04	3.26
✓	✓	✓		3.55	3.43	3.43	3.47
✓	✓		✓	3.80	3.97	3.70	3.82
▲	✓	✓	✓	2.60	2.70	2.43	2.58
✓	✓	✓	✓	**4.43**	**4.47**	**4.32**	**4.40**

compose the concepts, and the generated video is almost static. Although DB-LoRA [48] mostly follows the designated prompt to integrate visual concepts, there are significant drifts of visual concepts from the original inputs. BiCo best composes the visual concepts according to the given prompt while maintaining the visual concept consistency with the input image and video.

4.3. Diagnostic Experiments

To provide a better understanding of BiCo's components, we conduct both quantitative ablations and a case study.

Quantitative Ablations. We adopt the human evaluation method in §4.2.1 with another 24 volunteers and the same test cases. The results are presented in Tab. 2.

Case Study. We further illustrate the functions of BiCo's components with a concrete visual concept composition sample with an image and a video input in Fig. 6.

Baseline. We start from a simple baseline with only the global binder, omitting the hierarchical design, DAM, and TDS (#1). This naive baseline method does not achieve satisfactory performance due to limited concept binding capability and image-video compatibility.

Hierarchical Binder Structure. By integrating the hierarchical design of binders (Hrc., #2), the binding capability of our method is significantly enhanced with per-block binders for tailored concept-token association. The effectiveness is demonstrated by the improvement of *Concept Preservation* and *Motion Quality* in Tab. 2 and the better reproduction of the bird concept in Fig. 6 compared to #1.

Prompt Diversification. The prompt diversification operation (Div., #3) enhances the binding accuracy between concepts and prompt tokens under the one-shot training setting of BiCo. As Tab. 2 shows, the *Concept Preservation* score rises significantly compared to #1 with the integration of the prompt diversification operation. However, some unwanted details appear in Fig. 6, degrading the composition quality.

Absorbent Token. The absorbent token (Abs.) in AAM facilitates more accurate concept-prompt binding by suppressing prompt-irrelevant details during training, resulting in reduced unwanted elements comparing #4 to #3 and #7 to #5 in Fig. 6. The improvement of *Concept Preservation* and *Motion Quality* in Tab. 2 further verifies the effectiveness of

