A. Utilization of Large Language Models (LLMs)

In this study, Large Language Models (LLMs) are employed at the sentence level to assist in linguistic refinement. Their use was strictly confined to improving grammatical accuracy and overall readability of the manuscript. All research concepts, methodological designs, experimental processes, and analytical findings remain entirely original and have been solely contributed by the authors.

B. Detailed Experimental Settings

This section elaborates on the experimental setup, including all relevant hyperparameter choices.

B.1. Pre-training Settings

The results presented in Tab. 1 are derived using the pre-training configurations outlined in Tab. 4. Due to computational resource constraints, Exponential Moving Average (EMA) decay was not applied during the training of IOMM-L and IOMM-XL. All models were pre-trained on the Megalith-10M [35] and text-to-image-2M [18] datasets (except for IOMM-XL), comprising approximately 11 million images in total. Each image was resized so that its shorter edge was 512 pixels while preserving the original aspect ratio, then a central crop was applied to obtain a 512×512 image. Notably, since neither dataset provides images at a resolution of 1024×1024, we did not deploy high-resolution pre-training.

Table 4. **Pre-training settings.**

METHOD	**IOMM-B**	**IOMM-L**	**IOMM-XL**
Optimization			
Optimizer	AdamW		Muon
β	$(0.9, 0.95)$		$(0.9, 0.95)$
Learning rate	1e-4		1e-4
Max gradient norm	1.0		1.0
Weight decay	0.0		0.0
Training Configuration			
Generative Model Size	1.6B	2.7B	6B
Training data type	Image-only	Image-only	Image-only
EMA decay	0.999	-	-
Global batch size	1024	512	4096
Image token mask ratio r	0.85	0.85	0.45

B.2. Finetuning Settings

We fine-tuned the two models (B&L) at resolutions of 512 and 1024, respectively, using the pre-training settings specified in Tab. 4. The fine-tuning datasets include BLIP3o-60K [9], Echo-4o-Image [59], and ShareGPT-4o-Image [8], collectively comprising approximately 210,000 high-resolution images (except for IOMM-XL). All images in these datasets are at 1024×1024 resolution. For fine-tuning at both 512 and 1024 resolutions, we applied central cropping to resize images to the target resolution.

B.3. UMM Finetuning Settings

The results presented in Tab. 2 were obtained using the fine-tuning configurations specified in Tab. 6. For OpenUni-L, we performed full fine-tuning on both the connector module and the generative model. In contrast, for Qwen-Image-20B, we applied Low-Rank Adaptation (LoRA) [20] to fine-tune the model. Both models utilized a frozen understanding module. Additionally, due to computational constraints, Exponential Moving Average (EMA) decay was not implemented for Qwen-Image-20B.

Table 5. **Finetuning settings.**

METHOD	**IOMM-B**		**IOMM-L**		**IOMM-XL**
Resolution	512	1024	512	1024	512
Optimization					
Optimizer	AdamW		AdamW		Muon
β	$(0.9, 0.95)$		$(0.9, 0.95)$		$(0.9, 0.95)$
Learning rate	1e-4		1e-4		1e-4
Max gradient norm	1.0		1.0		1.0
Weight decay	0.0		0.0		0.0
Generative Model Size	1.6B	1.6B	2.7B	2.7B	6B
Training Configuration					
Training data type	Mix	Mix	Mix	Mix	Mix
EMA decay	0.999	0.999	-	-	-
Global batch size	256	96	256	96	256
Image token mask ratio r	0.85	0.85	0.85	0.85	0.45
Mix ratio λ	0.5	0.5	0.5	0.5	0.5

Table 6. **UMM finetuning settings.**

METHOD	OpenUni-L	Qwen-Image-20B
Optimization		
Optimizer	AdamW	AdamW
β	(0.9,0.95)	(0.9,0.95)
Learning rate	1e-4	1e-4
Max gradient norm	1.0	1.0
Weight decay	0.0	0.0
Training Configuration		
Training data type	Mix/Image-only/Pair	Mix/Image-only/Pair
EMA decay	0.999	-
Global batch size	256	48
Epochs	12	5
Image token mask ratio r	0.85	0.85
Mix ratio λ	0.5	0.5
LoRA Configuration		
LoRA rank	-	64
LoRA alpha	-	64
LoRA dropout	-	0.0

C. More Results

C.1. DPGBench Evaluation Results

The Tab. 7 shows the detailed results of the DPGBench evaluation shown in Tab. 1.

