Supplementary Material
Corgi: Cached Memory Guided Video Generation

Overview

In this supplement, we first provide human evaluation to measure the effectiveness of our Corgi method in comparison to other state-of-the-art approaches (Sec. A), and then we extend the ablation study to evaluate our method design (Sec. B). We discussed the limitations and negative impact in Sec. C. Additional preliminary details for subject-guided finetuning are provided in Sec. D and details of our customized dataset MainCharacter21 are in Sec. E. Additional video visualizations are included in the supplementary material folder.

A. Human Evaluation

We conduct human evaluations for our method against several baselines. Each pair of generated results was evaluated by five participants, including both experts in the field and individuals without specific background knowledge. Our evaluation set includes 21 pairs of generated results from *Corgi* and open-source baseline methods (FreeNoise [27] and Gen-L-Video [35]), as well as generated multi-scene video samples from closed-source methods (three videos from Animate-A-Story [9] and five videos from VideoDirectGPT [20]), where we directly use the provided prompts for generation. For comparisons with FreeNoise [27] and Gen-L-Video [35], we collected 105 responses (21 video pairs, evaluated by 5 participants), and for Animate-A-Story [9] and VideoDirectGPT [20], we had 15 (3 video pairs, evaluated by 5 participants) and 25 (5 video pairs, evaluated by 5 participants) responses, respectively. We mixed our generated results with those from baselines, presenting the participants with story prompts and corresponding videos generated by these methods in a randomized order. Participants were prompted to compare the consistency, faithfulness, diversity, and overall video quality of the multi-scene videos, asking, e.g., "Which video is more consistent/faithful/diverse/has higher quality?" We present the proportion of samples where a higher number of users preferred our examples as being better in Tab. 4. The results show that our *Corgi* method consistently outperforms the baseline methods across key metrics. Particularly notable are its high preference scores in both short-term and long-term consistency, as well as diversity score and overall video quality, with a remarkable 92% preference over VideoDirectGPT [20] for overall quality. Although *Corgi* shows a lower preference in visual faithfulness and long-term consistency compared to Animate-A-Story [9], this may be due to the limited comparison set, as we had access to only one group of conditioning images from Animate-A-Story [9]. These results show the effectiveness of *Corgi* in multi-scene video generation.

Table 4. **Human Preference.** We conduct a human evaluation to compare *Corgi* against four baseline methods: FreeNoise (**F**) [27], Gen-L-Video (**G**) [35], Animate-A-Story (**A**) [9], and VideoDirectorGPT (**V**) [20]. In each paired comparison, our method was preferred predominantly (over 50%) over the baselines across various metrics. It is important to note that **F** and **G** do not utilize input images for conditioning, hence visual faithfulness was not evaluated for these methods. For **A** due to limited access to only one set of images used for generating a single video, we report the visual faithfulness score solely for this specific comparison.

Evaluation (%)		Ours > **F** [27]	Ours > **G** [35]	Ours > **A** [9]	Ours > **V** [20]
Consistency	Short	87.62	77.14	66.67	88.00
	Long	84.76	94.28	46.67	92.00
Faithfulness	Visual	–	–	40.00	96.00
	Textual	63.81	78.09	60.00	84.00
Creativity		63.81	70.48	53.33	84.00
Overall Quality		81.90	84.76	66.67	92.00

B. Additional Ablation Study

In the main paper, we provide ablation studies to evaluate the impact of coverage-based selective caching (Sec. 4.4). Here we ablate two other method design choices of *Corgi*: cached latent conditioning and clip-by-clip sampling.

Cached Latent Conditioning. In our proposed method, cached latent conditioning plays an important role in controlling the generation process across video clips. To evaluate the effectiveness of this design choice, we conduct ablation studies to compare different scenarios:

1. Removing linear weight degradation (as in Eqn. 4) and maintaining a constant degree of influence across all frames, thus $\lambda_k = 0.02$ (**Constant**).
2. Setting the initial weight (λ_0) too low while still maintaining the linear weight degradation, reducing the cached latent influence, which may result in generated videos that are not visually faithful to the input subjects, $\lambda_0 = 0.002$ (**Low**).
3. Setting the initial weight (λ_0) too high while still maintaining the linear weight degradation, resulting in cached latents having an excessive influence on the generated frames, potentially limiting diversity, $\lambda_0 = 0.5$ (**High**).
4. Using the default setting with linear weight degradation, $\lambda_0 = 0.02$ (**Linear**).