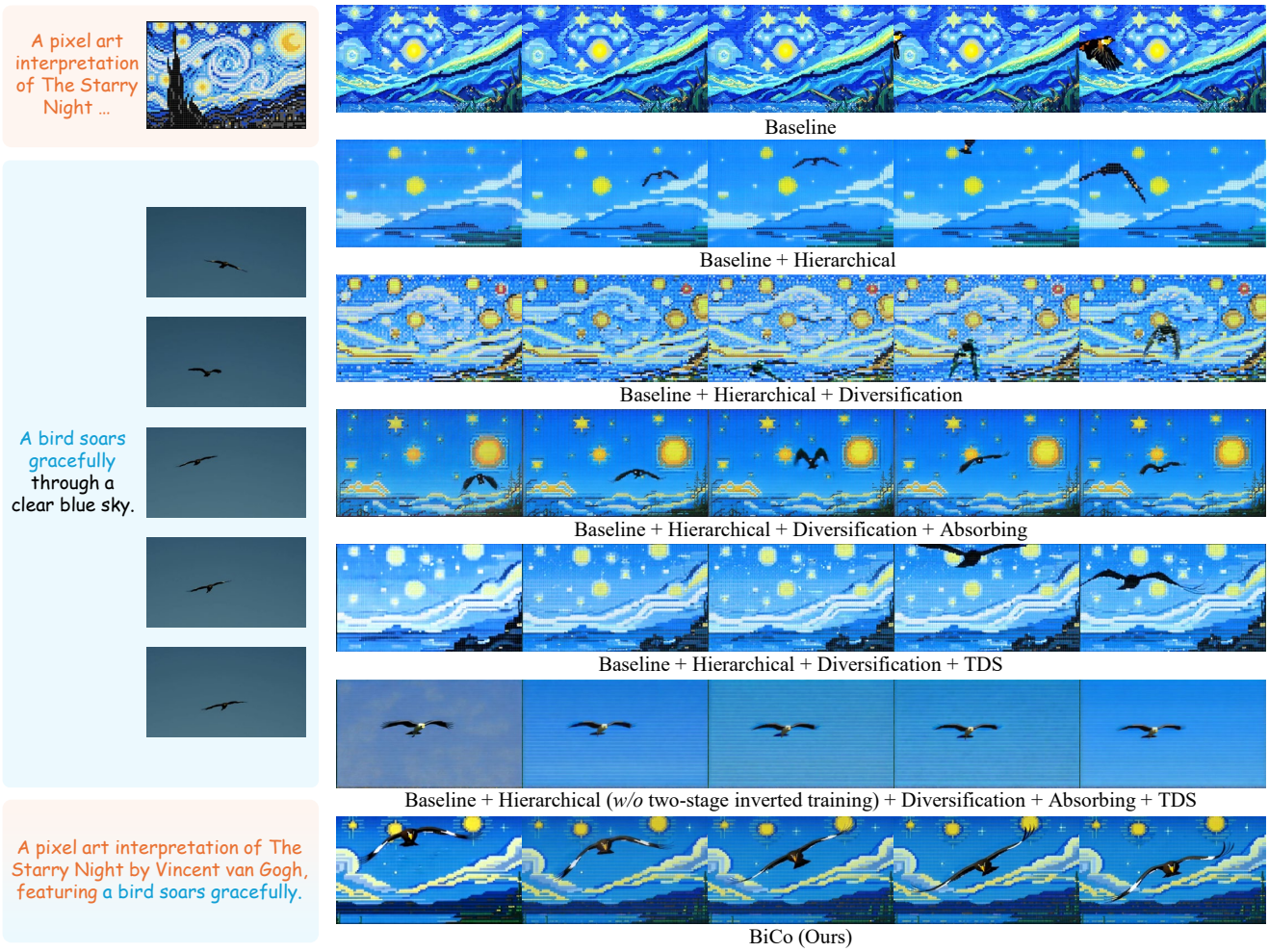

Figure 6. **Case Study for Components** (§4.3). The input visual concepts and composed prompts are on the left.

the absorbent token.

TDS. By decoupling the training process of spatial and temporal concepts in videos, TDS prominently enhances the *Overall Quality* in Tab. 2 comparing #5 to #3 and #7 to #4. The qualitative results in Fig. 6 also improves with better concept detail preservation from both the input image and video. These results validate its effectiveness for improving the compatibility between image and video concepts.

Two-stage Inverted Training Strategy. The two-stage inverted training strategy plays an essential part in training the hierarchical binders. By first training the global binder with a focus on high noise levels, the strategy provides a better initialization for the full training stage and stabilizes the training process. Without such a training strategy, the optimization becomes hard and unstable, resulting in considerably degraded results in #6 of both Tab. 2 and Fig. 6.

4.4. Qualitative Results

We present various creative visual concept composition results with BiCo in Figs. 1 and 7, including the composition of non-object concepts (*e.g.* style and motion), and composing multiple visual concepts. As observed, BiCo consistently achieves satisfactory concept consistency, prompt fidelity, and motion quality, validating the superiority of our design. More results can be found in the supplementary.

4.5. Other Applications

Thanks to the powerful concept binding capability and flexible token manipulation pattern of BiCo, we can utilize BiCo to perform other creative applications for visual content creation. As the upper part of Fig. 8 illustrates, BiCo possesses the capability of decoupling complex concepts from the visual inputs, such as all the dogs from the input with multiple dogs and cats. This is achieved by keeping only the dog-related tokens in the designated prompt and discarding the cat-related ones when generating the target video with the trained binder. In addition, BiCo can also perform text-guided visual editing, as displayed in the lower part of Fig. 8. To edit the input image or video, we first perform concept binding and then compose the designated