C.2. WISE Evaluation Results

The Tab. 8 shows the detailed results of the WISE evaluation shown in Tab. 1.

Table 7. **DPGBench evaluation results.** Here BLIP3-o-8B* donates the model that is trained with an 30 million proprietary data.

METHOD	Global	Entity	Attribute	Relation	Other	Overall
Gen. Only						
SDv1.5 [42]	74.63	74.23	75.39	73.49	67.81	63.18
SD3-Medium [14]	87.90	91.01	88.83	80.70	88.68	84.08
SDXL [40]	83.27	82.43	80.91	86.76	80.41	74.65
PixArt-α [7]	74.97	79.32	78.60	82.57	76.96	71.11
FLUX.1-dev [2]	74.35	90.00	88.96	90.87	88.33	83.84
Unified Models						
Janus [48]	82.33	87.38	87.70	85.46	86.41	79.68
Janus-Pro-1B [11]	87.58	88.63	88.17	88.98	88.30	82.63
Janus-Pro-7B [11]	86.90	88.90	89.40	89.32	89.48	84.19
MetaQuery-B [38]	-	-	-	-	-	80.04
MetaQuery-L [38]	-	-	-	-	-	81.10
MetaQuery-XL [38]	-	-	-	-	-	82.05
BLIP3-o-4B [9]	-	-	-	-	-	79.36
BLIP3-o-8B* [9]	-	-	-	-	-	81.60
Ours						
IOMM-B 512	91.33	89.39	90.07	86.89	87.78	82.95
IOMM-B 1024	86.20	88.39	87.69	90.11	87.05	80.71
IOMM-L 512	83.28	83.61	84.69	83.46	79.83	76.09
IOMM-L 1024	79.27	82.00	80.93	82.81	78.68	72.26

C.3. Different training recipe

The results presented in Tab. 9 correspond to the training configurations depicted in Fig. 1b. All models underwent approximately 5 epochs of pre-training on a dataset comprising 11 million images, followed by 10 epochs of fine-tuning on a dataset of approximately 210,000 images. Notably, the model pre-trained exclusively on image-only data and fine-tuned on a mixed data achieved superior performance across most metrics in the GenEval benchmark.

C.4. Image Editing Results

Fig. 4 compares the image editing capabilities of models pre-trained exclusively on image-only data (right) and those pre-trained on image-text pairs (middle). The sole distinction between these models lies in their pre-training data type; all other hyperparameters and fine-tuning settings remain consistent. Despite in a *zero-shot setting*, the model pre-trained with image-only data demonstrates superior consistency with the original input image. For instance, in the first row, the right image closely resembles the raw input, while in the second and third rows, the right images maintain nearly identical gestures to the original.

C.5. UMM finetune result

Beyond generation quality, we evaluate world knowledge and reasoning using the WISE benchmark. As shown in the final column of Tab. 2 (with detailed breakdowns in Tab. 10), both text-image pair and mixed-data fine-tuning provide a substantial performance uplift for OpenUni-L (up to 0.10) and a modest improvement for Qwen-Image (0.01). In contrast, fine-tuning with image-only data proves detrimental across nearly all scenarios, significantly impairing the models' prompt-following ability—an effect particularly pronounced in larger models (see App. C.6 for a detailed analysis).

C.6. Generation results comparison of UMM finetuning

As illustrated in Fig. 5, fine-tuning enhances the model's performance on tasks requiring reasoning. Although the understanding module was frozen during fine-tuning, the model's improved alignment between images and text enables more accurate

Table 8. **WISE evaluation results.** Here BLIP3-o-8B* donates the model that is trained with an 30 million proprietary data.

