MIST: Moment-Aligned Invariant Stability Transform for Robust Flow Matching

Liang Peng*, Deqing Li*, Yujia Wu, Hao Meng, Kuan Cao, Yu Wu, Xiaoxiao Xu, Lin Qu

Alibaba Group, Hangzhou, China **Equal Contribution* *Contact: jingjie.pl@alibaba-inc.com*

Motivation

Classifier-Free Guidance (CFG) enhances quality in flow-matching models, but high guidance scales **violate optimal transport dynamics**, causing:

- **Mode Collapse** — Overly simplified, stylized outputs
- **Over-Saturation** — Extreme color distortion and artifacts
- **Variance Explosion** — Uncontrolled trajectory divergence

Velocity Moment Decomposition

CFG velocity: $v_t^{\text{CFG}} = v_t(x) + w \cdot \delta v_t$

We decompose the CFG-induced distributional shift into two velocity-moment components:

$$\mathcal{D}_{\text{shift}}(p_t^w) \sim \begin{cases} \underbrace{w \cdot \mathbb{E}_{x\sim p_t}[\delta v_t]}_{\text{Linear Barycentric Drift (1st Moment)}} \\ \underbrace{\frac{1}{2}w^2 \cdot \mathbb{E}_{x\sim p_t}\left[\delta v_t^\top \mathcal{M}(x)\, \delta v_t\right]}_{\text{Quadratic Energetic Instability (2nd Moment)}} \end{cases}$$

Linear Barycentric Drift (1st Moment)

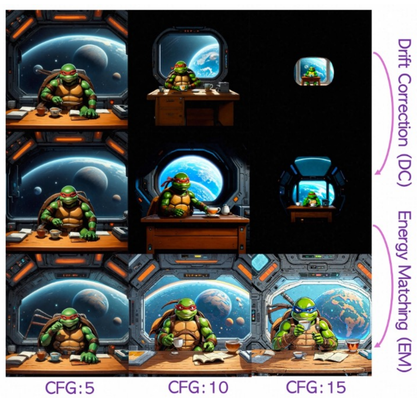
Scales linearly with w. Rigid translation of probability mass away from the learned manifold.

Quadratic Energetic Instability (2nd Moment)

Scales with w^2. Injects surplus kinetic energy, violating minimal-cost transport.

Effect of Each Component

Progressive correction at high guidance scales:



CFG → +DC (Drift Correction) → +DC+EM (Energy Matching). DC alone alleviates mode collapse; full method restores integrity.

Key Properties

- **Training-free:** No retraining or fine-tuning
- **Plug-and-play:** Compatible with any flow-matching model
- **Hierarchical:** Global alignment + Local regulation
- **Unified:** Works for both image and video generation

Contributions

1. We identify velocity distribution shift as the root cause of CFG instability, decomposing it into Linear Drift (1st moment) and Energetic Instability (2nd moment).
2. We propose MIST, a training-free plug-and-play method combining Invariant Alignment (IA) and Stability Thresholding (ST).
3. MIST outperforms CFG and variants across SD3, Flux-dev, and Wan2.2, establishing new SOTA for both T2I and T2V.

MIST: Proposed Framework

MIST (Moment-aligned Invariant Stability Transform) is a **training-free, plug-and-play** guidance method for flow-matching models:

Component 1: Invariant Alignment (IA)

Drift Correction (DC)

Targets O(w) linear drift. Projects onto zero-mean subspace:

$$\delta v_t^{dc} = \delta v_t - \mathbb{E}_{x\sim p_t}[\delta v_t]$$

Anchors guided velocity to unconditional prior.

Energy Matching (EM)

Targets O(w²) quadratic energy. Renormalizes kinetic energy:

$$v_t^{IA} = \mu + \left(v_t^{dc} - \mu\right) \cdot \frac{\sigma\left(v_t(x)\right)}{\sigma\left(v_t^{dc}\right) + \epsilon}$$

Prevents variance collapse and over-saturation.

Component 2: Stability Thresholding (ST)

Temporal Decay (TD)

Monotonicity constraint on guidance magnitude. Clips norms violating energy decay (dE/dt ≤ 0).

$$\hat{\delta v}_t = \delta v_t \cdot \min\left(1, \frac{\|\delta v_{t+1}\|_2}{\|\delta v_t\|_2}\right)$$

Enforces Lipschitz continuity for smooth convergence.

