ConceptWeaver: Weaving Disentangled Concepts with Flow

Jintao Chen[1,2*], Aiming Hao[2*], Xiaoqing Chen[2], Chengyu Bai[1], Chubin Chen[2], Yanxun Li[2], Jiahong Wu[2‡], Xiangxiang Chu[2], and Shanghang Zhang[1‡]

[1] State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University, China
[2] AMAP, Alibaba Group, China
cjt@stu.pku.edu.cn, shanghang@pku.edu.cn

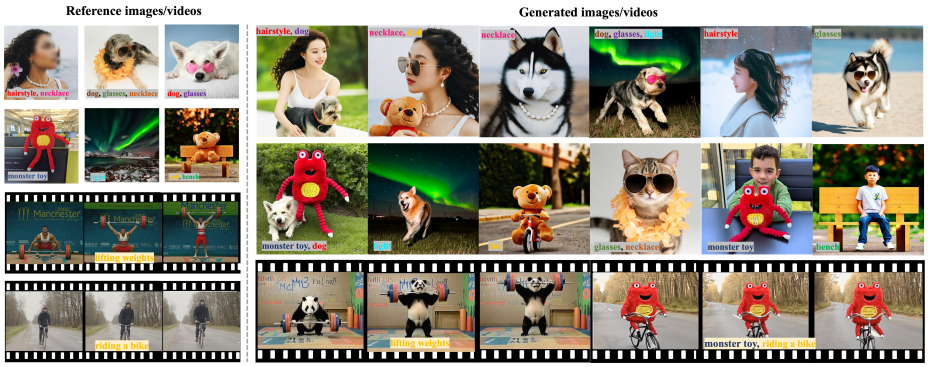

Fig. 1: Compositional Synthesis with ConceptWeaver. Our framework learns a visual concept, like a shirt pattern, from a single reference image by optimizing a personalized semantic offset. This offset is then injected the concept into new scenes by our stage-aware CWG, enabling compositional synthesis across diverse contexts.

Abstract. Pre-trained flow-based models excel at synthesizing complex scenes, yet lack a direct mechanism for disentangling and customizing their underlying concepts from one-shot real-world sources. To demystify this process, we first introduce a novel differential probing technique to isolate and analyze the influence of individual concept tokens on the velocity field over time. This investigation yields a critical insight: the generative process is not monolithic but unfolds in three distinct stages. An initial **Blueprint Stage** establishes low-frequency structure, followed by a pivotal **Instantiation Stage** where content concepts emerge with peak intensity and become naturally disentangled, creating an optimal window for manipulation. A final concept-insensitive **Refinement Stage** then synthesizes fine-grained details. Guided by this discovery, we propose **ConceptWeaver**, a framework for one-shot concept disentanglement. ConceptWeaver learns concept-specific semantic offsets from a single reference image using a stage-aware optimization strategy that aligns with the three-stage framework. These learned offsets are then deployed during inference via our novel ConceptWeaver Guidance (CWG) mechanism,

[1] * Equal contribution; † Project Lead; ‡ Co-corresponding author.

which strategically injects them at the appropriate generative stage. Extensive experiments validate that ConceptWeaver enables high-fidelity, compositional synthesis and editing, demonstrating that understanding and leveraging the intrinsic, staged nature of flow models is key to unlocking precise, multi-granularity content manipulation. We will release our code at `https://github.com/JasperChennn/ConceptWeaver`.

Keywords: One-shot Concept Disentanglement · Stage-aware Optimization · ConceptWeaver Guidance

1 Introduction

Modern generative models [5, 11, 24, 44, 49], functioning like digital looms, can weave together diverse elements—objects, attributes, and motions—into highly realistic visual tapestries from text prompts. This remarkable capability implies an underlying compositional structure in their internal representations. However, moving beyond simply generating what is in a prompt to precisely controlling how each concept appears remains a central challenge. Achieving true personalized content creation requires the ability to disentangle these visual building blocks, extracting them from reference images and seamlessly recombining them in novel compositions. This is the key to unlocking the next level of creative and controllable synthesis [2, 42, 52].

Early personalization techniques, such as textual inversion [7] or parameter fine-tuning [16, 20, 27, 39], achieve single-concept embedding but often face a fundamental trade-off between subject fidelity and prompt controllability. More recent works [3, 8, 17, 36, 53] have sought to disentangle multiple concepts from a single image. A significant breakthrough in this area is TokenVerse [8], which leverages a specific architectural feature for control: the per-token modulation space—a pathway in certain Diffusion Transformer (DiT) [34] variants (e.g., Flux [24]) where text embeddings directly modulate channels. By learning offsets in this space, TokenVerse enables impressive, mask-free multi-concept personalization.

Despite its strengths, TokenVerse has two inherent **limitations**. First, it depends on a text-conditioned modulation path, an architectural feature absent from most DiTs [34], where modulation typically encodes only non-semantic signals such as the timestep t. This reliance constrains its applicability to a narrow subset of models. Second, by focusing solely on a spatial locus of control (i.e., which parameters to modify), it neglects the temporal dimension of concept formation—failing to address when a concept should be manipulated to maximize effect and minimize interference. This dual gap in architectural generality and temporal awareness underpins the motivation for our work.

Building on these observations, our work addresses both the temporal challenge and the generality gap identified above. We begin with a novel differential probing technique to isolate and chart the influence of individual concepts on the velocity field [28, 31] over time. This analysis yields a critical and universal insight: *the generative process is not monolithic but unfolds in a distinct three-stage framework*. An initial **Blueprint Stage** establishes coarse structure, followed by

a pivotal **Instantiation Stage** where content concepts emerge with peak intensity and become naturally disentangled. A final, concept-insensitive **Refinement Stage** then adds fine-grained detail. This discovery reveals a principled, optimal window for manipulation. Unlike recent flow-editing works that decompose editing trajectories or prompts to reduce entanglement [50, 54], our analysis focuses on when reference concepts emerge in the velocity field and how to inject them at the corresponding generative stage.

Guided by this temporal insight, we introduce **ConceptWeaver**, a one-shot learning and guidance framework that operationalizes the probing results without requiring access to a modulation space. ConceptWeaver learns concept-specific semantic offsets from a single reference data using a stage-aware optimization strategy that concentrates training exclusively on each concept's peak-influence stage. During inference, our ConceptWeaver Guidance (CWG) mechanism injects these learned offsets only during the corresponding stage, aligning control with the model's intrinsic dynamics rather than imposing uniform, heuristic constraints. By combining architecture-agnostic applicability with precise temporal alignment, ConceptWeaver enables high-fidelity compositional synthesis..