As shown in Tab. 5 and Fig. 6, linear weight degradation enables for a gradual transition, allowing the generated frames to deviate from the initial frame while still maintaining visual faithfulness to the input subjects. However, maintaining a constant degree of influence across all frames, without the linear weight degradation, leads to an overly rigid adherence to the cached latents. This affects the natural transition of the generated videos, resulting in minimal motion movement throughout the clips. Setting the cached latents weight too high limits diversity by overly constraining the content to the initial frame cached latents, while a too low weight diminishes visual faithfulness and consistency as frames have little influence from cached latents, deviating from earlier frames, both compromising overall video consistency. While constant weight outperforms others in terms of short-term consistency and visual faithfulness as expected, it significantly affected diversity and long-term consistency.

Table 5. **Ablation on Cached Latent Conditioning.** We compare different scenarios: constant weight (**Constant**), low weight (**Low**), high weight (**High**) and linear weight degradation (**Linear**). The results show that our proposed linear weight degradation approach achieves the optimal tradeoff of consistency, faithfulness, and diversity.

	Consistency (↓)		Faithfulness (↑)		Diversity (↑)
Weight Setting	Short-term	Long-term	Visual	Textual	
Constant	**7.42** ± 4.37	17.93 ± 5.02	**86.44** ± 8.24	35.94 ± 5.73	38.64 ± 6.74
Low	21.36 ± 6.15	23.48 ± 4.63	75.89 ± 8.06	32.18 ± 7.93	49.27 ± 5.15
High	8.57 ± 5.82	25.14 ± 4.85	54.38 ± 9.53	21.49 ± 3.81	34.96 ± 7.36
Linear (ours)	12.58 ± 5.76	**11.63** ± 5.23	85.83 ± 6.38	**37.11** ± 4.27	**52.84** ± 3.28

Clip-by-clip Sampling. Furthermore, we conduct an ablation study to evaluate the impact of the self-attention operation with cached latents concatenation in clip-by-clip sampling. Keeping the same experiment settings for other parts, we evaluate **w/ concatenation** (Eqn. 5) and **w/o concatenation** (Eqn. 7), the results are in Tab. 6. Our ablation study shows that incorporating the proposed cached latent concatenation for self-attention improves performance. When the cached latent concatenation was omitted for self-attention, the ability to preserve the visual appearance of the input subjects was largely weakened and it frequently results in jittery motion and object distortions (Fig. 7).

Table 6. **Ablation on Clip-by-Clip Sampling.** We conduct an ablation study on self-attention concatenation during clip-by-clip sampling, comparing scenarios with and without cached latent concatenation. The results show that with concatenation improves video quality and consistency. The ✓denotes using concatenation.

	Consistency (↓)		Faithfulness (↑)		Diversity (↑)
Concatenation	Short-term	Long-term	Visual	Textual	
✓	**12.58** ± 5.76	**11.63** ± **5.23**	**85.83** ± 6.38	37.11 ± 4.27	**52.84** ± **3.28**
	14.31 ± 6.58	12.95 ± 4.17	74.23 ± 7.82	**40.03** ± **5.22**	50.17 ± 5.39

C. Limitations

While our method offers promising results in multi-scene video generation, it still has its limitations. For example, we observed that when novel subjects in the story prompts are not specified, e.g., in Fig. 9 (A), with only the corgi images as input, the generated results will merge the features of multiple subjects (corgi and squirrel). Another failure case we observed is if in the input images, there is always some part attached to the target subject (e.g., in Fig. 9 (B), the tree branch is attached to the owl), then this feature will be propagated via the cached latents to the final generated videos. Additionally, our diversity metric does not capture whether this diversity aligns with the intended story. As in some cases, it could be preferable for subsequent clips to have similar visual content. Quantifying "desirable" or "reasonable" diversity is subjective and context-dependent. An interactive UI is ideal but beyond our scope. Future work e.g. adaptive weighting or human-in-the-loop approaches for user-selected intermediate images could further improve quality. These challenges open up new opportunities for future research exploration.