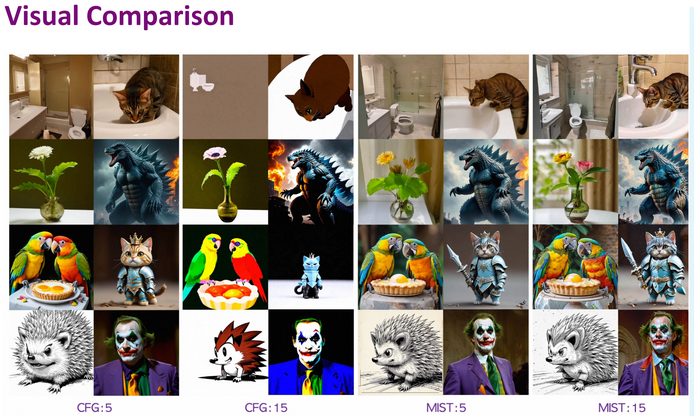
Spatial Suppression (SS)

Suppresses local singularities in high-curvature regions. Per-location ratio clipping:

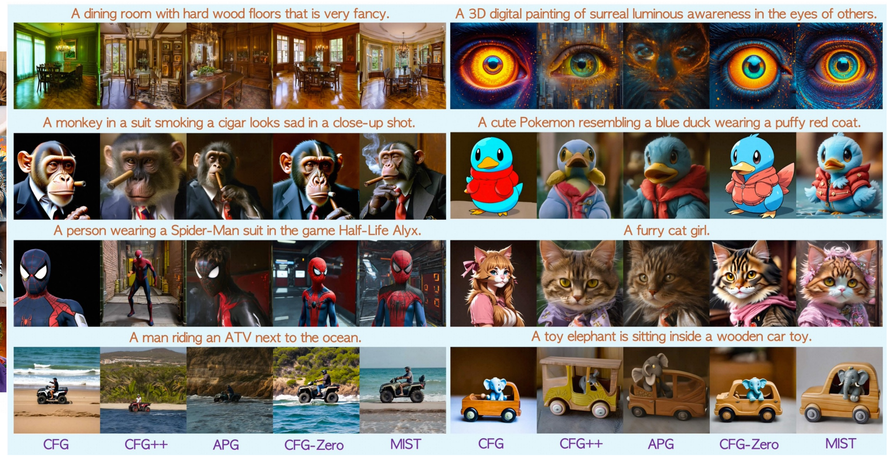
$$\delta v_{i,j}^{ST} = \hat{\delta} v_{i,j} \cdot \text{clip}\left(\frac{\gamma}{\rho_{i,j}}, 0, 1\right) \qquad \rho_{i,j} = \frac{w\left\|\hat{\delta} v_{i,j}\right\|_2}{\left\|v_{i,j}(x)\right\|_2 + \epsilon}$$

Prevents artifacts at boundaries and textures.

Visual Comparison



At w=15, CFG collapses while MIST maintains fidelity. At w=5, MIST also improves visual quality.



Quantitative Results

Table 1. Quantitative comparisons on HPD v2 benchmark. "G.S." is guidance scale; "P.S." is PickScore; "Aes." is aesthetic; "I.R." is ImageReward; "U.R." is UnifiedReward.

Model	G.S.	P.S.	Aes.	CLIP	HPS	I.R.	U.R.
CFG (Ho & Salimans, 2022)	5.0	22.78	5.984	37.03	29.66	1.0818	3.3988
CFG++ (Chung et al., 2025)		20.93	5.814	32.75	23.98	-0.0159	2.5197
APG (Sadat et al., 2025)		21.82	5.985	35.06	24.81	0.4964	2.9532
CFG-Zero (Fan et al., 2025)		22.84	6.014	36.89	**30.31**	1.0876	3.4190
MIST		**22.95**	**6.022**	**37.26**	30.22	**1.1126**	**3.4230**
CFG (Ho & Salimans, 2022)	10.0	22.44	5.866	36.57	29.21	1.0361	3.3662
CFG++ (Chung et al., 2025)		22.26	6.020	36.37	28.23	0.8044	3.1727
APG (Sadat et al., 2025)		22.42	6.040	36.24	27.37	0.8606	3.2494
CFG-Zero (Fan et al., 2025)		22.72	5.972	37.00	30.64	1.1558	3.4431
MIST		**23.02**	**6.053**	**37.33**	**31.29**	**1.2216**	**3.4958**
CFG (Ho & Salimans, 2022)	15.0	21.43	5.507	34.31	25.15	0.4922	2.9521
CFG++ (Chung et al., 2025)		22.60	6.051	36.98	29.76	1.0092	3.3510
APG (Sadat et al., 2025)		22.67	6.065	36.67	28.68	0.9934	3.3685
CFG-Zero (Fan et al., 2025)		22.25	5.824	36.60	29.27	1.0363	3.2907
MIST		**22.98**	**6.067**	**37.31**	**31.60**	**1.2482**	**3.4812**

Table 2. Ablation study on IA (Invariant Alignment) and ST (Stability Thresholding). Both components bring improvements.