In summary, our contributions are as follows:

- A novel differential probing technique that reveals, for the first time, a universal three-stage framework for concept formation in flow-based models, identifying a principled window for manipulation.
- ConceptWeaver, a one-shot framework featuring a stage-aware optimization strategy and a novel CWG mechanism, designed to leverage this temporal discovery for principled concept control.
- Extensive experiments demonstrating superior performance in high-fidelity, compositional synthesis, proving the efficacy of our approach.

2 Related Works

Flow Methods. Diffusion models [14, 33, 40, 41] have revolutionized generative tasks with high-quality outputs, but their inverse processes rely on Markov chains leading to slow sampling, and score-based formulations require expensive gradient estimations for high-dimensional data. To address these limitations, Flow Matching (FM) was proposed as a lightweight alternative framework [9, 24, 28, 29, 31]. It bypasses explicit noise addition steps and directly learns a continuous flow mapping noise distributions to data distributions—optimizing a velocity field that guides smooth transitions from random noise to real data. Unlike diffusion models, FM enables efficient sampling via ODE solvers (faster than Markov chains) and avoids gradient-based score estimation. Subsequent works [6, 25, 48] extended FM to conditional tasks: FlowEdit [22] leverages conditional Flow Matching to enable precise image editing by aligning source and target velocity fields. It uses the difference between source and target velocity fields to guide inverse ODE sampling [32], solving the control issues in diffusion-based editing (e.g., drift from desired edits).

Personalization Methods. Personalized generation [7, 10, 16, 20, 27, 30, 39] aims to synthesize images that faithfully incorporate a user-specified concept (e.g., a person, object or style) while adhering to a given textual description [19,45,47,52]. This task typically relies on a small set of reference images to generalize the target concept into new contexts. Existing personalization methods are generally categorized by three inversion spaces. In the *textual space*, the concept is represented as a learnable pseudo-word embedding (e.g., Textual Inversion [7, 23]). The *parameter space* methods fine-tune the diffusion model itself to internalize the concept, with DreamBooth [39] being a prime example, often augmented by parameter-efficient techniques like LoRA [16]. Despite these advancements, a core challenge remains: balancing subject fidelity with text controllability. Many methods that excel at capturing fine-grained visual details often overfit, leading to outputs that merely reproduce original images rather than adapting to prompt-specified structural or contextual variations—a limitation known as the "copy-paste phenomena" [4, 30, 37].

Concept Decoupling. Recent personalized generation efforts aim to extract and reuse multiple visual concepts from minimal reference data [3,8,20,53]. However, cleanly separating semantically distinct attributes (e.g., shape, appearance, pose, context) remains challenging. For example, Break-a-Scene [1] uses manually annotated spatial masks for multi-concept learning, which limits scalability and conflates non-spatial attributes not addressable by masks. Alternative approaches like Inspiration Tree [43] and ConceptExpress [12] associate a single image with several learned tokens. Yet, the implicit semantic assignment to individual tokens during optimization often lacks user control, leading to unclear or inconsistent visual facets governed by each token, which hinders reliable composition. BiCo [21] further explores image/video concept composition by binding visual concepts to corresponding prompt token segments with binder modules, improving flexible concept reuse across visual sources. By probing concept emergence in the velocity field, ConceptWeaver learns prompt-conditioned concept offsets and injects them at their most effective generative stage, providing a complementary temporal perspective for concept decoupling and composition.

3 Preliminaries and Analysis

3.1 Conditional Velocity Field

Flow models [24, 28, 31, 44] generation as learning a continuous-time velocity field that transports samples from a simple prior distribution to the target data distribution. Given a data example $\boldsymbol{x}_0 \sim \mathcal{D}$ and its fully noised counterpart $\boldsymbol{x}_1 \sim \mathcal{N}(0, \boldsymbol{I})$, the forward trajectory is defined by:

$$\boldsymbol{x}_t = (1-t)\,\boldsymbol{x}_0 + t\boldsymbol{x}_1,\ t \in [0,1]\,, \tag{1}$$

yielding the velocity:

$$\boldsymbol{v}_t = \frac{\partial \boldsymbol{x}_t}{\partial t} = \boldsymbol{x}_1 - \boldsymbol{x}_0. \tag{2}$$

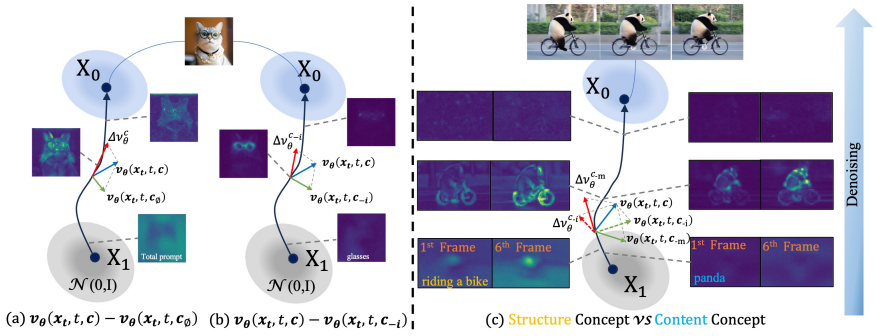

Fig. 2: Probing Concept Formation Dynamics. While standard prompt guidance (a) provides a dense, entangled signal, our differential probing technique (b) isolates the dynamic influence of individual concepts. This analysis uncovers a consistent three-stage framework (c): an early **Blueprint Stage** dominated by structural concepts (e.g.,"riding a bike"), a pivotal mid-stage **Instantiation Stage** where content concepts (e.g., "panda") are decoupled and peak in intensity, and a final **Refinement Stage** where signals fade.

In the text-conditioned setting, let $\boldsymbol{c}$ denote the text prompt. The prompt is processed by a text encoder into a sequence of token embeddings $\{\boldsymbol{e}_1, \cdots, \boldsymbol{e}_K\}$, each corresponding to an individual lexical unit within $\boldsymbol{c}$. These embeddings are injected into the velocity predictor:

$$\hat{\boldsymbol{v}}_t = \boldsymbol{v}_\theta\left(\boldsymbol{x}_t, t, \{\boldsymbol{e}_k\}_{k=1}^{K}\right), \tag{3}$$

directly modulating the instantaneous direction and magnitude of motion in latent space. This conditioning shapes the velocity field $\boldsymbol{v}_\theta(\cdot)$ into a semantic trajectory aligned with the prompt, but also intrinsically *couples* all concepts in $\boldsymbol{c}$ into a single, entangled guidance force.