Negative Impact. While our method aims to enable multi-scene video generation, there is a risk that it could be exploited to create misleading or inappropriate content, which underscores the need for robust filters and stricter regulatory frameworks to prevent misuse in the future.

D. Finetuning Preliminary

Here we provide additional preliminary details for the subject-guided finetuning [30]. Diffusion models are a type of probabilistic generative model designed to learn data distributions. They achieve this by progressively denoising a sample initially drawn from a Gaussian distribution, effectively reducing its noise through each step of the process. As denoted in Sec. 3.2, with pretrained T2I diffusion model $\hat{\mathbf{x}}_\theta$ and conditioning vector $\mathbf{c} = \tau_\theta(\mathbf{p})$, and initial noise map $\boldsymbol{\epsilon}$ drawn from a normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$, as well as the ground-truth image $\mathbf{x}$, the original training objective is:

$$\mathbb{E}_{\mathbf{x},\mathbf{c},\boldsymbol{\epsilon},t}\left[w_t\|\hat{\mathbf{x}}_\theta(\alpha_t\mathbf{x}+\sigma_t\boldsymbol{\epsilon},\mathbf{c})-\mathbf{x}\|_2^2\right], \tag{8}$$

α_t, σ_t, w_t control the noise schedule and sample quality. We follow Dreambooth [30] and leverage the class-specific prior preservation loss during finetuning:

$$\mathbb{E}_{\mathbf{x},\mathbf{c},\boldsymbol{\epsilon},\boldsymbol{\epsilon}',t}[w_{t'}\|\hat{\mathbf{x}}_\theta(\alpha_{t'}\mathbf{x}_{\text{pr}}+\sigma_{t'}\boldsymbol{\epsilon}',\mathbf{c}_{\text{pr}})-\mathbf{x}_{\text{pr}}\|_2^2], \tag{9}$$

where $\mathbf{x}_{\text{pr}} = \hat{\mathbf{x}}(\mathbf{z}_{t_1}, \mathbf{c}_{\text{pr}})$ from the pretrained and frozen T2I model. $\mathbf{z}_{t_1} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is random initial noise and $\mathbf{c}_{\text{pr}} := \tau_\theta(f(\text{"a [name of class]"}))$ is a conditioning vector. The loss of T2I finetuning is the combination of the both training objectives above:

$$\begin{aligned}\mathbb{E}_{\mathbf{x},\mathbf{c},\boldsymbol{\epsilon},\boldsymbol{\epsilon}',t}[w_t\|\hat{\mathbf{x}}_\theta(\alpha_t\mathbf{x}+\sigma_t\boldsymbol{\epsilon},\mathbf{c})-\mathbf{x}\|_2^2+\\ \lambda w_{t'}\|\hat{\mathbf{x}}_\theta(\alpha_{t'}\mathbf{x}_{\text{pr}}+\sigma_{t'}\boldsymbol{\epsilon}',\mathbf{c}_{\text{pr}})-\mathbf{x}_{\text{pr}}\|_2^2].\end{aligned} \tag{10}$$