METHOD	Cultural	Time	Space	Biology	Physics	Chemistry	Overall
Gen. Only							
SDv1.5 [42]	0.34	0.35	0.32	0.28	0.29	0.21	0.32
SDv2.1 [42]	0.30	0.38	0.35	0.33	0.34	0.21	0.32
SD3-Medium [14]	0.42	0.44	0.48	0.39	0.47	0.29	0.42
SDXL [40]	0.43	0.48	0.47	0.44	0.45	0.27	0.43
SD3.5-Large [14]	0.44	0.50	0.58	0.44	0.52	0.31	0.46
PixArt-α [7]	0.45	0.50	0.48	0.49	0.56	0.34	0.47
FLUX.1-dev [7]	0.48	0.58	0.62	0.42	0.51	0.35	0.50
Unified Models							
Show-o [56]	0.28	0.40	0.48	0.30	0.46	0.30	0.35
Janus [48]	0.16	0.26	0.35	0.28	0.30	0.14	0.23
Janus-Pro-1B [11]	0.20	0.28	0.45	0.24	0.32	0.16	0.26
Janus-Pro-7B [11]	0.30	0.37	0.49	0.36	0.42	0.26	0.35
MetaQuery-B [38]	0.44	0.49	0.58	0.41	0.49	0.34	0.46
MetaQuery-L [38]	0.56	0.57	0.62	0.48	0.63	0.42	0.55
MetaQuery-XL [38]	0.56	0.55	0.62	0.49	0.63	0.41	0.55
BAGEL [12]	0.44	0.55	0.68	0.44	0.60	0.39	0.52
BLIP3-o-4B [9]	-	-	-	-	-	-	0.50
BLIP3-o-8B* [9]	-	-	-	-	-	-	0.62
Ours							
IOMM-B 512	0.50	0.56	0.66	0.49	0.72	0.46	0.55
IOMM-B 1024	0.44	0.50	0.64	0.46	0.63	0.43	0.50
IOMM-L 512	0.48	0.56	0.63	0.49	0.64	0.51	0.53
IOMM-L 1024	0.44	0.48	0.59	0.43	0.58	0.44	0.48

Table 9. **Training recipe comparison.** The GenEval score of the models pre-trained with different training recipes. **Bold** denotes the best performance and underline denotes the second best performance.

Finetuning Recipe	Single Obj.	Two Obj.	Counting	Colors	Position	Color Attri.	Overall (↑)
Pre-trained with Text-Image Pair Data							
Image	**1.00**	**0.95**	0.63	0.87	0.50	0.72	0.78
Pair	0.99	0.92	0.76	0.91	0.87	0.69	0.86
Mix	0.99	0.91	0.80	0.92	0.90	0.75	0.88
Pre-trained with Image-Only Data							
Image	0.99	0.84	0.24	0.75	0.37	0.45	0.61
Pair	0.99	0.91	0.77	0.93	0.87	0.75	0.87
Mix	0.99	0.92	**0.83**	**0.94**	**0.91**	0.75	**0.89**

generation of desired details. What's more, a qualitative comparison between the original Qwen-Image model and our fine-tuned version. Our method enhances the model's ability to generate images with richer visual detail and improved alignment to the textual prompt.