3.2 Prompt Semantic Guidance

When stronger prompt adherence is desired, one can operate directly in the velocity field to reinforce the semantic trajectory induced by the text. Among such approaches, Classifier-Free Guidance (CFG) [15] is widely adopted. CFG computes a single *global semantic shift* in the velocity field by contrasting the model's velocity predictions guided by the text prompt with those generated from a fully null prompt, and then amplifies this shift during inference:

$$\boldsymbol{v}_{\text{CFG}} = \boldsymbol{v}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c}_\emptyset) + w \cdot (\boldsymbol{v}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c}) - \boldsymbol{v}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c}_\emptyset)), \tag{4}$$

where $w > 1$ controls the guidance scale. The difference term $\Delta\boldsymbol{v}_{\text{prompt}} = \boldsymbol{v}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c}) - \boldsymbol{v}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c}_\emptyset)$ acts as a prompt semantic guidance in velocity field, steering the generative trajectory toward the overall semantics of $\boldsymbol{c}$. By construction, this guidance reflects the combined influence of all concepts described in the prompt, providing holistic guidance without distinguishing their individual contributions. The inability to isolate concept-specific influence motivates the following concept disentanglement task.

3.3 Concept Decoupling

Real-world visual scenes are inherently compositional, blending high-level structural concepts (e.g., spatial layouts) with content concepts (e.g., identities, textures). Pre-trained flow-based generative models, trained on such data, demonstrate a remarkable ability to synthesize novel, coherent scenes by recombining these elements.

This powerful compositional capability strongly suggests that the model's internal representations must go beyond memorizing concept combinations; coherent compositional generalization plausibly requires that constituent concepts be encoded in a partially disentangled form within the underlying velocity field. We therefore hypothesize that *such a disentangled structure exists, enabling the manipulation and recomposition of individual concepts.*

The **Concept Decoupling** task examines this hypothesis in the challenging **one-shot** setting. The practical goal is to achieve fine-grained control over composition. Formally, given N reference images $(\boldsymbol{I}_1, ..., \boldsymbol{I}_N)$, each paired with a textual prompt comprising a multiple concepts set $\mathcal{C}_i$, the task is to synthesize a novel target image $\tilde{\boldsymbol{I}}$ by selectively recombining concepts from these sets. For instance, given $\boldsymbol{I}_1$ describing "cat" and "glasses" and $\boldsymbol{I}_2$ describing "dog" and "a beach," the goal is to generate $\tilde{I}$ depicting "dog" (from $\mathcal{C}_2$) wearing "glasses" (from $\mathcal{C}_1$).

Achieving this level of control, however, requires a deeper understanding of the model's internal dynamics. This reframes the challenge: the central analytical task becomes to characterize the individual contribution of each semantic token to the conditional velocity field. A significant barrier here is the model's inherent non-linearity; the full velocity field likely involves complex interactions, making a simple additive decomposition insufficient [13,26]. This exposes a fundamental question at the heart of generative control: *How can we reliably isolate and quantify the contribution of an individual concept, when its influence is deeply entangled with others in a dynamic, non-linear system?* Answering this question would allow us to chart the temporal lifecycle of concepts—their emergence, peak contribution, and decay—providing the necessary foundation for principled, fine-grained compositional control.

4 Methods

Our method for one-shot concept disentanglement proceeds in three steps. First, we introduce a probing technique to analyze the temporal dynamics of concept formation (Sec. 4.1). Guided by this analysis, we then present **ConceptWeaver**, a framework for learning concept-specific offsets from a single reference (Sec. 4.2). Finally, we detail our stage-aware guidance mechanism that applies these offsets for high-fidelity compositional synthesis (Sec. 4.3).

4.1 Concept-Specific Velocity Field Probing

To investigate how individual concept $\boldsymbol{e}_i$ contributes to the velocity field, we introduce a **differential probing technique** inspired by CFG. Instead of con-

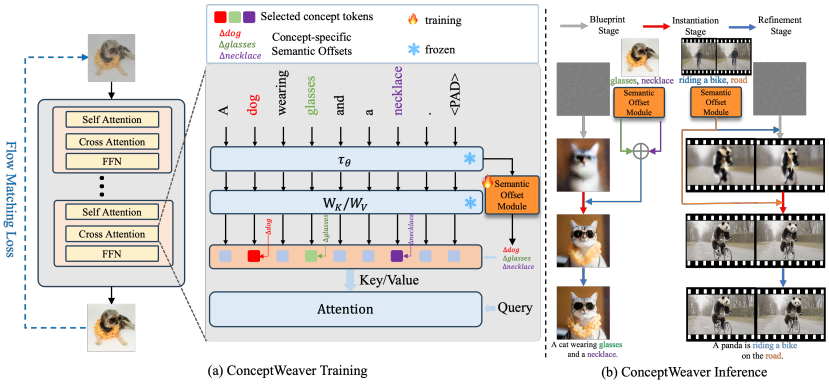

Fig. 3: Flowchart of proposed ConceptWeaver. Our framework consists of two main phases: training and inference. (a) **During training**, a lightweight Semantic Offset Module learns to represent a visual concept by adding a learnable offset to the key/value pairs of its corresponding concept token. This is optimized with a stage-aware loss. (b) **During inference**, our ConceptWeaver Guidance (CWG) applies these offsets in a stage-aware manner: structural offsets are injected during the Blueprint Stage, and content offsets during the Instantiation Stage, enabling precise compositional control.

trasting with a fully null prompt, we compute the difference against a *partially-masked prompt* $\boldsymbol{c}_{-i} = \boldsymbol{c} - \Delta\boldsymbol{c}_{\text{concept}_i}$, where $\Delta\boldsymbol{c}_{\text{concept}_i} = \boldsymbol{e}_i \cdot \boldsymbol{\delta}_i$. (only the i-th concept token $\boldsymbol{e}_i$ is retained and all other components are zero) This yields an estimate of the *concept-specific velocity shift*:

$$\Delta\boldsymbol{v}_{\text{concept}_i} = \boldsymbol{v}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c}) - \boldsymbol{v}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c}_{-i}). \tag{5}$$

which serves as an approximate measure of the dynamic influence of $\boldsymbol{e}_i$ throughout generation, despite the non-linearity of the underlying model.