Model	P.S.	Aes.	CLIP
CFG	22.44	5.866	36.57
+IA	22.92	6.032	37.16
+ST	22.72	5.950	37.22
+Both	**23.02**	**6.053**	**37.33**

Table 3. Ablation study on ST strategies. "SS" is spatial suppression and "TM" is temporal decay.

Model	P.S.	Aes.	CLIP
CFG	22.44	5.866	36.57
+SS	22.71	5.935	37.21
+TM	22.61	5.929	36.94
+Both	**22.72**	**5.950**	**37.22**

Table 4. Quantitative comparisons on DPG benchmark (Hu et al., 2024). "Attr." refers to "Attribute". MIST achieves state-of-the-art results on the overall metric.

Model	G.S.	Global	Entity	Attr.	Relation	Other	Overall
CFG (Ho & Salimans, 2022)	5.0	84.50	90.27	**88.38**	93.65	82.80	84.36
CFG++ (Chung et al., 2025)		79.79	82.89	80.39	90.35	70.40	74.70
APG (Sadat et al., 2025)		82.98	86.89	85.11	92.22	75.60	80.03
CFG-Zero (Fan et al., 2025)		84.50	90.48	88.28	93.60	81.90	84.97
MIST		**85.11**	**90.64**	88.37	**93.71**	**82.90**	**85.16**
CFG (Ho & Salimans, 2022)	10.0	82.29	90.49	88.23	93.49	83.70	84.51
CFG++ (Chung et al., 2025)		84.80	87.57	85.76	91.78	78.00	81.67
APG (Sadat et al., 2025)		**85.33**	89.22	86.97	93.38	79.10	83.09
CFG-Zero (Fan et al., 2025)		83.43	90.92	88.62	93.82	82.30	85.29
MIST		84.19	**91.41**	**88.64**	**94.13**	**85.20**	**85.87**
CFG (Ho & Salimans, 2022)	15.0	78.34	87.44	84.52	91.59	80.10	79.93
CFG++ (Chung et al., 2025)		**85.94**	88.37	86.60	92.28	79.50	82.96
APG (Sadat et al., 2025)		85.26	90.01	87.70	93.55	80.50	84.31
CFG-Zero (Fan et al., 2025)		81.00	89.85	87.51	93.16	82.00	83.40
MIST		83.97	**91.83**	**88.58**	**94.55**	**85.70**	**86.40**

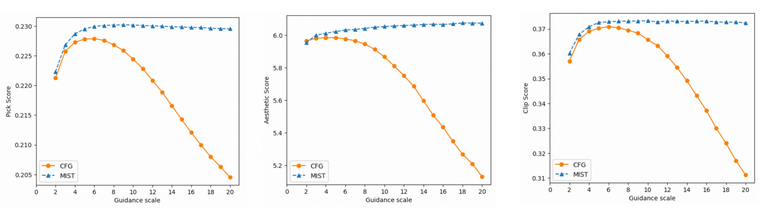


Table 8. Comparisons on Vbench benchmark. We use the recent Wan2.2 models as our base model. Compared to vanilla CFG, MIST improves both frame aesthetics and overall video quality.

Model	Guidance	Aesthetic Quality	Motion Smoothness	Overall Consistency	Spatial Relationship	Temporal Style	Quality Score	Semantic Score	Total Score
Wan2.2 5B (Wan et al., 2025)	CFG 4.0	58.69	**98.69**	24.81	75.38	24.81	83.02	71.19	80.65
	CFG 9.0	59.09	98.22	25.36	**80.67**	24.82	83.36	**74.74**	81.64
	MIST 9.0	**59.69**	98.53	**25.55**	80.15	**25.02**	**83.89**	74.05	**81.92**
Wan2.2 A14B (Wan et al., 2025)	CFG 4.0	62.69	98.20	26.14	79.86	23.92	83.93	75.81	82.30
	CFG 9.0	62.64	97.73	26.23	**80.95**	**24.26**	83.63	76.66	82.24
	MIST 9.0	**62.82**	**98.23**	**26.24**	80.54	24.13	**84.07**	**76.76**	**82.61**