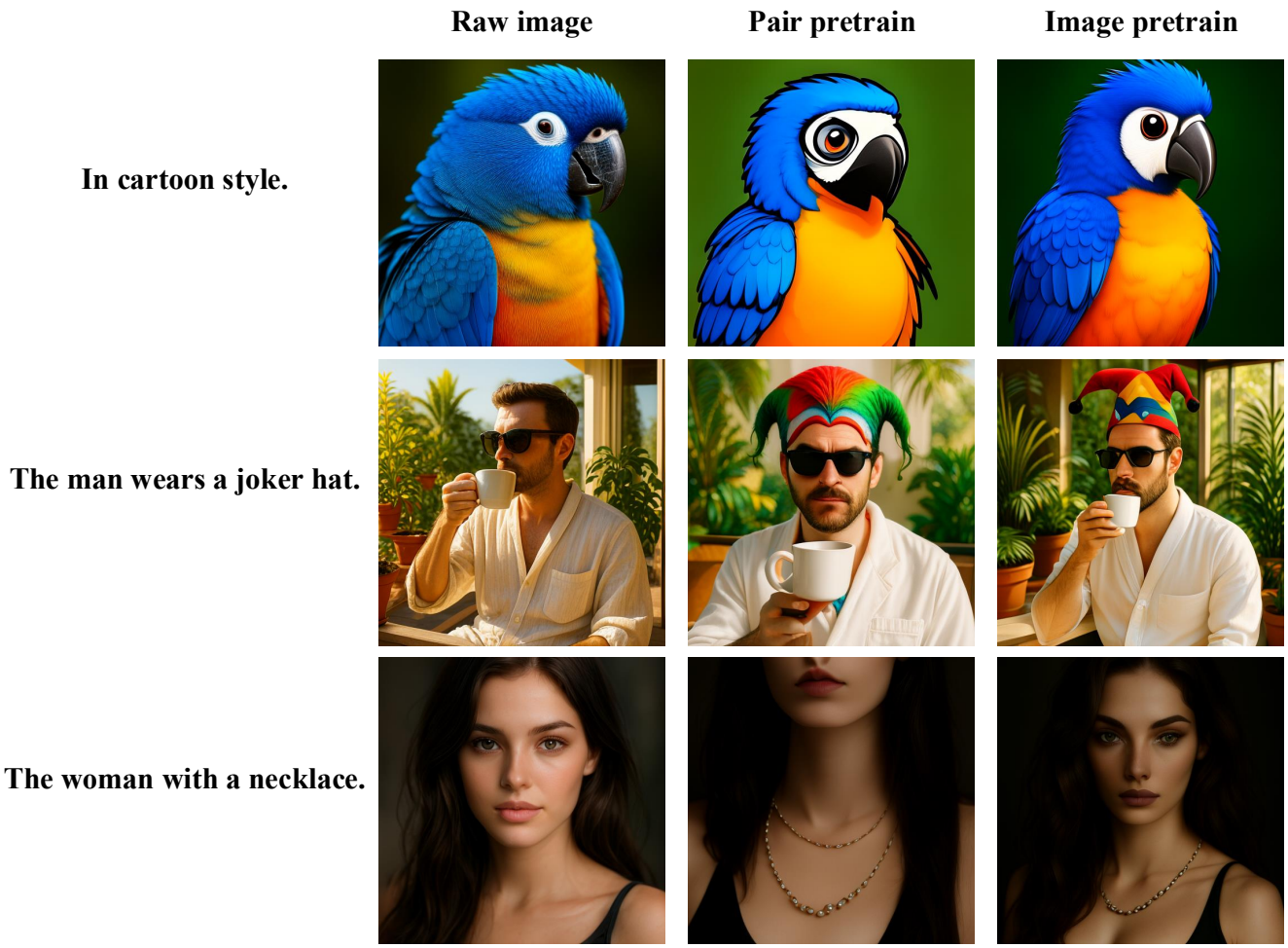


Figure 4. **Image editing ability with different pre-training method.**

Table 10. **UMM finetuning WISE results.** Notation A⊕B denotes the result obtained by combining methods A and B.

METHOD	Res.	NFEs	Cultural	Time	Space	Biology	Physics	Chemistry	Overall
OpenUni-L [50]	512	20×2	0.51	0.45	0.58	0.39	0.50	0.30	0.52
⊕Image finetuning	512	20×2	0.46	0.52	0.66	0.49	0.51	0.29	0.49
⊕Pair finetuning	512	20×2	0.63	0.58	0.74	0.57	0.71	0.44	0.62
⊕Mix finetuning	512	20×2	0.60	0.58	0.70	0.51	0.64	0.46	0.59
Qwen-Image [49]	512	50×2	-	-	-	-	-	-	-
⊕Image finetuning	512	50×2	0.39	0.42	0.56	0.32	0.50	0.28	0.41
⊕Pair finetuning	512	50×2	0.62	0.62	0.76	0.56	0.74	0.36	0.62
⊕Mix finetuning	512	50×2	0.62	0.64	0.81	0.56	0.70	0.36	0.63
Qwen-Image [49]	1024	50×2	0.62	0.63	0.77	0.57	0.75	0.40	0.62
⊕Image finetuning	1024	50×2	0.28	0.35	0.52	0.40	0.40	0.28	0.35
⊕Pair finetuning	1024	50×2	0.63	0.63	0.77	0.62	0.72	0.37	0.63
⊕Mix finetuning	1024	50×2	0.64	0.63	0.78	0.57	0.73	0.38	0.63