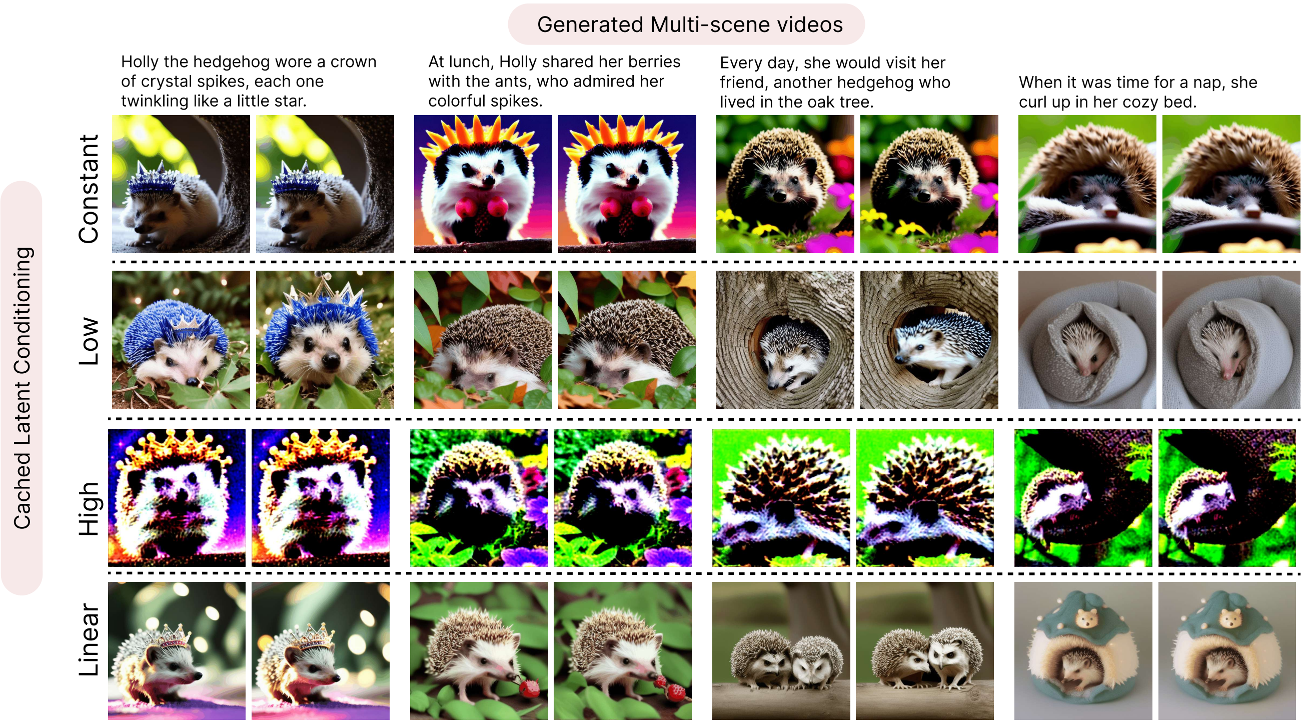

Generated Multi-scene videos
Holly the hedgehog wore a crown of crystal spikes, each one twinkling like a little star.
At lunch, Holly shared her berries with the ants, who admired her colorful spikes.
Every day, she would visit her friend, another hedgehog who lived in the oak tree.
When it was time for a nap, she curl up in her cozy bed.
Cached Latent Conditioning
Constant
Low
High
Linear

Figure 6. **Ablation study on Cached Latent Conditioning.** We examine different weight settings for cached latent conditioning: constant weight across all frames **(Constant)**, low weight **(Low)**, high weight **(High)**, linear weight degradation **(Linear)**. The **Linear** approach achieves the best balance between consistency, faithfulness, and diversity. **Constant** leads to overly rigid adherence and the videos have minimum motion and appear similar to static images rather than dynamic video sequences, **High** limits diversity and the generated results look unrealistic, and **Low** diminishes visual faithfulness to input subjects.

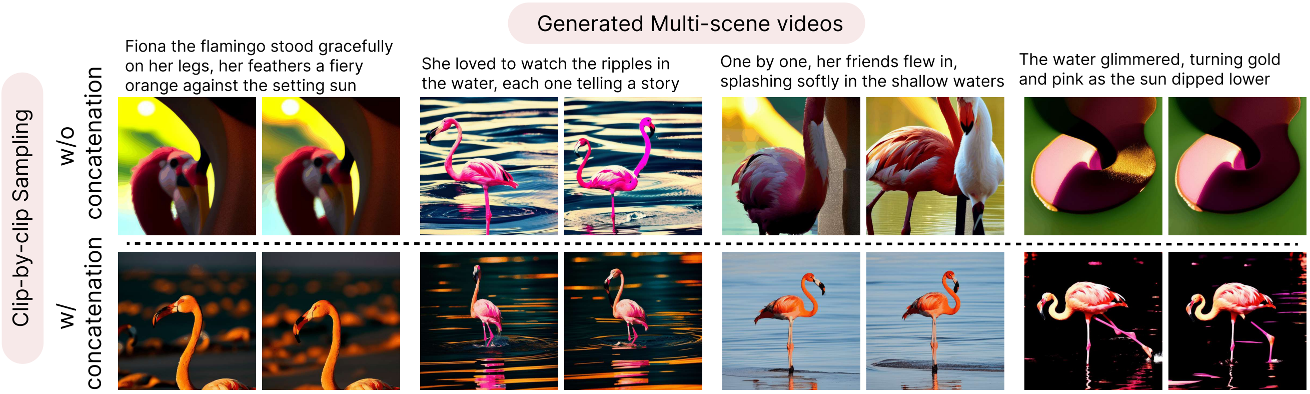

Generated Multi-scene videos
Fiona the flamingo stood gracefully on her legs, her feathers a fiery orange against the setting sun
She loved to watch the ripples in the water, each one telling a story
One by one, her friends flew in, splashing softly in the shallow waters
The water glimmered, turning gold and pink as the sun dipped lower
Clip-by-clip Sampling
w/o concatenation
w/ concatenation