For our analysis of both image and video generation, we utilize **Wan2.2** [44] as our base Flow Matching model. As shown in Figure 2, such probing reveals temporal variations that are hidden under standard holistic guidance. While the standard shift $\Delta\boldsymbol{v}_{\text{prompt}}$ in (a) provides a dense, entangled signal throughout generation, the concept-specific shift $\Delta\boldsymbol{v}_{\text{glasses}}$ in (b) evolves dynamically: starting as a weak, diffuse signal in early timesteps, peaking sharply and being tightly localized at mid-stage, and then fading into sparse high-frequency noise in later timesteps. Beyond this example, our probing shows a clear distinction between structural and content concepts. Structural concepts (e.g., "riding a bike") exert their strongest influence in the early-to-mid stages, establishing the scene's layout and motion. Content concepts (e.g., "panda" or "glasses") follow a more concentrated lifecycle — weak and diffuse at first, sharply peaking in both intensity and spatial localization at mid-stage, then rapidly diminishing during refinement.

This directly observed, non-monotonic behavior allows us to move beyond a simple "coarse-to-fine" understanding and characterize the generative process through a more structured, three-stage framework:

- **Blueprint Stage** (early t): Concepts exhibit weak, diffuse signals across the image as the model focuses on establishing a coarse structural layout, providing a foundation for subsequent concept refinement.
- **Instantiation Stage** (mid t): The content concept reaches peak intensity and strong spatial localization, marking a critical window where its semantic identity is fully materialized. In this stage, the signal is potent and disentangled from other elements, allowing precise and independent manipulation without affecting unrelated regions.
- **Refinement Stage** (late t): Concept signals gradually diminish while the model concentrates on synthesizing fine-grained textures and enhancing global coherence.

This discovery—that **concepts are not uniformly exerted but rather emerge and disentangle within specific temporal stages**—forms the cornerstone of our work.

4.2 ConceptWeaver: Learning Concept Offsets

The core finding of our probing analysis is that a concept's influence can be isolated by measuring the velocity change after removing its corresponding token. This operation—effectively applying a negative offset to the prompt—motivates a new, unified perspective on concept manipulation: the principle of **concept offsets**. In our probing, the "existence shift" (none $\rightarrow$ generic) is measured by applying a negative semantic offset $\Delta \boldsymbol{c}_{\text{concept}_i}$. We extend this offset-based formulation from *analysis* to *synthesis*: instead of a predefined removal offset, we learn a reference-conditioned offset, $\Delta \boldsymbol{c}_{\text{ref}}$, that induces a *customization shift* (generic $\rightarrow$ specific). To this end, we introduce ConceptWeaver, a one-shot framework to learn a set of *concept-specific semantic offsets* from a single reference. As illustrated in Figure 3 (a), this is achieved with a lightweight, learnable **Semantic Offset Module**. This module learns an additive offset $\Delta \boldsymbol{e}_i$ for each selected concept token $\boldsymbol{e}_i$, which is then injected into the key/value representations of the model's cross-attention layers. These offsets, forming the semantic offsets $\Delta \boldsymbol{c}_{\text{ref}}$, are optimized by minimizing a flow matching loss between the model's prediction and the ground-truth velocity, aiming to collapse the conditional velocity field toward the reference distribution:

$$\mathcal{L}_{flow} = \mathbb{E}\left[\left\|\boldsymbol{v}_\theta\left(\boldsymbol{x}_t, t, \boldsymbol{c} + \Delta \boldsymbol{c}_{\text{ref}}\right) - \boldsymbol{v}_t\right\|^2\right]. \tag{6}$$

Crucially, this optimization is **stage-aware**, directly leveraging our findings from the probing analysis. For content concepts like object identity (e.g., "glasses", "necklace"), the optimization prioritizes loss computations within the pivotal Instantiation Stage. In contrast, for structural concepts like motion or layout (e.g., "riding a bike"), it prioritizes the Blueprint Stage. By aligning the optimization with the model's own internal focus during its stage of peak influence, this targeted approach naturally forces the shift to encode a disentangled representation of the concept, without requiring an explicit disentanglement objective.

Upon convergence, this process yields a set of disentangled concept offsets, each encoding the visual attributes of a concept from the reference. These learned offsets can be deployed for novel synthesis as shown in Figure 3 (b).

4.3 ConceptWeaver Guidance

Building upon standard CFG, we introduce ConceptWeaver Guidance (CWG), a flexible mechanism that extends the velocity field with our learned concept offsets $\{\Delta \boldsymbol{e}_i\}$. CWG refines coarse, prompt-level control into a fine-grained, concept-aware process by decomposing the guidance into a weighted sum of individual concept shifts.

Let N be the number of customized concepts, each with a learned offset $\Delta \boldsymbol{c}_{\text{ref},i} = \Delta \boldsymbol{e}_i \cdot \boldsymbol{\delta}_i$ and a corresponding guidance scale $w_{c,i}$. The full guidance formulation is:

$$\boldsymbol{v}_{\text{CWG}} = \boldsymbol{v}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c}_\emptyset) + w \cdot \Delta \boldsymbol{v}_{\text{prompt}} + \sum_{i=1}^{N} w_{\text{c},i} \cdot \Delta \boldsymbol{v}_{\text{ref},i}, \tag{7}$$

Here, $\Delta \boldsymbol{v}_{\text{prompt}}$ is the standard prompt-level shift, while $\Delta \boldsymbol{v}_{\text{ref},i} = \boldsymbol{v}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c} + \Delta \boldsymbol{c}_{\text{ref},i}) - \boldsymbol{v}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c})$ is the reference-concept-specific shift for the i-th concept. This per-concept weighting scheme provides fine-grained control over the contribution of each concept. By adjusting the guidance scales $\{w_{c,i}\}$ independently, the relative influence of each learned offset can be precisely modulated. This capability is crucial for balancing potentially competing concepts and achieving coherent compositional synthesis, particularly in multi-concept scenarios.

Crucially, guided by our analysis, CWG applies these learned shifts in a time-dependent manner. Structural concept offsets are injected during the **Blueprint Stage** to shape low-frequency layouts, while content concept offsets are introduced during the **Instantiation Stage** to leverage their peak signal intensity. As demonstrated in Figure 4, this stage-aware, multi-level guidance enables customized rendering of target concepts, faithfully preserving and integrating their attributes across diverse subjects and scenes.