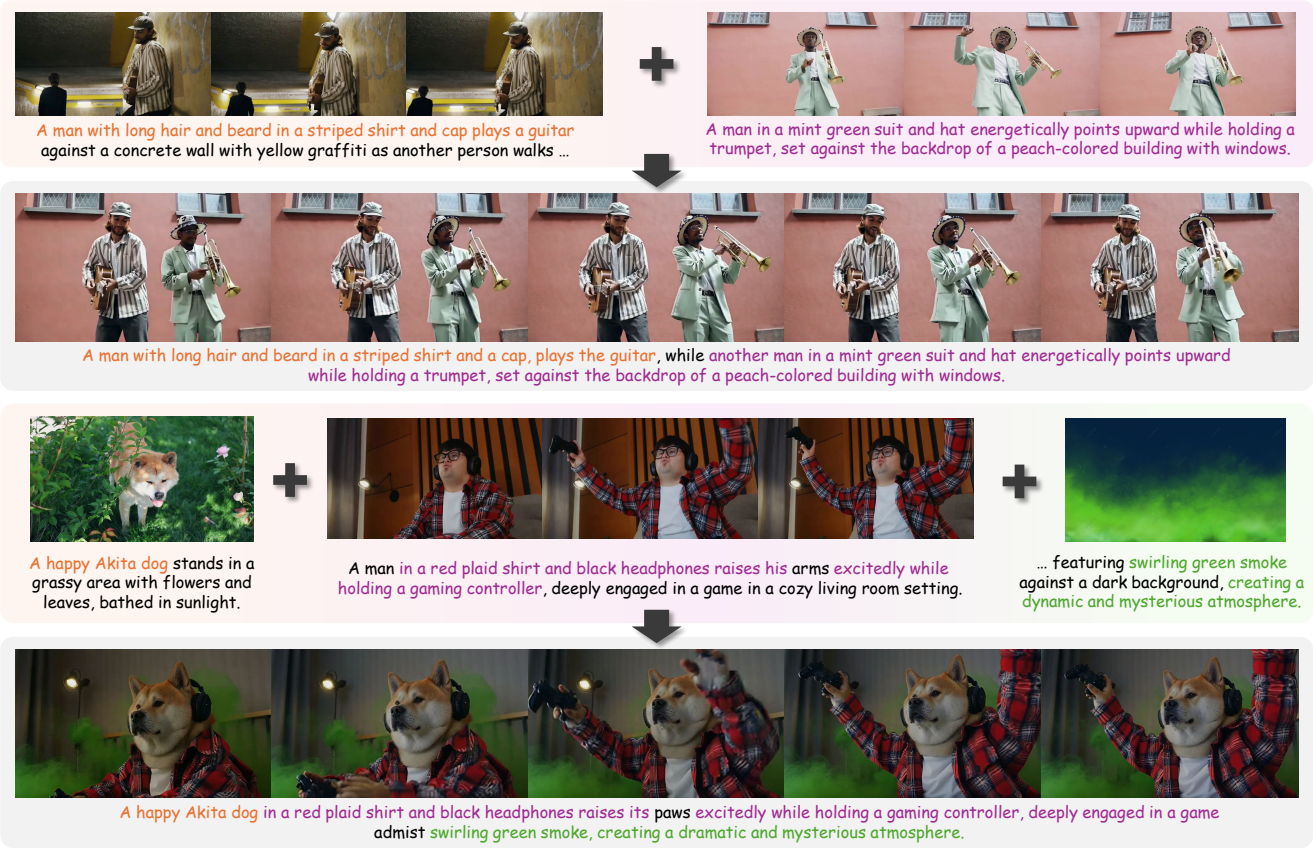

Figure 7. **Qualitative Results** (§4.4). In each case, the upper row shows the visual inputs, and the lower row presents the composed video.

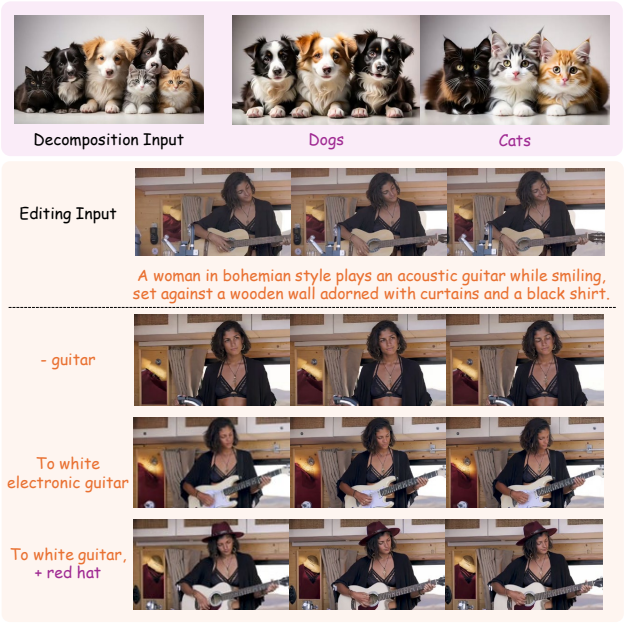

Figure 8. **Other Applications** (§4.5). BiCo can also perform other tasks like image/video decomposition and text-guided editing.

prompt tokens. For the unchanged visual elements, we pass the corresponding prompt tokens through the binder. For the edited parts, the prompt tokens are directly used to compose the designated prompt without updates.

5. Conclusion and Discussion

In this work, we propose BiCo, a one-shot method that can accurately extract complex visual concepts and flexibly combine concepts from both images and videos. It first binds visual concepts with the corresponding prompt tokens and then composes the target prompt with bound tokens from various sources to generate the composed video. It includes a hierarchical binder structure to achieve complex visual concept decomposition, DAM for more accurate concept-token binding, and TDS for enhanced image-video compatibility. Extensive results across various scenarios have validated the effectiveness of BiCo. We believe that BiCo will boost the community's creativity by providing a handy tool to achieve versatile visual concept composition.

Limitations. BiCo treats each token equally in the concept composition process. Nevertheless, the significance of each token for T2V generation is unevenly distributed. Some tokens that represent subjects and motions play a more important role than the function words. We plan to integrate adaptive designs to highlight those critical tokens in future work. More discussions are included in the supplementary.