Figure 7. **Ablation study on Clip-by-Clip Sampling.** We compare the impact of cached latent conditioning on the generated videos. The model without cached latent **(w/o concatenation)** suffers from jittery motion and object distortions, while the model with cached latent **(w/ concatenation)** maintains visual appearance of input subjects and generates more stable and high-quality videos. This demonstrates the effectiveness of the proposed clip-by-clip sampling approach in preserving visual consistency and faithfulness to the input subjects.

E. MainCharacter21

In our study, we introduce the **MainCharacter21** dataset, including 21 unique subjects, with each subject represented by 3 to 5 images. Fig. 8 shows three images of each subject. We show a list of sample instruction prompts (Tab. 7) used to generate story prompts, along with three example story prompts (Tab. 8, 9, 10) created by MLLM [22, 23] given instruction prompts. It is important to note that, as we use rare tokens (e.g. V*) plus subjects during the T2I finetuning stage, we similarly added the rare tokens to the story prompts before subjects and pronouns in

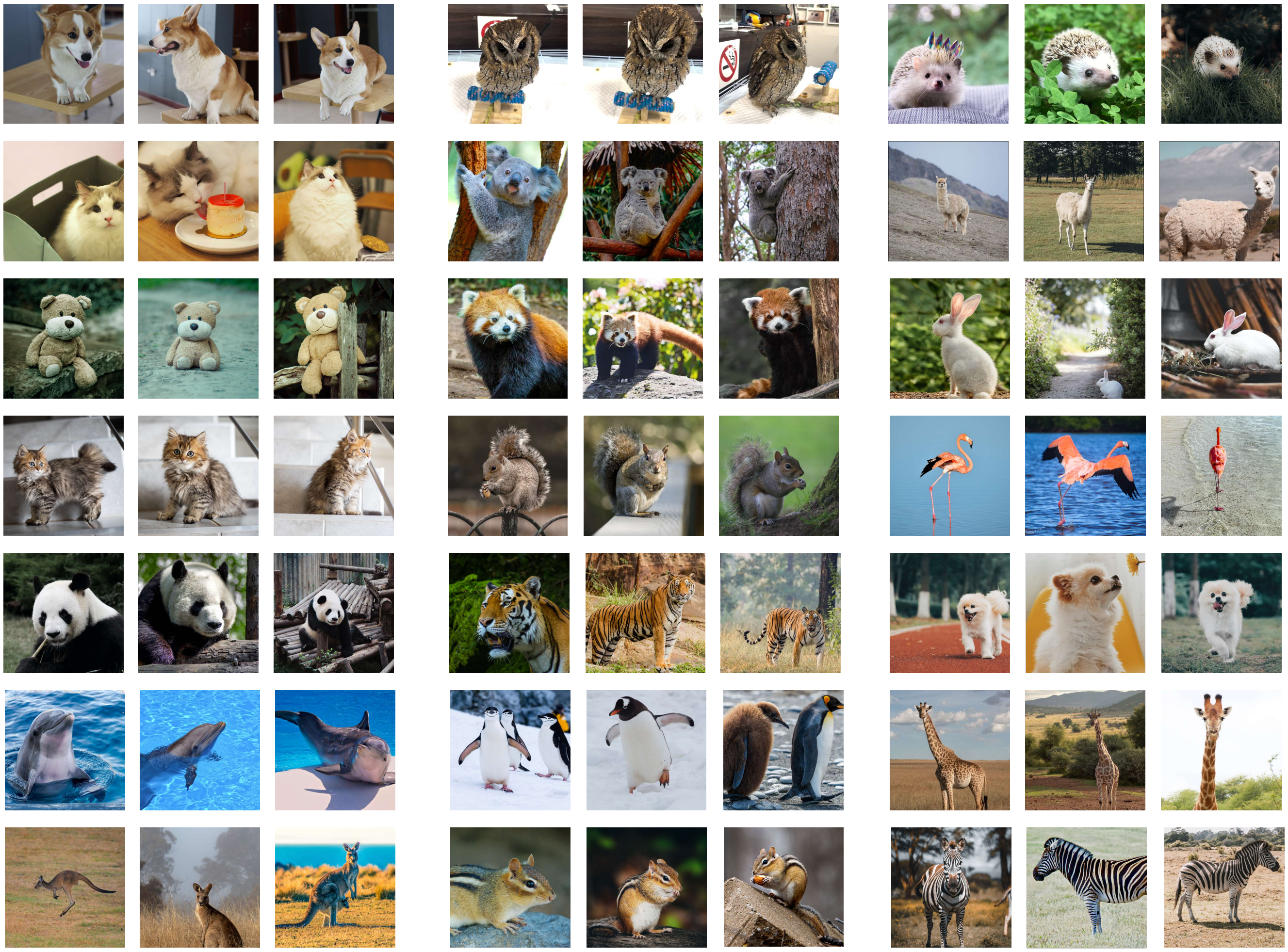
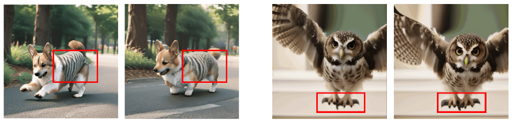