5 Expriments

5.1 Experimental Setups

Datasets. For the single-concept dataset, we curated 20 reference images from Dreambench++ [35] and Dreambooth [39], comprising 11 object/content concepts (for appearance personalization) and 9 layout/structural concepts (for spatial control), and additionally collected 10 motion concepts from the internet for motion concept transfer. For the multi-concept dataset, we curated 15 multi-concept references from the internet, with each reference paired with 2–4 constituent concepts drawn from the single-concept pool above so that both appearance and structural compositions are covered. To stress-test compositional generation, we used GPT-5.5 to generate 50 composition prompts spanning diverse combinations of subjects, attributes, motions, and scenes, and sampled 5

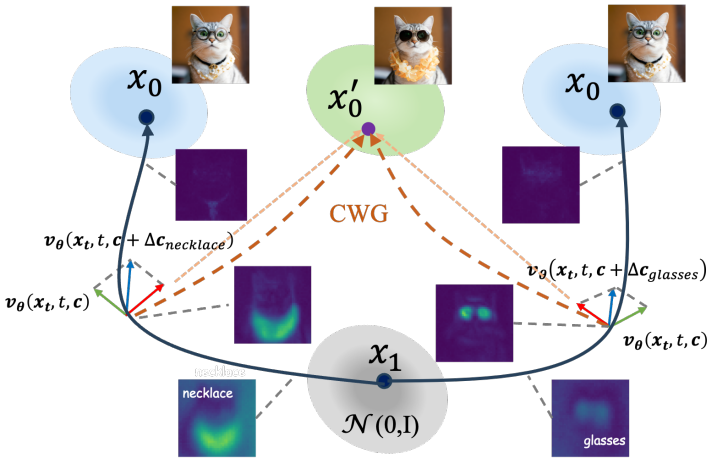

Fig. 4: Mechanism of ConceptWeaver Guidance (CWG). Our guidance modifies the standard generative path (to $\boldsymbol{x}_0$) by injecting learned concept shifts (red vectors) at stage-aware intervals. This timed intervention, visualized by the heatmaps, steers the trajectory from noise ($\boldsymbol{x}_1$) to a customized target ($\boldsymbol{x}'_0$), enabling precise compositional control.

images per prompt, totaling 250 generated images for quantitative evaluation. All our self-collected data were captioned using LLM. We used LLM to create 5 test prompts each for single-concept generation and multi-concept fusion (including motion concept transfer).

Evaluation Metrics. Consistent with prior work, our evaluation comprises Following previous studies, we evaluate our method on the following three aspects:

1. **Textual Alignment:** Assessed by the average CLIP similarity ($S^{\mathrm{T}}_{\mathrm{CLIP}}$) [38] between the generated images and their textual prompts.
2. **Concept Alignment:** Measured by the CLIP ($S^{\mathrm{I}}_{\mathrm{CLIP}}$) and DINO ($S_{\mathrm{DINO}}$) [51] similarities between the generated images and the reference images. For multi-subject generations, we report the mean similarity, calculated across individual subject-reference pairs.
3. **MLLM-based Evaluation:** We employ GPT-4o [18] to score the single-concept generation (S^{4o}_{GPT}) and GPT-5.5 to score the more challenging multi-concept composition ($S^{5.5}_{GPT}$), capitalizing on their advanced visual understanding capabilities. For each setting, the corresponding model is prompted to assign a score from 0 to 4 for every generated instance, based on the provided concept name(s), reference image(s), and the generated image(s). Higher scores indicate better alignment.

Baselines Previous methods mostly focused on only one aspect. For concept decoupling and multi-concept generation in images, we choose MS-Diffusion [46] and reproduced TokenVerse [8] (as TokenVerse*) and DreamBooth [39] on Wan2.2-TI2V-5B. For concept video generation, we chose VACE-14B [19] as a comparative method. See the Appendix **A** for specific settings.

Implementation Details. In ConceptWeaver, we employ Wan2.2-TI2V-5B as our base architecture. For the images, we maintain the training resolution at 704×704, inference resolution at 704×704 for small number of concepts and 704×1280 for multi concepts generation. We perform data augmentation methods such as random cropping and mirror flipping. For the concept videos generation, we maintained a training resolution of 704×1280, 49 frames per second (fps) at 16. All case prompts can be expanded and enhanced using templates. The semantic offset module's rank was set to 16, and rank alpha to 32. The training learning rate was $1e-4$, and each case was trained for 2000 steps.

Table 1: Quantitative comparison on single- and multi-concept generation. The metrics evaluate textual alignment ($S^{\mathrm{T}}_{\mathrm{CLIP}}$) and concept alignment ($S^{\mathrm{I}}_{\mathrm{CLIP}}$, S_{DINO}, S_{GPT}). ↑ indicates higher is better. Best results are highlighted in **bold**.

Method	Single-concept generation				Multi-concept generation			
	$S^{\mathrm{T}}_{\mathrm{CLIP}}\uparrow$	$S^{\mathrm{I}}_{\mathrm{CLIP}}\uparrow$	$S_{\mathrm{DINO}}\uparrow$	$S^{4o}_{GPT}\uparrow$	$S^{\mathrm{T}}_{\mathrm{CLIP}}\uparrow$	$S^{\mathrm{I}}_{\mathrm{CLIP}}\uparrow$	$S_{\mathrm{DINO}}\uparrow$	$S^{5.5}_{GPT}\uparrow$
TokenVerse*	35.20	61.75	26.31	3.20	33.06	54.89	26.33	2.47
MS-Diffusion	35.02	75.58	**70.69**	3.60	33.98	68.90	66.59	3.08
DreamBooth	34.50	73.14	68.96	3.70	33.42	66.80	65.79	2.98
Ours	**35.22**	**76.13**	69.57	**3.80**	**35.36**	**77.06**	**69.02**	**4.14**

5.2 Main Results

Our proposed method demonstrates outstanding performance across multiple generation tasks. For **single-concept generation**, as shown in Table 1, our method achieves state-of-the-art results, leading in both concept alignment ($S^{\mathrm{I}}_{\mathrm{CLIP}}$, S_{DINO}, S^{4o}_{GPT}) and text alignment ($S^{\mathrm{T}}_{\mathrm{CLIP}}$), which validates its strong fundamental generation capabilities. For **multi-concept composition**, Table 1 shows that ConceptWeaver consistently surpasses all baselines across concept alignment and text alignment metrics, with a particularly large gain on $S^{5.5}_{GPT}$. Consistently, Figure 5 shows that our method not only effectively integrates multiple concepts but also excels at avoiding overfitting to the reference images. Furthermore, in the **Reference To Video (R2V) task** (Figure 6), ConceptWeaver successfully disentangles target concepts and transfers them to new contexts. More comparison results are shown in Appendix **B**.