References

[1] Rameen Abdal, Or Patashnik, Ekaterina Deyneka, Hao Chen, Aliaksandr Siarohin, Sergey Tulyakov, Daniel Cohen-Or, and Kfir Aberman. Zero-shot dynamic concept personalization with grid-based lora, 2025. 2

[2] Rameen Abdal, Or Patashnik, Ekaterina Deyneka, Hao Chen, Aliaksandr Siarohin, Sergey Tulyakov, Daniel Cohen-Or, and Kfir Aberman. Zero-shot dynamic concept personalization with grid-based lora, 2025. 3

[3] Rameen Abdal, Or Patashnik, Ivan Skorokhodov, Willi Menapace, Aliaksandr Siarohin, Sergey Tulyakov, Daniel Cohen-Or, and Kfir Aberman. Dynamic concepts personalization from single videos. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, New York, NY, USA, 2025. Association for Computing Machinery. 2

[4] Rameen Abdal, Or Patashnik, Ivan Skorokhodov, Willi Menapace, Aliaksandr Siarohin, Sergey Tulyakov, Daniel Cohen-Or, and Kfir Aberman. Dynamic concepts personalization from single videos. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, New York, NY, USA, 2025. Association for Computing Machinery. 3

[5] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia 2023 Conference Papers*, New York, NY, USA, 2023. Association for Computing Machinery. 3

[6] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016. 5

[7] 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, Jianqiang 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. 4, 12

[8] Xiuli Bi, Jian Lu, Bo Liu, Xiaodong Cun, Yong Zhang, Weisheng Li, and Bin Xiao. Customttt: Motion and appearance customized video generation via test-time training. *AAAI*, 39(2):1871–1879, 2025. 3

[9] Shengqu Cai, Eric Ryan Chan, Yunzhi Zhang, Leonidas Guibas, Jiajun Wu, and Gordon Wetzstein. Diffusion self-distillation for zero-shot customized image generation. In *CVPR*, pages 18434–18443, 2025. 2

[10] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9630–9640, 2021. 5

[11] Bowen Chen, Mengyi Zhao, Haomiao Sun, Li Chen, Xu Wang, Kang Du, and Xinglong Wu. Xverse: Consistent multi-subject control of identity and semantic attributes via dit modulation, 2025. 2

[12] Jin Chen, Xinxiao Wu, Yao Hu, and Jiebo Luo. Spatial-temporal causal inference for partial image-to-video adaptation. *AAAI*, 35(2):1027–1035, 2021. 5

[13] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. Perception prioritized training of diffusion models. In *CVPR*, pages 11462–11471, 2022. 4

[14] Yusuf Dalva, Hidir Yesiltepe, and Pinar Yanardag. Lorashop: Training-free multi-concept image generation and editing with rectified flow transformers, 2025. 2

[15] Sara Dorfman, Dana Cohen-Bar, Rinon Gal, and Daniel Cohen-Or. Ip-composer: Semantic composition of visual concepts. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers*, New York, NY, USA, 2025. Association for Computing Machinery. 2

[16] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*, pages 12606–12633. PMLR, 2024. 2

[17] 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. 5, 13

[18] Daniel Garibi, Shahar Yadin, Roni Paiss, Omer Tov, Shiran Zada, Ariel Ephrat, Tomer Michaeli, Inbar Mosseri, and Tali Dekel. Tokenverse: Versatile multi-concept personalization in token modulation space. *ACM TOG*, 44(4), 2025. 2, 3

[19] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, WUYOU XIAO, Rui Zhao, Shuning Chang, Weijia Wu, Yixiao Ge, Ying Shan, and Mike Zheng Shou. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. In *NeurIPS*, pages 15890–15902. Curran Associates, Inc., 2023. 3

[20] Yoav HaCohen, Nisan Chiprut, Benny Brazowski, Daniel Shalem, Dudu Moshe, Eitan Richardson, Eran Levin, Guy Shiran, Nir Zabari, Ori Gordon, Poriya Panet, Sapir Weissbuch, Victor Kulikov, Yaki Bitterman, Zeev Melumian, and Ofir Bibi. Ltx-video: Realtime video latent diffusion, 2024. 2, 3

[21] Shaozhe Hao, Kai Han, Zhengyao Lv, Shihao Zhao, and Kwan-Yee K. Wong. Conceptexpress: Harnessing diffusion models for single-image unsupervised concept extraction. In *ECCV*, pages 215–233, Cham, 2024. Springer Nature Switzerland. 3

[22] Xuanhua He, Quande Liu, Shengju Qian, Xin Wang, Tao Hu, Ke Cao, Keyu Yan, and Jie Zhang. Id-animator: Zero-shot identity-preserving human video generation, 2024. 3

[23] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus), 2023. 5

[24] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pages 6840–6851. Curran Associates, Inc., 2020. 2