Figure 8. **MainCharacter21.** This figure illustrates our dataset MainCharacter21, including images from 21 distinct subjects, with three sample images per subject.



(A) A corgi saw a squirrel and chases after it. (B) A baby owl learns how to fly.

Figure 9. **Limitations.** Feature disentanglement for image-conditioned video generation still remains challenging. As shown in (A), the features of a corgi and a squirrel are mistakenly combined when the input images only includes the corgi. Additionally, in (B), the base T2I model's limitations in contextual understanding and a tendency to overfit to features that appear across all images used for fine-tuning result in incorrect feature attachment.

the prompts were adjusted accordingly during inference.

Table 7. Sample Instruction Prompts

`Inspired by the photo, write a story for a children's book, consisting of 7 sentences.`
`Write a 9-sentence tale about two individuals reuniting under surprising circumstances using the image as inspiration.`
`Narrate a 4-sentence adventure about discovering something invaluable, drawing inspiration from the image.`
`Craft a 5-sentence story about unexpected turns in life, drawing from the image's atmosphere.`
`Using the image as a foundation, write a 9-sentence tale about a life lesson.`

Table 8. Sample Story Prompts 1

Sidney the squirrel scurried around the park, his little heart full of glee. He found a perfect acorn, shiny and brown, right for tea. His fluffy tail flickered as he nibbled away, happy as can be. He played peek-a-boo with the children, who laughed merrily. Sidney had a secret stash, hidden under the oak tree. He’d jump from branch to branch, the leaves whispering, "Catch me!" His friends, the birds, would sing as he danced. When it rained, he snuggle in his cozy warm and dry. And as the stars appeared, Sidney would dream of tomorrow’s joyous spree.

Table 9. Sample Story Prompts 2

Holly the hedgehog wore a crown of crystal spikes, each one twinkling like a little star. She loved to explore the garden, her crown catching the light and casting rainbows everywhere. She snuffled through the leaves, her tiny feet padding softly on the earth. Every day, Holly would visit her friend, another hedgehog who lived in the oak tree. At lunch, Holly shared her berries with the ants, who admired her colorful spikes. In the evening, Holly would sit and watch the stars, her crown shimmering along with them. When it was time for a nap, she curl up in her cozy bed.

Table 10. Sample Story Prompts 3

Fiona the flamingo stood gracefully on her legs, her feathers a fiery orange against the setting sun. She loved to watch the ripples in the water, each one telling a story. One by one, her friends flew in, splashing softly in the shallow waters. The water glimmered, turning gold and pink as the sun dipped lower. Fiona and her friend danced in the twilight, creating a whirlpool of colors with their wings. As the stars began to twinkle, she settled down, nestling together in the warm sand.