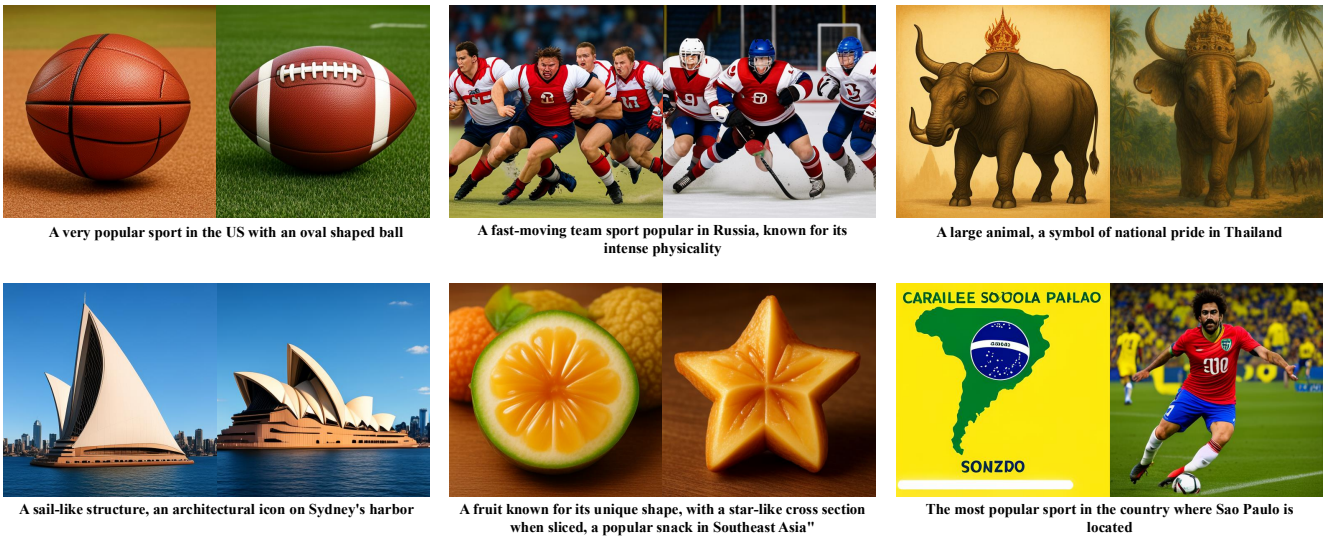
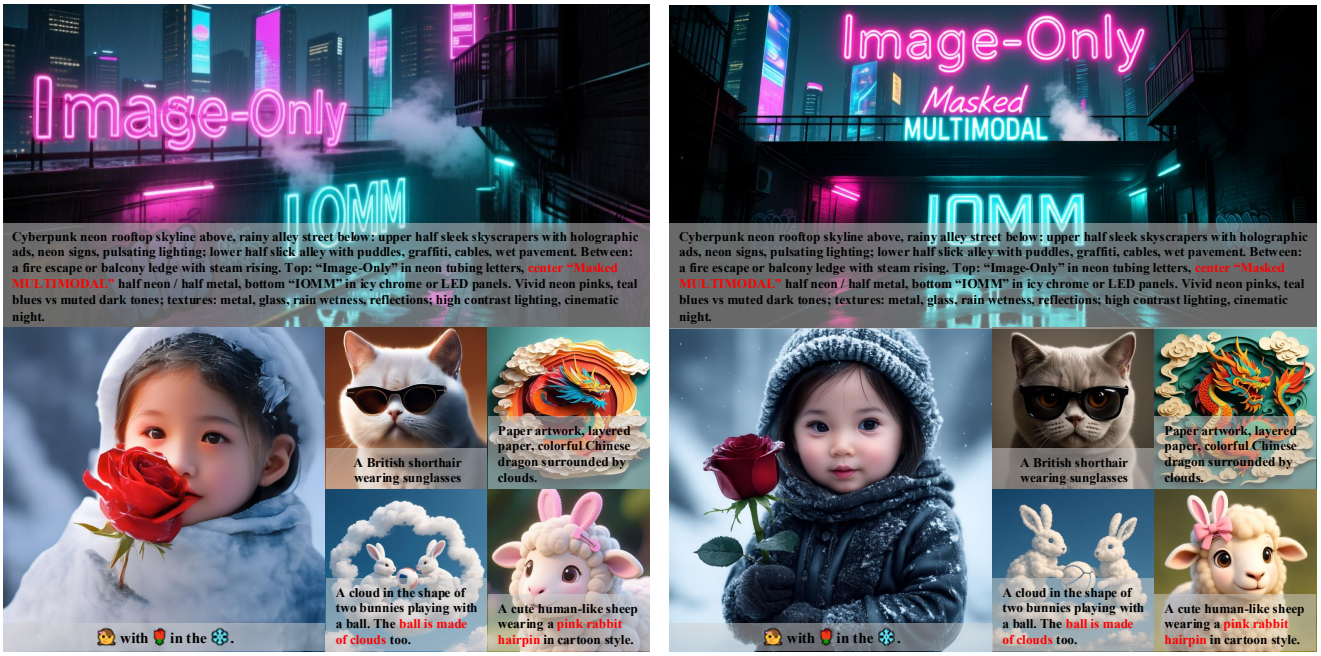