[25] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *ICLR*, 2022. 2, 3

[26] Chi-Pin Huang, Yen-Siang Wu, Hung-Kai Chung, Kai-Po Chang, Fu-En Yang, and Yu-Chiang Frank Wang. Videomage: Multi-subject and motion customization of text-to-video diffusion models. In *CVPR*, pages 17603–17612, 2025. 2, 3

[27] Lianghua Huang, Wei Wang, Zhi-Fan Wu, Yupeng Shi, Huanzhang Dou, Chen Liang, Yutong Feng, Yu Liu, and Jingren Zhou. In-context lora for diffusion transformers, 2024. 2

[28] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models, 2020. 2

[29] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, Kathrina Wu, Qin Lin, Junkun Yuan, Yanxin Long, Aladdin Wang, Andong Wang, Changlin Li, Duojun Huang, Fang Yang, Hao Tan, Hongmei Wang, Jacob Song, Jiawang Bai, Jianbing Wu, Jinbao Xue, Joey Wang, Kai Wang, Mengyang Liu, Pengyu Li, Shuai Li, Weiyan Wang, Wenqing Yu, Xinchi Deng, Yang Li, Yi Chen, Yutao Cui, Yuanbo Peng, Zhentao Yu, Zhiyu He, Zhiyong Xu, Zixiang Zhou, Zunnan Xu, Yangyu Tao, Qinglin Lu, Songtao Liu, Dax Zhou, Hongfa Wang, Yong Yang, Di Wang, Yuhong Liu, Jie Jiang, and Caesar Zhong. Hunyuanvideo: A systematic framework for large video generative models, 2025. 2, 3

[30] Xianghao Kong, Hansheng Chen, Yuwei Guo, Lvmin Zhang, Gordon Wetzstein, Maneesh Agrawala, and Anyi Rao. Taming flow-based i2v models for creative video editing. *arXiv preprint arXiv:2509.21917*, 2025.

[31] Xianghao Kong, Qiaosong Qi, Yuanbin Wang, Anyi Rao, Biaolong Chen, Aixi Zhang, Si Liu, and Hao Jiang. Profashion: Prototype-guided fashion video generation with multiple reference images. *arXiv preprint arXiv:2505.06537*, 2025. 2

[32] Zhe Kong, Yong Zhang, Tianyu Yang, Tao Wang, Kaihao Zhang, Bizhu Wu, Guanying Chen, Wei Liu, and Wenhan Luo. Omg: Occlusion-friendly personalized multi-concept generation in diffusion models. In *ECCV*, pages 253–270, Cham, 2024. Springer Nature Switzerland. 2, 3

[33] Gihyun Kwon and Jong Chul Ye. Tweediemix: Improving multi-concept fusion for diffusion-based image/video generation. In *ICLR*, 2025.

[34] Gihyun Kwon, Simon Jenni, Dingzeyu Li, Joon-Young Lee, Jong Chul Ye, and Fabian Caba Heilbron. Concept weaver: Enabling multi-concept fusion in text-to-image models. In *CVPR*, pages 8880–8889, 2024. 2, 3

[35] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. 2

[36] Lijiang Li, Huixia Li, Xiawu Zheng, Jie Wu, Xuefeng Xiao, Rui Wang, Min Zheng, Xin Pan, Fei Chao, and Rongrong Ji. Autodiffusion: Training-free optimization of time steps and architectures for automated diffusion model acceleration. In *ICCV*, pages 7082–7091, 2023. 4

[37] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *ICLR*, 2023. 2

[38] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow, 2022. 2

[39] Joanna Materzyńska, Josef Sivic, Eli Shechtman, Antonio Torralba, Richard Zhang, and Bryan Russell. Newmove: Customizing text-to-video models with novel motions. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, pages 1634–1651, 2024. 3

[40] Yuxi Mi, Zhizhou Zhong, Yuge Huang, Qiuyang Yuan, Xuan Zhao, Jianqing Xu, Shouhong Ding, Shaoming Wang, Rizen Guo, and Shuigeng Zhou. Data synthesis with diverse styles for face recognition via 3dmm-guided diffusion. In *CVPR*, pages 21203–21214, 2025. 2

[41] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4172–4182, 2023. 2, 3

[42] Xiangyu Peng, Zangwei Zheng, Chenhui Shen, Tom Young, Xinying Guo, Binluo Wang, Hang Xu, Hongxin Liu, Mingyan Jiang, Wenjun Li, Yuhui Wang, Anbang Ye, Gang Ren, Qianran Ma, Wanying Liang, Xiang Lian, Xiwen Wu, Yuting Zhong, Zhuangyan Li, Chaoyu Gong, Guojun Lei, Leijun Cheng, Limin Zhang, Minghao Li, Ruijie Zhang, Silan Hu, Shijie Huang, Xiaokang Wang, Yuanheng Zhao, Yuqi Wang, Ziang Wei, and Yang You. Open-sora 2.0: Training a commercial-level video generation model in $200k, 2025. 2, 3

