

Qwen3-VL-235B-A22B

KI-Modellportrait · Quellenstand 2026-10-31

QV

Monogramm-Fallback; kein sauber belegtes Logo-Asset (Wikimedia-Dateien mit Marken- und Lizenz-Risiko, siehe Integrationsnotiz)
Monogramm-Darstellung, kein Logo-Asset übernommen.

Offenes Vision-Language-Flaggschiff der Qwen-Reihe von Alibaba Cloud: Mixture-of-Experts mit 235 Mrd. Parametern (22 Mrd. aktiv), als Instruct- und als Thinking-Edition veröffentlicht.

Quellen geprüft am 31. Oktober 2026 · Redaktion: Philipp Paulik

Entwickler: [Alibaba Cloud](#) →

Technische Einordnung

Qwen3-VL-235B-A22B ist die größte Variante der Qwen3-VL-Serie und verarbeitet Text, Bilder und Videos zu Textausgaben. Es gibt zwei Editionen mit eigenen Gewichten und API-IDs: Instruct für direkte Antworten und Thinking für multimodales Reasoning mit Denkphase. Dieses Porträt behandelt beide Editionen der 235B-A22B-Größe; kleinere Qwen3-VL-Varianten sind separate Modelle.

Zugang & Nutzung

Offene Gewichte (Hugging Face, ModelScope) unter Apache 2.0; gehostete API über Alibaba Cloud Model Studio mit den Modell-IDs qwen3-vl-235b-a22b-instruct und qwen3-vl-235b-a22b-thinking (Regionen u.a. Peking, Singapur, Frankfurt, Virginia); zusätzlich Qwen Chat und bei Hugging Face gelistete Drittanbieter (u.a. Novita, DeepInfra).

Apache License 2.0 für die Modellgewichte beider Editionen. Davon getrennt gelten die Nutzungsbedingungen der jeweiligen Plattform (Model Studio, Qwen Chat, Drittanbieter).

Grenzen & richtig lesen

Das API-Kontextfenster (131.072 Token) ist kleiner als das native Kontext der Gewichte (256K, erweiterbar bis 1M). Function Calling ist in der API nur in der Region Peking ausgewiesen, Fine-tuning über Model Studio nicht. Benchmark-Angaben stammen vom Hersteller. API-Preise werden in CNY ausgewiesen und unterscheiden sich je nach Region. Selbsthosting der BF16-Gewichte erfordert Multi-GPU-Hardware.

Benchmark-Belege

Für dieses Portrait ist noch kein ausreichend zugeordnetes Ergebnis erfasst. Fehlende Werte sind keine Null. Wir übernehmen keine Werte aus ähnlich benannten Modell- oder Benchmark-Versionen.

Steckbrief

Entwickler

Qwen-Team, Alibaba Cloud

Veröffentlichung

23. September 2025 (Instruct und Thinking)

Architektur

Mixture-of-Experts, 235 Mrd. Parameter gesamt, 22 Mrd. pro Token aktiviert; HF-Metadaten nennen 236 Mrd. Parameter in BF16

Lizenz der Gewichte

Apache 2.0

Eingabe → Ausgabe

Text, Bild, Video → Text

Kontext (Gewichte)

Nativ 256K Token, erweiterbar bis 1M

Kontext (API)

Fenster 131.072 Token; max. Eingabe 129.024 (Instruct) bzw. 126.976 (Thinking); max. Ausgabe 32.768

OCR

32 Sprachen

API-IDs

qwen3-vl-235b-a22b-instruct, qwen3-vl-235b-a22b-thinking

Angaben laut den unten verlinkten Primärquellen.

Quellen & Stand

1. [Hugging Face: Qwen/Qwen3-VL-235B-A22B-Instruct](#)
Geprüft: 2026-10-31
2. [Hugging Face: Qwen/Qwen3-VL-235B-A22B-Thinking](#)
Geprüft: 2026-10-31
3. [GitHub: QwenLM/Qwen3-VL](#)
Geprüft: 2026-10-31
4. [Alibaba Cloud Community: Qwen3-VL: Sharper Vision, Deeper Thought, Broader Action](#)
Geprüft: 2026-10-31
5. [Alibaba Cloud Model Studio: qwen3-vl-235b-a22b-instruct - Model Info](#)
Geprüft: 2026-10-31
6. [Alibaba Cloud Model Studio: qwen3-vl-235b-a22b-thinking - Model Info](#)
Geprüft: 2026-10-31

Redaktionelle Zusammenfassung mit KI-Unterstützung, verantwortet von Philipp Paulik. Kein eigener Leistungstest. Änderbare Anbieterangaben gelten für den angegebenen Quellenstand.

KI Fachblatt · Redaktion: Philipp Paulik · Quellenstand: 2026-10-31

Vollständiges Webportrait: <https://kifachblatt.de/ki-modelle/qwen3-vl-235b-a22b>

Statischer Export, kein Live-Datenblatt. Preise, Zugriff und Anbieterangaben können sich ändern. Quellen und ihre jeweiligen Abrufstände stehen oben. [Impressum](#) · [Transparenz](#)