5.3 User Study

We conducted a user study with 10 participants using a 5-point Likert scale (1–5) to evaluate **Concept Consistency**, **Text Adherence**, and **Aesthetic Quality**. As shown in Table 2, ConceptWeaver achieves the highest average scores across all metrics. Notably, it scores **4.06** in Concept Consistency, significantly outperforming MS Diffusion (3.75) and Dreambooth (3.02). Similar improvements are observed in Text Adherence (**3.78**) and Aesthetic Quality (**4.12**), confirming ConceptWeaver's superior ability to balance concept fidelity with prompt controllability and visual appeal.

Fig. 5: Comparison on multi-concept composition generation. The figure illustrates different methods' performance in combining multiple concepts, focusing on overfitting to original reference concepts and the ability to disentangle target concepts.

Fig. 6: Comparison of R2V task results. Contrasts VACE method with our approach in terms of concept disentanglement, structural preservation, and "copy-paste phenomena".

5.4 Ablation Study

We ablate our stage-aware mechanisms and the concept guidance scale $w_{c,i}$.

Effect of Stage-Aware Components. We compare our full model against two variants: (1) **w/o Stage-aware Opt.**, where the concept offset is trained uniformly over all timesteps; and (2) **w/o Stage-aware Guid.**, where the guidance is applied uniformly across all stages. Figure 7 shows that removing Opt. degrades concept fidelity, while removing Guid. introduces structural artifacts and concept bleeding. Table 3 confirms this quantitatively: $S^{5.5}_{GPT}$ drops from 4.14 to 3.61 and 3.34 respectively. The two designs are complementary—Opt. learns a faithful semantic offset from the reference, and Guid. preserves text–concept correspondence at inference. More examples are in Appendix **C**.

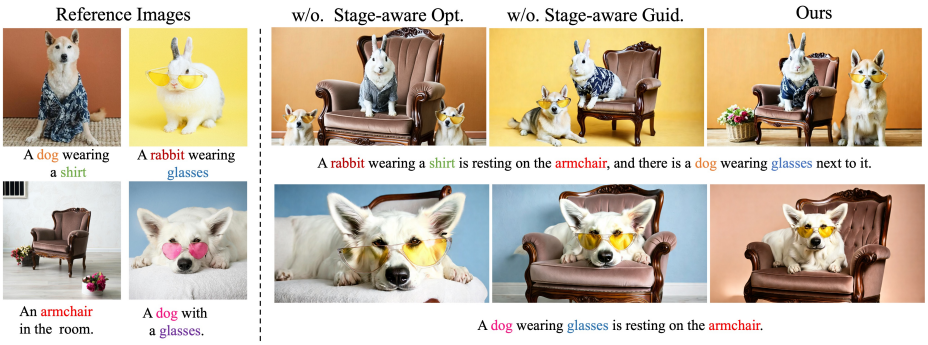

Fig. 7: Qualitative Ablation of Stage-Aware Components. We compare our full model with variants lacking key components. (a) w/o Stage-aware Opt.: Naive training fails to learn a high-fidelity concept. (b) w/o Stage-aware Guid.: Applying guidance uniformly across all stages causes structural artifacts and concept bleeding. (c) Ours: The synergy of both strategies yields the best compositional quality.

Table 2: User Study Results for Concept Generation (Values represent average scores on a 1-5 scale).

Metric	**ConceptWeaver**	MS Diffusion	Dreambooth
Concept Consistency	**4.06**	3.75	3.02
Text Adherence	**3.78**	3.14	2.88
Aesthetic Quality	**4.12**	3.75	3.05

Sensitivity to Stage-Split Partitions. Table 3 also reports several stage-length splits against our standard 10/10/10. Ending the Blueprint or Instantiation stage too early (e.g., 5/10/15, 10/5/15) causes concept drift and clearly lowers $S^{5.5}_{GPT}$, while more balanced partitions (8/12/10, 12/8/10) approach but do not match ours. This indicates that ConceptWeaver is robust around the balanced configuration while still benefiting from a stage-aware schedule.

Analysis of Concept Guidance Scale. The scale $w_{c,i}$ in Eq. 7 balances the learned concept against the base text prompt. As shown in Figure 8, a low w_c yields a generic concept driven by the prompt, and increasing w_c progressively aligns the generated attributes (e.g., the style of the *glasses*/*necklace*, the identity of the *dog*) with the reference, providing a fine-grained knob for personalization strength. More examples are in Appendix **D**.

5.5 Applications

Motion Concept Transfer. Our framework achieves one-shot motion transfer by treating motion as a structural concept. As our probing (Sec. 4.1) shows its influence peaks in the Blueprint Stage, we apply our stage-aware optimization and guidance for motion exclusively in this early stage. This effectively decouples motion from appearance, enabling the high-fidelity transfer of dynamics from a

Table 3: Quantitative ablation on multi-concept generation. Top: removing each stage-aware component. Middle: sensitivity to the three-stage partition; each "$T_1/T_2/T_3$" lists the timesteps assigned to the Blueprint / Instantiation / Refinement stages. Our standard setting (10/10/10) is highlighted.

Setting	$S^{\mathrm{T}}_{\mathrm{CLIP}}\uparrow$	$S^{\mathrm{I}}_{\mathrm{CLIP}}\uparrow$	$S_{\mathrm{DINO}}\uparrow$	$S^{5.5}_{GPT}\uparrow$
w/o Stage-aware Opt.	34.87	73.99	67.83	3.61
w/o Stage-aware Guid.	34.41	74.05	66.46	3.34
5/10/15	35.03	75.67	68.54	3.79
10/5/15	34.74	75.93	68.47	3.67
8/12/10	35.11	**77.07**	68.68	4.01
12/8/10	35.09	76.83	68.70	3.93
Ours (10/10/10)	**35.36**	77.06	**69.02**	**4.14**

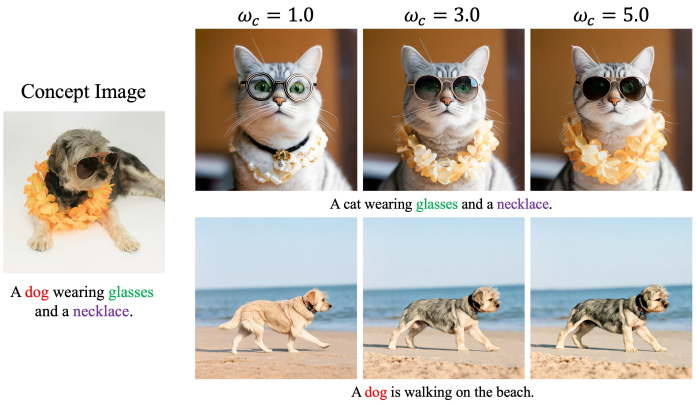

Fig. 8: Analysis of the Concept Guidance Scale w_c. As w_c increases, the visual attributes of the generated concepts align more closely with the reference image.