[43] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *CVPR*, pages 724–732, 2016. 5

[44] Ryan Po, Guandao Yang, Kfir Aberman, and Gordon Wetzstein. Orthogonal adaptation for modular customization of diffusion models. In *CVPR*, pages 7964–7973, 2024. 3

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 5

[46] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10674–10685, 2022. 2

[47] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. 2

[48] 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. 2, 5, 6, 13

[49] Viraj Shah, Nataniel Ruiz, Forrester Cole, Erika Lu, Svetlana Lazebnik, Yuanzhen Li, and Varun Jampani. Ziplora: Any subject in any style by effectively merging loras. In *ECCV*, pages 422–438, Cham, 2024. Springer Nature Switzerland. 3

[50] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 2

[51] Meituan LongCat Team, Xunliang Cai, Qilong Huang, Zhuoliang Kang, Hongyu Li, Shijun Liang, Liya Ma, Siyu Ren, Xiaoming Wei, Rixu Xie, and Tong Zhang. Longcat-video technical report, 2025. 3

[52] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*. Curran Associates, Inc., 2017. 2

[53] Yael Vinker, Andrey Voynov, Daniel Cohen-Or, and Ariel Shamir. Concept decomposition for visual exploration and inspiration. *ACM Trans. Graph.*, 42(6), 2023. 3

[54] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models, 2025. 2, 3, 5

[55] Wenchuan Wang, Mengqi Huang, Yijing Tu, and Zhendong Mao. Dualreal: Adaptive joint training for lossless identity-motion fusion in video customization, 2025. 2, 3, 5, 13

[56] 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. 2, 3, 5, 13

[57] Felix Wimbauer, Bichen Wu, Edgar Schoenfeld, Xiaoliang Dai, Ji Hou, Zijian He, Artsiom Sanakoyeu, Peizhao Zhang, Sam Tsai, Jonas Kohler, Christian Rupprecht, Daniel Cremers, Peter Vajda, and Jialiang Wang. Cache me if you can: Accelerating diffusion models through block caching. In *CVPR*, pages 6211–6220, 2024. 4

[58] Jiaqi Xu, Xinyi Zou, Kunzhe Huang, Yunkuo Chen, Bo Liu, MengLi Cheng, Xing Shi, and Jun Huang. Easyanimate: A high-performance long video generation method based on transformer architecture, 2024. 2, 3

[59] Danni Yang, Ruohan Dong, Jiayi Ji, Yiwei Ma, Haowei Wang, Xiaoshuai Sun, and Rongrong Ji. Exploring phrase-level grounding with text-to-image diffusion model. In *ECCV*, pages 161–180, Cham, 2024. Springer Nature Switzerland. 2

[60] Yang Yang, Wen Wang, Liang Peng, Chaotian Song, Yao Chen, Hengjia Li, Xiaolong Yang, Qinglin Lu, Deng Cai, Boxi Wu, and Wei Liu. Lora-composer: Leveraging low-rank adaptation for multi-concept customization in training-free diffusion models, 2024. 3

[61] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. In *ICLR*, 2025. 2, 3

[62] Ruihan Zhang, Borou Yu, Jiajian Min, Yetong Xin, Zheng Wei, Juncheng Nemo Shi, Mingzhen Huang, Xianghao Kong, Nix Liu Xin, Shanshan Jiang, Praagya Bahuguna, Mark Chan, Khushi Hora, Lijian Yang, Yongqi Liang, Runhe Bian, Yunlei Liu, Isabela Campillo Valencia, Patricia Morales Tredinick, Ilia Kozlov, Sijia Jiang, Peiwen Huang, Na Chen, Xuanxuan Liu, and Anyi Rao. Generative ai for film creation: A survey of recent advances. In *CVPRW*, pages 6266–6278, 2025. 2

[63] Yue Zhang, Zhizhou Zhong, Minhao Liu, Zhaokang Chen, Bin Wu, Yubin Zeng, Chao Zhan, Yingjie He, Junxin Huang, and Wenjiang Zhou. Musetalk: Real-time high-fidelity video dubbing via spatio-temporal sampling, 2025. 2

[64] Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jia-Wei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. Motiondirector: Motion customization of text-to-video diffusion models. In *ECCV*, pages 273–290, Cham, 2024. Springer Nature Switzerland. 3

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 further ablates the two-stage inverted training strategy.
- §D illustrates more qualitative comparisons with previous methods.
- §E performs another case study to facilitate the understanding of different components of BiCo.
- §F further discusses the limitations with failure cases and the societal impacts of BiCo.