Figure 5. **Generation results of OpenUni-L before and after finetuning.** The left one is the image generated by the original OpenUni-L, while the right one is generated by the OpenUni-L after finetuning.



(a) **Baseline Qwen-Image generation.** (b) **Our fine-tuned Qwen-Image generation.**

Figure 6. **(a, b)** Qualitative comparison between the original Qwen-Image model and our fine-tuned version. Our method enhances the model's ability to generate images with richer visual detail and improved alignment to the textual prompt.

C.7. Prompts details

The prompts used in Fig. 1a are as follows, from left to right, top to bottom.

- Hyper-detailed macro photograph of a mechanical hummingbird crafted from gold filigree and sapphire gears, sipping nectar from a chrome rose; studio lighting, 200 mm macro lens, razor-sharp focus with creamy bokeh.
- A photo of a bear made entirely of autumn leaves.
- A fox wearing a suit and tie reading a newspaper at a café.
- a tiny astronaut hatching from an egg on the moon
- A man sipping coffee on a sunny balcony filled with potted plants, wearing linen clothes and sunglasses, basking in the

morning light.

- A cloud in the shape of two bunnies playing with a ball. The ball is made of clouds too.
- Portrait of a noble samurai android wearing lacquered carbon-fiber armor and cherry-blossom patterns; Rembrandt lighting, 50 mm f/1.2, hyperreal pores and brushed metal textures.
- A hot air balloon in the shape of a heart. Grand Canyon
- A captivating photograph of an exquisite wooden dragon sculpture, skillfully carved with intricate details and realistic scales. The dragon is poised on a tree branch, its grand wings spread wide, revealing a mesmerizing woodland landscape below. The sky is painted with a symphony of soft blues and yellows, as the sun casts its final rays beyond the horizon. The dragon's glass eyes lend it a lifelike presence.
- Close-up portrait of a young woman with light skin and long brown hair, looking directly at the camera. Her face is illuminated by dramatic, slatted sunlight casting shadows across her features, creating a pattern of light and shadow. Her eyes are a striking green, and her lips are slightly parted, with a natural pink hue. The background is a soft, dark gradient, enhancing the focus on her face. The lighting is warm and golden.
- A lone figure in dark robes ascends worn stone steps toward a glowing light in an ancient temple entrance. Ornate arches, lush greenery, and intricate carvings adorn the scene, evoking a mystical, high-fantasy atmosphere reminiscent of works by artists like Randy Vargas, with cinematic lighting and epic storytelling.
- A whimsical scene featuring a plush toy bear wearing a blue sweater, positioned in the foreground, holding a butterfly on its raised arm. The bear is surrounded by a field of vibrant blue flowers, likely nemophila, creating a lush and colorful foreground. In the background, Mount Fuji rises majestically, its snow-capped peak sharply contrasting against a clear blue sky. The mountain is framed by fluffy white clouds and a line of dark green trees at its base. The butterfly, with its intricate black and orange wings, adds a touch of realism to the playful composition.
- A candid midday portrait of a young East Asian woman with dark braided hair, laughing softly at the camera while cradling a steaming mug of coffee. She wears a tattered band t-shirt with a faded punk logo, frayed gray collar, and missing sleeve button. The background shows peeling floral wallpaper and a rusted folding chair beneath a window with harsh noon sunlight. Shot as a grainy film photograph with high contrast and sharp focus on her animated expression.
- professional portrait photo of an anthropomorphic cat wearing fancy gentleman hat and jacket walking in autumn forest.