Fig. 9: Motion Concept Transfer with ConceptWeaver. We learn a motion concept from a single reference video (up) and apply it to new subjects defined by text prompts (down). The generated videos demonstrate high-fidelity motion replication while accurately synthesizing the new subject's appearance, showcasing effective disentanglement of motion and content.

reference video to new subjects, as illustrated in Figure 9. More visualization examples are shown in the Appendix **E**.

Customized Image Editing. ConceptWeaver can be extended to real image editing by integrating with the FlowEdit [22]. We first learn a concept offset $\Delta \boldsymbol{c}_{\mathrm{ref}}$

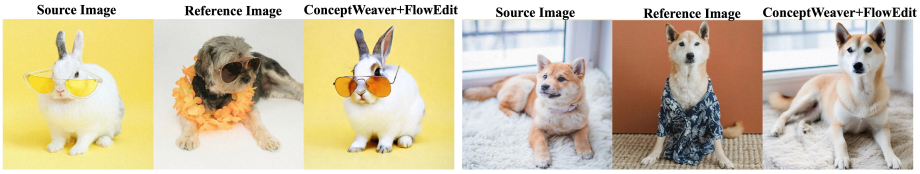

Fig. 10: Image Editing with ConceptWeaver and FlowEdit. By integrating with FlowEdit, we apply a visual concept learned from a reference image to a source image. The resulting edited image preserves the source's structure while faithfully incorporating the specific appearance of the learned concept.

from a reference image. This offset is then used to create a customized target velocity, $\boldsymbol{v}_{\text{target}} = \boldsymbol{v}_\theta(\boldsymbol{x}_t, t, \boldsymbol{c}_{\text{edit}} + \Delta\boldsymbol{c}_{\text{ref}})$, which guides the FlowEdit process. This synergy allows FlowEdit to preserve the source image's structure while our stage-aware guidance injects the high-fidelity appearance of the learned concept. As shown in Figure 10, this enables customized, one-shot editing of real images. More visualization examples are shown in the Appendix **F**.

6 Conclusion

In this work, we addressed the challenge of one-shot concept decoupling in flow models. We introduced a novel differential probing technique that revealed a universal three-stage generative framework—**Blueprint**, **Instantiation**, and **Refinement**—which governs how concepts are formed over time. This key insight demonstrated that concepts naturally disentangle during a pivotal mid-stage, providing a principled and highly effective window for manipulation. Guided by this discovery, we proposed ConceptWeaver, a framework that operationalizes our findings. It learns concept-specific offsets via a stage-aware optimization and applies them with a CWG mechanism. By aligning closely with the model's intrinsic dynamics rather than forcing external constraints, ConceptWeaver achieves high-fidelity compositional control from a single reference image. Our work highlights a shift towards principled, analysis-driven model control. While we manually categorize concepts, a promising future direction is to learn these temporal profiles automatically. We believe that understanding and leveraging the inherent structure of generative models is the key to unlocking the next frontier of precise and creative content synthesis.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62476011), and by the Beijing Natural Science Foundation (L252060).

References

1. Avrahami, O., Aberman, K., Fried, O., Cohen-Or, D., Lischinski, D.: Break-a-scene: Extracting multiple concepts from a single image. In: SIGGRAPH Asia 2023 Conference Papers. pp. 1–12 (2023)

2. Avrahami, O., Patashnik, O., Fried, O., Nemchinov, E., Aberman, K., Lischinski, D., Cohen-Or, D.: Stable flow: Vital layers for training-free image editing. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 7877–7888 (2025)
3. Chen, B., Zhao, M., Sun, H., Chen, L., Wang, X., Du, K., Wu, X.: Xverse: Consistent multi-subject control of identity and semantic attributes via dit modulation. arXiv preprint arXiv:2506.21416 (2025)
4. Chen, T.S., Siarohin, A., Menapace, W., Fang, Y., Lee, K.S., Skorokhodov, I., Aberman, K., Zhu, J.Y., Yang, M.H., Tulyakov, S.: Multi-subject open-set personalization in video generation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6099–6110 (2025)
5. Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., et al.: Scaling rectified flow transformers for high-resolution image synthesis. In: Forty-first international conference on machine learning (2024)
6. Fan, W., Zheng, A.Y., Yeh, R.A., Liu, Z.: Cfg-zero*: Improved classifier-free guidance for flow matching models. arXiv preprint arXiv:2503.18886 (2025)
7. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022)
8. Garibi, D., Yadin, S., Paiss, R., Tov, O., Zada, S., Ephrat, A., Michaeli, T., Mosseri, I., Dekel, T.: Tokenverse: Versatile multi-concept personalization in token modulation space. ACM Transactions On Graphics (TOG) **44**(4), 1–11 (2025)
9. Gat, I., Remez, T., Shaul, N., Kreuk, F., Chen, R.T., Synnaeve, G., Adi, Y., Lipman, Y.: Discrete flow matching. Advances in Neural Information Processing Systems **37**, 133345–133385 (2024)
10. Gu, Y., Wang, X., Wu, J.Z., Shi, Y., Chen, Y., Fan, Z., Xiao, W., Zhao, R., Chang, S., Wu, W., et al.: Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. Advances in Neural Information Processing Systems **36**, 15890–15902 (2023)
11. HaCohen, Y., Chiprut, N., Brazowski, B., Shalem, D., Moshe, D., Richardson, E., Levin, E., Shiran, G., Zabari, N., Gordon, O., et al.: Ltx-video: Realtime video latent diffusion. arXiv preprint arXiv:2501.00103 (2024)
12. Hao, S., Han, K., Lv, Z., Zhao, S., Wong, K.Y.K.: Conceptexpress: Harnessing diffusion models for single-image unsupervised concept extraction. In: European Conference on Computer Vision. pp. 215–233. Springer (2024)
13. Hertz, A., Mokady, R., Tenenbaum, J., Aberman, K., Pritch, Y., Cohen-Or, D.: Prompt-to-prompt image editing with cross attention control. arXiv preprint arXiv:2208.01626 (2022)
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
15. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
16. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
17. Huang, Q., Fu, S., Liu, J., Jiang, H., Yu, Y., Song, J.: Resolving multi-condition confusion for finetuning-free personalized image generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 3707–3714 (2025)
18. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)