Please refer to the supplemental page for video results.

A. Detailed Prompts for DAM

In the prompt diversification process, we utilize a powerful VLM Qwen2.5-VL [7] 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 3. **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 Ablations on Two-stage Inverted Training Strategy

We provide additional quantitative ablation results under the same settings in §4.3 to facilitate the understanding of the two-stage inverted training strategy. Results are shown in Tab. 3, 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. Additional Qualitative Comparisons

We provide more composed videos in Fig. 9 for additional qualitative comparisons with other methods. Fig. 9a demonstrates a motion transfer task, where Textual Inversion [17] and DreamVideo [56] fails to combine the visual concepts. DualReal [55] suffers from inadequate visual concept preservation and unintended concept leakage (*e.g.*, the green leaves). Although DB-LoRA [48] 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. 9b illustrates a creative style transfer task to integrate the line art sketch style with the subject in a video. All previous methods [17, 48, 55, 56] fail in this task to learn and compose the style concept. This sample further verifies the flexible versatile controllability of BiCo.

E. Extra Case Study

We further illustrate the functions of BiCo's components with another concrete visual concept composition sample in Fig. 10. 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 ac-

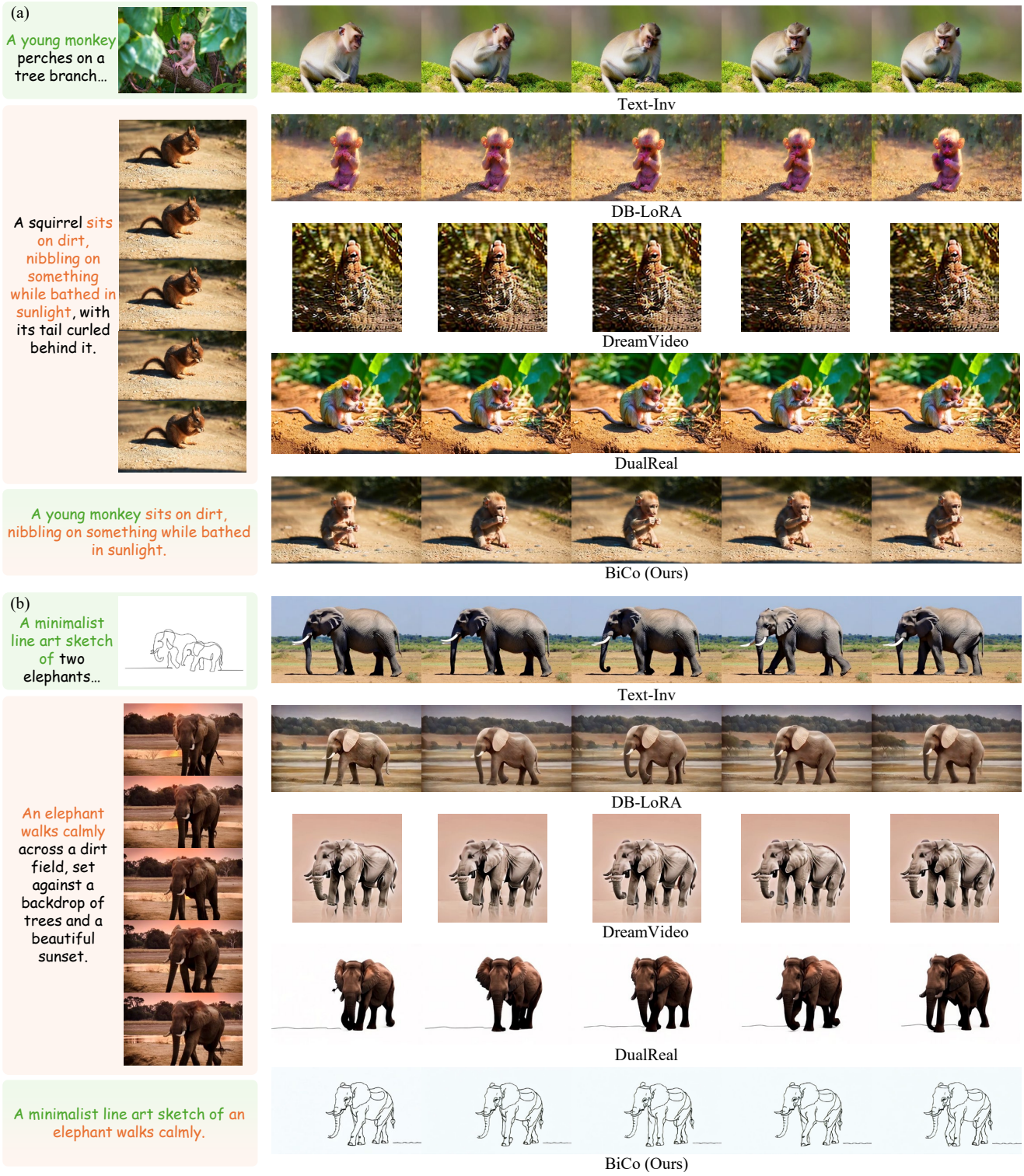

Figure 9. **Additional Qualitative Comparisons** (§D). The input visual concepts and composed prompts are on the left.

curacy of concept-prompt binding, better preserving background details in the composed videos. The effectiveness

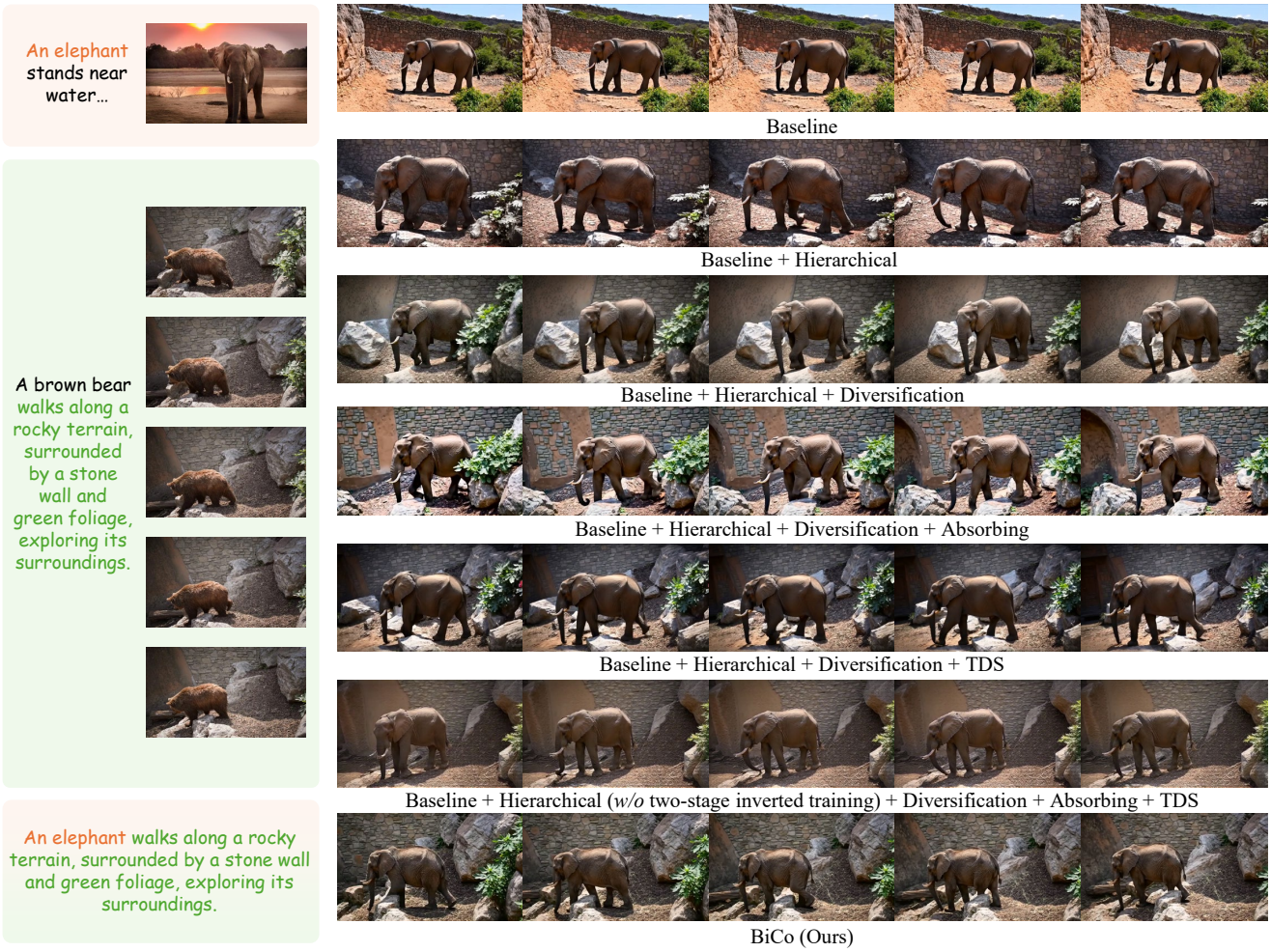

Figure 10. **Extra Case Study for Components** (§E). The input visual concepts and composed prompts are on the left.

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 supplemental page.

F. More Discussions

F.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. 11, 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. 11 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.

Figure 11. **Failure Cases** (§F.1). In each case, the upper row shows the visual inputs, and the lower row presents the composed video.

F.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.