19. Jiang, Z., Han, Z., Mao, C., Zhang, J., Pan, Y., Liu, Y.: Vace: All-in-one video creation and editing. arXiv preprint arXiv:2503.07598 (2025)
20. Jin, J., Yu, Z., Shen, Y., Fu, Z., Yang, J.: Latexblend: Scaling multi-concept customized generation with latent textual blending. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23585–23594 (2025)
21. Kong, X., Zhang, Z., Guo, Y., Zhao, Z., Zhang, S., Rao, A.: Composing concepts from images and videos via concept-prompt binding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14800–14810 (2026)
22. Kulikov, V., Kleiner, M., Huberman-Spiegelglas, I., Michaeli, T.: Flowedit: Inversion-free text-based editing using pre-trained flow models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19721–19730 (2025)
23. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.Y.: Multi-concept customization of text-to-image diffusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1931–1941 (2023)
24. Labs, B.F.: Flux. `https://github.com/black-forest-labs/flux` (2024)
25. Li, G., Yang, Y., Song, C., Zhang, C.: Flowdirector: Training-free flow steering for precise text-to-video editing. arXiv preprint arXiv:2506.05046 (2025)
26. Liang, Q., Liu, Z., Ostrow, M., Fiete, I.: How diffusion models learn to factorize and compose. Advances in Neural Information Processing Systems **37**, 15121–15148 (2024)
27. Ling, P., Bu, J., Zhang, P., Dong, X., Zang, Y., Wu, T., Chen, H., Wang, J., Jin, Y.: Motionclone: Training-free motion cloning for controllable video generation. arXiv preprint arXiv:2406.05338 (2024)
28. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
29. Lipman, Y., Havasi, M., Holderrieth, P., Shaul, N., Le, M., Karrer, B., Chen, R.T., Lopez-Paz, D., Ben-Hamu, H., Gat, I.: Flow matching guide and code. arXiv preprint arXiv:2412.06264 (2024)
30. Liu, L., Ma, T., Li, B., Chen, Z., Liu, J., Li, G., Zhou, S., He, Q., Wu, X.: Phantom: Subject-consistent video generation via cross-modal alignment. arXiv preprint arXiv:2502.11079 (2025)
31. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
32. Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., Zhu, J.: Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. Advances in neural information processing systems **35**, 5775–5787 (2022)
33. Mao, F., Hao, A., Chen, J., Liu, D., Feng, X., Zhu, J., Wu, M., Chen, C., Wu, J., Chu, X.: Omni-effects: Unified and spatially-controllable visual effects generation. arXiv preprint arXiv:2508.07981 (2025)
34. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
35. Peng, Y., Cui, Y., Tang, H., Qi, Z., Dong, R., Bai, J., Han, C., Ge, Z., Zhang, X., Xia, S.T.: Dreambench++: A human-aligned benchmark for personalized image generation. arXiv preprint arXiv:2406.16855 (2024)
36. Po, R., Yang, G., Aberman, K., Wetzstein, G.: Orthogonal adaptation for modular customization of diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7964–7973 (2024)

37. Polyak, A., Zohar, A., Brown, A., Tjandra, A., Sinha, A., Lee, A., Vyas, A., Shi, B., Ma, C.Y., Chuang, C.Y., et al.: Movie gen: A cast of media foundation models. arXiv preprint arXiv:2410.13720 (2024)
38. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
39. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 22500–22510 (2023)
40. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
41. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456 (2020)
42. Tewel, Y., Gal, R., Chechik, G., Atzmon, Y.: Key-locked rank one editing for text-to-image personalization. In: ACM SIGGRAPH 2023 conference proceedings. pp. 1–11 (2023)
43. Vinker, Y., Voynov, A., Cohen-Or, D., Shamir, A.: Concept decomposition for visual exploration and inspiration. ACM Transactions on Graphics (TOG) **42**(6), 1–13 (2023)
44. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
45. Wang, W., Huang, M., Tu, Y., Mao, Z.: Dualreal: Adaptive joint training for lossless identity-motion fusion in video customization. arXiv preprint arXiv:2505.02192 (2025)
46. Wang, X., Fu, S., Huang, Q., He, W., Jiang, H.: Ms-diffusion: Multi-subject zero-shot image personalization with layout guidance. arXiv preprint arXiv:2406.07209 (2024)
47. Wei, Y., Zhang, S., Qing, Z., Yuan, H., Liu, Z., Liu, Y., Zhang, Y., Zhou, J., Shan, H.: Dreamvideo: Composing your dream videos with customized subject and motion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6537–6549 (2024)
48. Xie, C., Li, M., Li, S., Wu, Y., Yi, Q., Zhang, L.: Dnaedit: Direct noise alignment for text-guided rectified flow editing. arXiv preprint arXiv:2506.01430 (2025)
49. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: Cogvideox: Text-to-video diffusion models with an expert transformer. arXiv preprint arXiv:2408.06072 (2024)
50. Yoon, S.H., Li, M., Beaudouin, G., Wen, C., Azhar, M.R., Wang, M.: Splitflow: Flow decomposition for inversion-free text-to-image editing. Advances in Neural Information Processing Systems **38**, 153207–153225 (2026)
51. Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L.M., Shum, H.Y.: Dino: Detr with improved denoising anchor boxes for end-to-end object detection. arXiv preprint arXiv:2203.03605 (2022)
52. Zhang, S., Zhuang, J., Zhang, Z., Shan, Y., Tang, Y.: Flexiact: Towards flexible action control in heterogeneous scenarios. In: Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers. pp. 1–11 (2025)

53. Zhong, W., Yang, H., Liu, Z., He, H., He, Z., Niu, X., Zhang, D., Li, G.: Mod-adapter: Tuning-free and versatile multi-concept personalization via modulation adapter. arXiv preprint arXiv:2505.18612 (2025)
54. Zhou, Z., Deng, Y., He, X., Dong, W., Tang, F.: Multi-turn consistent image editing. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15792–15801 (2025)